Studies in Optics and Optoelectronics

By
Steven John Feinman Byrnes

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Physics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Feng Wang, *Chair*
Professor Yuen-Ron Shen
Professor Richard Saykally

Fall 2012

Abstract

Studies in Optics and Optoelectronics

by

Steven John Feinman Byrnes

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Feng Wang, Chair

This thesis will detail four projects aimed at understanding and applying the principles of optics and optoelectronics.

In Chapter 1, we describe phase-sensitive sum-frequency vibrational spectroscopy (PS-SFVS), a nonlinear optical technique that can probe the molecular structure of the top few monolayers of a liquid-vapor interface. We use this technique to investigate the air-water interface, using a number of water samples with different dissolved salts. The information is used to draw inferences about the surface propensity of these salt ions—information that can shed light on both atmospheric chemistry and water solvation theory. We also give a detailed description of the experimental methodology for PS-SFVS, its rationale, and the issues that can arise.

PS-SFVS measurements, such as those described in Chapter 1, can be fruitfully used by comparing them with the signal predicted by molecular simulation. However, the relationship between a molecular configuration and its nonlinear optical signal is not thoroughly understood in the theoretical chemistry community. In particular, the procedures used in the literature to predict an PS-SFVS signal within a molecular simulation have been ambiguous, depending on arbitrary parameters. In Chapter 2, we review PS-SFVS theory at a fundamental level, then map it to modern simulation methods, thereby explaining the ambiguities as consequences of improper truncation of a multipole expansion. A molecular-dynamics simulation of the water-air interface is used as an example, illustrating the consequences of different simulation methods and suggesting which ones should be most accurate.

Chapter 3 explores a different aspect of nonlinear optics: The compression and characterization of ultrafast pulses of light. These pulses have been explored for a variety of scientific and technological applications. Ideally, an optical pulse can be reduced in duration up to the limit imposed by its spectral bandwidth via the uncertainty principle. However, the presence of "nonlinear chirp" (different frequencies arriving at different times in a nonlinear fashion), which is especially common in mode-locked fiber lasers, can be a major factor preventing the shortening of a pulse. We describe a new technology, a type of patterned glass phase plate, that promises to reduce nonlinear chirp in a convenient, adjustable, inexpensive, and high-throughput manner. After showing simulations, we describe how we made the plate, and then how we used frequency-resolved optical gating (FROG) to watch the plate change

the duration and shape of a pulse from a fiber laser.

Finally, Chapter 4 discusses a new architecture for solar cells that uses the field effect, rather than the traditional p-n junction, to separate charge. This could be advantageous for semiconductor materials that are difficult to dope to both p- and n-type, such as oxides, sulfides, and nanoparticles. We discuss the underlying physics and rule-of-thumb design principles, along with both finite element simulations and experimental verifications.

Contents

List of Figures

Related publications and patents

Some of the work described herein has been published elsewhere.

- The results of Chapter 1 were published in:

  > C. S. Tian†, S. J. Byrnes†, H.-L. Han, Y. R. Shen, "Surface Propensities of Atmospherically Relevant Ions in Salt Solutions Revealed by Phase-Sensitive Sum Frequency Vibrational Spectroscopy." J. Phys. Chem. Lett. **2**, 1946 (2011).

- The results of Chapter 2 were published in:

  > S. J. Byrnes, P. L. Geissler, Y. R. Shen, "Ambiguities in Surface Nonlinear Spectroscopy Calculations." Chem. Phys. Lett. **516**, 115 (2011).

  (Parts of this text are reproduced from the publication, with permission from the coauthors and thanks to Elsevier B.V.)

- The work discussed in Chapter 3 are protected by a provisional patent (filed 2012). A publication is in prep.

- The results of Chapter 4 were published in:

  > W. Regan†, S. Byrnes†, W. Gannett, O. Ergen, O. Vazquez-Mena, F. Wang, A. Zettl, "Screening-Engineered Field-Effect Solar Cells." Nano Lett. (2012).

  Work is also protected by a provisional patent (filed 2011).

Acknowledgments

I am indebted to many people for their help during this work.

First and foremost, I thank my advisors Feng Wang and Ron Shen, who guided every step of these projects with care, intelligence, expertise, and dedication. They have been excellent and trusted mentors, and their examples will continue to inspire me.

I thank Phill Geissler for his enthusiastic and skillful guidance and supervision during my excursion into theoretical chemistry (Chapter 2). I also thank Joyce Noah-Vanhoucke for help with Appendix 2.D, and Patrick Varilly for help with LAMMPS.

I thank Richard Saykally for offering excellent advice, help, and guidance on numerous occasions. I also thank Dale Otten for helpful discussions about nonlinear optics.

I thank my colleagues from the Wang and Shen groups for their encouragement, discussions, advice, teaching, and help on countless occasions. I am particularly lucky to have learned nonlinear optics from Chuan-Shan Tian, a generous and patient teacher who understood every aspect of his work down to the deepest levels.

I especially thank Will Regan, my main collaborator for the work described in Chapter 4. He has been always full of great ideas, and has worked tirelessly to make them happen. I also thank Will Gannett, Kris Erickson, and Alex Zettl for their help in this work.

My graduate-school tuition and stipend were generously funded by a National Defense Science and Engineering Graduate Fellowship, and by a National Science Foundation Graduate Research Fellowship. I thank these institutions, as well as the US taxpayers who ultimately fund them!

Last but not least, I am extremely lucky to have two loving parents, a wonderful brother, and a new wife to share my life with!

1 Phase-sensitive sum-frequency spectroscopic measurement of air-aqueous solution interfaces

1.1 Overview

1.1.1 SFG

Sum-frequency generation (SFG) is a nonlinear-optical process where light at frequencies ω_1 and ω_2 interact to generate light at the "sum frequency" $\omega_{\mathrm{SF}} = \omega_1 + \omega_2$. In the electric-dipole-order approximation, the SFG process is described by:

$$\vec{P}_{\mathrm{SF}} = \overset{\leftrightarrow}{\chi}^{(2)} : \vec{E}_1 \vec{E}_2 \tag{1.1}$$

where $\vec{P}_{\mathrm{SF}}$ is the sum-frequency polarization, $\vec{E}_1$ and $\vec{E}_2$ are the electric fields at frequency ω_1 and ω_2, and the equation holds at every point in space. (Written out with Cartesian indices j, ℓ, m, the equation is: $P_{\mathrm{SF},j} = \chi^{(2)}_{j\ell m} E_{1\ell} E_{2m}$.) To lowest order, SFG is symmetry-forbidden in a centrosymmetric medium; therefore, SFG primarily probes the surface.

SFG is a generalization of *second-harmonic generation* (SHG) (where $\omega_1 = \omega_2$). Although SFG is experimentally more difficult than SHG (two beams must be separately generated, then aligned both spatially and temporally), it is a more flexible and powerful technique. In this work, we use SFG as a surface-sensitive vibrational spectroscopy, by sweeping ω_1 across an infrared vibrational mode.

1.1.2 The air/water interface

This study uses SFG to investigate the air/water interface. Water interfaces play key roles in many areas of science, from protein folding to corrosion chemistry, but many aspects of them remain poorly understood [1, 2]. One branch of this field involves the behavior of dissolved ions at an air/water interface: In what concentrations are the ions present at the surface, and how exactly are they incorporated into the molecular surface structure of water? These questions are particularly important in atmospheric chemistry, because some chemical reactions in the atmosphere are known to occur in a heterogeneous fashion, where a vapor molecule reacts directly with a dissolved molecule at the surface of a water aerosol droplet [1, 3–5] (Fig. 1.1).

The simplest understanding of inorganic ions at the water surface is the Onsager model [6], in which the water and air are treated as a smooth, featureless dielectrics. By classical electrostatics, any charge in the water is repelled away from the interface by the image-charge effect. However, recent studies have confirmed that this picture is an oversimplification, and some ionic species can, in fact, be found at the outermost surface of water [7]. The effects are difficult to predict theoretically, as an ion's surface affinity is highly sensitive to poorly-determined simulation parameters [8]. Therefore, experimental measurements of the surface of water solutions are particularly important.

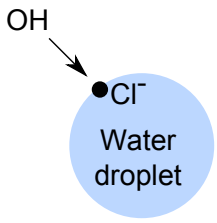

Figure 1.1 – Postulated heterogeneous reaction between a *vapor* molecule (neutral OH derived from UV-generated ozone) and a *dissolved* ion (Cl^-) at the outermost surface. (After Ref. [3].)

SFG is one of only a few techniques that can reliably probe the top few monolayers of a liquid, offering information unavailable by any other technique. For example, surface tension [9], surface potential [10], and zeta potential [10, 11] measurements provide some insight over the whole integrated surface profile, but cannot specify any finer details. X-ray photoemission [12] and resonant ultraviolet SHG [7] measurements provide information about the concentration profile near the topmost surface, but unlike SFG, provide no insight into the structure or bonding environment at the surface.

Therefore, this study provides SFG measurements of the surface of water, including both neat water and water with a variety of dissolved ions, focusing particularly on ionic species that are important in atmospheric chemistry: Cations Na^+, K^+, NH_4^+, and anions Cl^-, NO_3^-, SO_4^{2-}.

1.2 SFG details

1.2.1 SFG as a surface probe

In a centrosymmetric medium, such as bulk water, SFG is symmetry-forbidden in the electric-dipole approximation [13]. In particular, if the medium is centrosymmetric, then symmetry demands that

$$\vec{P}_{\mathrm{SF}} = \overleftrightarrow{\chi}^{(2)} : \vec{E}_1\vec{E}_2 \implies (-\vec{P}_{\mathrm{SF}}) = \overleftrightarrow{\chi}^{(2)} : (-\vec{E}_1)(-\vec{E}_2)$$

from which it follows that $\overleftrightarrow{\chi}^{(2)} = 0$. At a surface or interface, however, the SFG signal is is *not* symmetry forbidden, due to structural asymmetry at the surface—for example, there may be more water molecules pointing up than down at the surface. One cause of structural asymmetry is the exigencies of how the hydrogen-bonding network terminates at the surface. Another possible cause of structural asymmetry, key to this study, is that an ionic double-layer can create a static electric field at the surface, which will reorient a fraction of the water molecules there.

An SFG surface response can also come from field gradients: As the index of refraction and local field factor change across the interface, the electric field changes rapidly over an

atomic scale, which may induce SFG in surface molecules [14]. However, for the purpose of this study, this aspect of the signal can be safely neglected, because the analysis is based on the *difference* between pure water and relatively-dilute salt solutions (ion concentration of a few percent), and these should have a similar field-gradient surface response.

At higher orders of perturbation theory, SFG is *not* forbidden in the bulk—an issue discussed further in Chapter 2. However, again, for the purpose of this study, the possibility of a bulk signal can be safely neglected, because it should be similar with or without salt in the concentrations used.

1.2.2 Phase matching

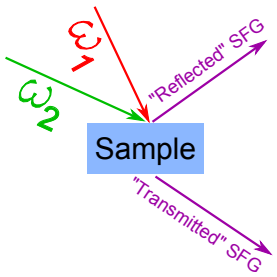

Figure 1.2 – The two directions in which coherent SFG is emitted.

For a planar interface, there are two phase-matched directions in which relatively strong, coherent SFG is emitted, shown in Fig. 1.2. If the interface is at $z = 0$, then coherent emission requires $k_{1x} + k_{2x} = k_{\mathrm{SF},x}$ and $k_{1y} + k_{2y} = k_{\mathrm{SF},y}$. These constraints, along with the fact that $|\vec{k}_{\mathrm{SF}}|$ is fixed by the frequency, mean that there are exactly two allowed $\vec{k}_{\mathrm{SF}}$ values for coherent light: The "reflected" and "transmitted" directions as shown in Fig. 1.2. In this study, only the reflected SFG was measured, which was sufficient for the purpose at hand. (The transmitted direction gives some complementary information [15], see also Chapter 2.)

1.2.3 Theory of SFVS

The study described herein uses a specific type of measurement called Infrared-Visible Sum-Frequency Vibrational Spectroscopy ("SFVS") (Fig. 1.3). In this setup, one of the incoming beams (ω_1) is an infrared frequency which is resonant (or nearly-resonant) with an OH vibrational mode of H_2O. The other incoming beam (ω_2) is visible and off-resonance, as is the sum-frequency ω_{SF}. The measurement involves sweeping the frequency ω_1 across the resonance, hence performing vibrational spectroscopy. This can yield valuable information about the intermolecular bonding and environment.

To understand the type of information gleaned from SFVS, we review the basic theory behind it.

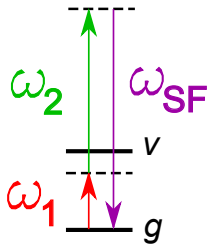

Figure 1.3 – Relation of SFG frequencies to molecular energy levels (the ground state g and the first excited vibrational state v), in an SFVS measurement.

When a molecule is placed in oscillating electric fields $\vec{E}_1 e^{-i\omega_1 t}$ and $\vec{E}_2 e^{-i\omega_2 t}$, its nonlinear response creates an oscillating dipole moment at $\omega_{\mathrm{SF}} = \omega_1 + \omega_2$. The strength of this response is characterized by the molecule's *hyperpolarizability* tensor, $\overleftrightarrow{\alpha}^{(2)}$:

$$\vec{p}^{(2)}(\omega_{\mathrm{SF}}) = \overleftrightarrow{\alpha}^{(2)} \vdots \vec{E}_1 \vec{E}_2$$

Using perturbation theory in the electric-dipole approximation [16],

$$\overleftrightarrow{\alpha}^{(2)}(\omega_1) = \overleftrightarrow{\alpha}^{(2),\mathrm{NR}} + \frac{\overleftrightarrow{A}}{\omega_1 - \omega_{vg} + i\Gamma_{vg}}$$

$$A_{j\ell m} \propto \langle g|\alpha^{(1)}_{j\ell}|v\rangle \langle v|p_m|g\rangle$$

where "NR" stands for nonresonant contributions, $\overleftrightarrow{\alpha}^{(1)}$ is ordinary Raman polarizability, $\vec{p}$ is the electric dipole moment operator, g and v are the ground and first-excited vibrational state (Fig. 1.3), ω_{vg} and Γ_{vg} are respectively the vibrational frequency and damping coefficient, j, ℓ, m are Cartesian indices, and all expressions are in laboratory coordinates. This expression clarifies several points:

- For a vibration to have an SFG signal, it must be *both* Raman *and* IR active. Therefore, for example, an individual water molecule can have an SFG response, but an individual centrosymmetric molecule cannot, as its Raman-active and IR-active modes are mutually exclusive. This restriction is just as expected in the electric dipole approximation, from the symmetry argument above.

- If a molecule's orientation is flipped, the Raman polarizability $\overleftrightarrow{\alpha}^{(1)}$ is unchanged, but the transition dipole moment operator $\vec{p}$ changes sign, and therefore so does $\overleftrightarrow{A}$. Consequently, the sign of $\overleftrightarrow{A}$ directly reflects the absolute orientations of molecules. This is a crucial point for our data interpretation (see Sec. 1.2.5).

- In a bulk system of randomly-oriented asymmetric molecules, such as bulk water, each individual molecule emits SFG, but the signals of molecules with opposite orientations

cancel each other out. So there is no net signal, again consistent with the expectation from symmetry.

For a bulk molecular system, each molecule is in a slightly different environment, and therefore has a slightly different vibrational frequency ω_{vg}. The expected *aggregate* SFG signal is:

$$\overleftrightarrow{\chi}^{(2)}(\omega_1) = \overleftrightarrow{\chi}^{(2)}_{\mathrm{NR}} + \int \frac{\overleftrightarrow{A}(\omega_{vg})}{\omega_1 - \omega_{vg} + i\Gamma(\omega_{vg})}\rho(\omega_{vg})d\omega_{vg} \tag{1.2}$$

where $\rho(\omega_{vg})d\omega_{vg}$ is the density of molecules with a particular vibrational frequency ω_{vg}, and $\overleftrightarrow{A}(\omega_{vg})$ is the average value for this group of molecules, which is zero if the molecules are pointing up and down equally, or has one sign or the other depending on which orientation dominates.

1.2.4 Motivation for phase-sensitive SFVS

After measuring $\overleftrightarrow{\chi}^{(2)}$, one wants to analyze Eq. (1.2) to learn about $A(\omega_{vg})$ and $\rho(\omega_{vg})$, i.e. what molecular environments are present and how the molecules in each environment are oriented. However, if the absolute value $|\overleftrightarrow{\chi}^{(2)}(\omega_1)|^2$ is known, but the complex phase is not, this analysis is quite difficult and error-prone [17]. In fact, even with no noise and just a few discrete lines, it is mathematically impossible to get a unique fit to Eq. (1.2) [18].

However, with the full complex function $\overleftrightarrow{\chi}^{(2)}(\omega_1)$ (both absolute value and phase), there is a simple and unambiguous approach to data analysis. In the case of the water surface, we can take $\Gamma \to 0$ in Eq. (1.2), because we expect inhomogeneous broadening to dominate over homogeneous. Therefore, taking the imaginary part of both sides and applying the Sokhotski-Plemelj theorem, Eq. (1.2) becomes:

$$\mathrm{Im}\,\overleftrightarrow{\chi}^{(2)}(\omega_1) = -\pi\overleftrightarrow{A}(\omega_1)\rho(\omega_1). \tag{1.3}$$

Thus, by checking the sign of $\mathrm{Im}\,\overleftrightarrow{\chi}^{(2)}(\omega_1)$, one can immediately see whether the group of molecules with vibration frequency equal to ω_1 tends to be oriented disproportionately towards or away from the surface. This will be the primary mode of data analysis as described in the following section.

1.2.5 Data interpretation for water solutions

A typical SFG intensity spectrum of water in the OH stretch region is shown in Fig. 1.4. The sharp peak on the right is called the "dangling OH" peak, which corresponds to OH bonds at the topmost surface with the hydrogen pointing towards vapor [17]. This study focuses on the broad, lower-energy signal associated with hydrogen-bonded OH's.

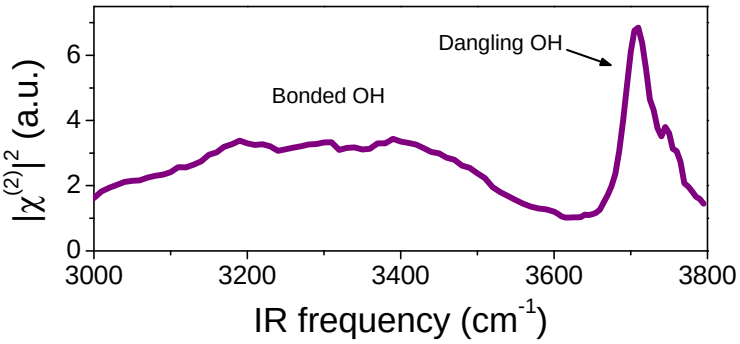

Figure 1.4 – Measured SFG intensity from neat water in the OH stretch region (smoothed and corrected for Fresnel factors.

The data analysis method for this work is summarized in Fig. 1.5. The relative surface affinity of the anion and cation will determine the sign of the surface electric field from the ionic double-layer. This electric field, in turn, will tend to reorient water molecules. From Eq. (1.3), $\mathrm{Im}\,\overleftrightarrow{\chi}^{(2)}$ in the bonded OH region will increase if the double-layer electric field tends to point towards the air, and will decrease if the double-layer electric field points away from the air. (We are assuming here, as usual, that $\mathrm{Im}\,\overleftrightarrow{\chi}^{(2)} > 0$ for the orientation where O→H points towards the air, as it does for the dangling OH peak. Strictly speaking, this assumption may or may not be correct, but it does not matter. In fact, $\overleftrightarrow{\chi}^{(2)}$ is measured with a sign ambiguity (see Sec. 1.3.4). The relationship between signal and absolute orientation is established instead by the dangling OH peak.)

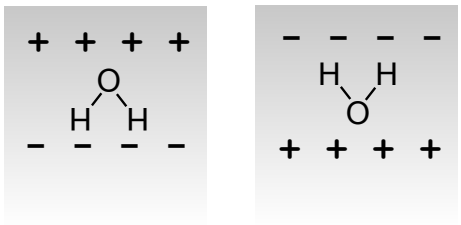

Figure 1.5 – Relation between surface affinity and SFG signal. Left: If the cation has higher surface affinity than the anion, then the electric field from this double-layer will tend to reorient water molecules into the orientation shown. Right: If the anion has higher surface affinity, the preferred water molecule orientation will be reversed. The left and right case can be distinguished by the sign of $\mathrm{Im}\,\overleftrightarrow{\chi}^{(2)}$ (Eq. (1.3)).

Therefore, by using different pairs of cations and anions, one can determine in each case whether the anion or cation is, on average, closer to the surface.

1.2.6 Fresnel factors and light polarizations

An SFG measurement does not capture $\overleftrightarrow{\chi}^{(2)}$ directly, but rather a related quantity called $\overleftrightarrow{\chi}^{(2)}_{\text{eff}}$, the *effective* surface susceptibility, which in this context differs from $\overleftrightarrow{\chi}^{(2)}$ solely due to the Fresnel factors that relate the fields within the high-index water to the corresponding fields in the adjacent air. The basic formula is [19]:

$$\vec{E}_B(\omega) = \overleftrightarrow{L}(\omega) \cdot \vec{E}_A(\omega) \tag{1.4}$$

where $\overleftrightarrow{L}$ is the tensorial Fresnel factor, $\vec{E}_A$ is the electric field on the air side of the interface, and $\vec{E}_B$ is the electric field on the water side. If the interface is normal to the z direction, and the light is incident in the $x-z$ plane, then the tensorial Fresnel factor is diagonal in the $x-y-z$ basis, and its nonzero components are [19]:

$$\begin{aligned} L_{XX}(\omega) &= \frac{2\epsilon_A(\omega)k_{Bz}(\omega)}{\epsilon_B(\omega)k_{Az}(\omega)+\epsilon_A(\omega)k_{Bz}(\omega)} \\ L_{YY}(\omega) &= \frac{2k_{Az}(\omega)}{k_{Az}(\omega)+k_{Bz}(\omega)} \\ L_{ZZ}(\omega) &= \frac{2\epsilon_A(\omega)\epsilon_B(\omega)k_{Az}(\omega)}{\epsilon_B(\omega)k_{Az}(\omega)+\epsilon_A(\omega)k_{Bz}(\omega)} \cdot \frac{1}{\epsilon'(\omega)} \end{aligned} \tag{1.5}$$

Here, ϵ' is the dielectric constant in the interfacial region between A and B (the top few monolayers), which is poorly defined, since dielectric properties are by definition averaged over a mesoscopic region. Nevertheless, a careful analysis relates the appropriate value of ϵ' to the local field factors, which can be roughly estimated in a simple slab model to give [19,20]:

$$\epsilon' = \frac{\epsilon_B(\epsilon_B+5)}{4\epsilon_B+2}. \tag{1.6}$$

Next, plugging in Eq. (1.4),

$$\chi^{(2)}_{\text{eff}} = [\overleftrightarrow{L}(\omega_{\text{SF}}) \cdot \hat{e}_{\text{SF}}] \cdot \overleftrightarrow{\chi}^{(2)} \vdots [\overleftrightarrow{L}(\omega_1) \cdot \hat{e}_1][\overleftrightarrow{L}(\omega_2) \cdot \hat{e}_2]$$

where $\hat{e}_i$ is the polarization vector of the corresponding wave on the air side.

All measurements in this study used "SSP" polarization. The designation "SSP" refers, respectively, to the light polarizations at ω_{SF}, ω_2, and ω_1 (infrared). SSP polarization is particularly useful for SFVS surface studies, although other polarizations give complementary information [15]. In the SSP polarization:

$$\chi^{(2)}_{\text{eff,ssp}} \propto \chi^{(2)}_{yyz} L_{YY}(\omega_{\text{SF}}) L_{YY}(\omega_2) L_{ZZ}(\omega_1)$$

(using the fact that $\chi^{(2)}_{yyx} = 0$ by symmetry.)

For this experiment, it is worth noting that $L_{ZZ}(\omega_1)$ is complex, as a result of water's complex dielectric constant at the OH stretch resonance. Therefore, the measured complex

phase for $\chi^{(2)}_{\text{eff}}$ is different than the inferred complex phase for $\chi^{(2)}$. This phase difference, however, is quite small—3.5° or less (Fig. 1.6)—so it does not noticeably affect the data, although it is taken into account anyway. (However, if $\epsilon' = \epsilon_B$ were used instead of Eq. (1.6), the phase correction would be as large as 20°, noticeably altering the data.)

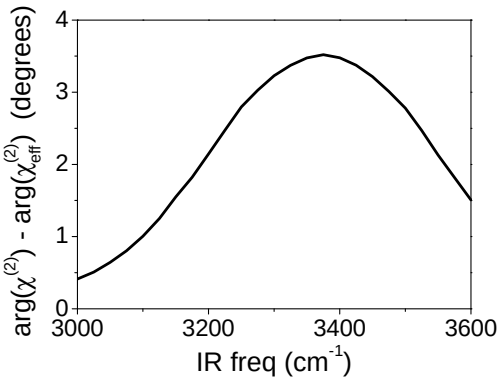

Figure 1.6 – Phase-shift due to complex Fresnel factors.

The absolute value of $L_{ZZ}(\omega_1)$ varies slightly with frequency, which is taken into account in the results. The L_{YY} factors are ignored altogether, being both real and frequency-independent. (The measurement of $\overleftrightarrow{\chi}^{(2)}$ is in arbitrary units anyway.)

1.3 Experimental methods

1.3.1 Generating the beams

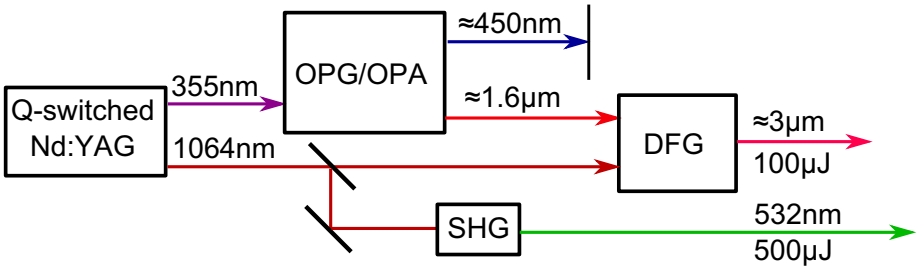

Figure 1.7 – Schematic of experimental setup for generating SFG input beams (ω_1 and ω_2). See text for description.

The two input beams for SFG are generated by a multi-step process, summarized in Fig. 1.7.

First, a commercial Q-switched Nd:YAG laser (EKSPLA Corporation) generates light with wavelength 1064nm, pulse length 20ps, and repetition 10Hz. Nonlinear crystals within

the laser convert some of the power into the third harmonic (355nm). The pulse power exiting the laser is 40mJ in the fundamental and 6mJ in the third harmonic.

Part of the 1064nm beam is directed into an SHG crystal. The resulting 532nm beam serves as the ω_2 input for SFG, with power about 500μJ per pulse.

Meanwhile, the 355nm beam is directed into an optical parametric generation and amplification (OPG/OPA) system [21]. By tuning the angles of the nonlinear crystals and of the bandwidth-narrowing grating, the signal and idler frequencies from the OPG/OPA can be varied. Afterwards, the pump beam and signal beam are filtered out, while the idler is combined with part of the 1064nm beam and directed into a DFG crystal ($LiNbO_3$). The resulting mid-infrared beam is separated with a germanium Brewster plate, and used as the ω_1 input for SFG, with power about 100μJ per pulse.

A delay line in the 1064nm beam ensures that it arrives at the DFG crystal simultaneously with the idler. Likewise, a delay line on the 532nm ω_2 beam ensures that it arrives at the sample simultaneously with the ω_1 beam.

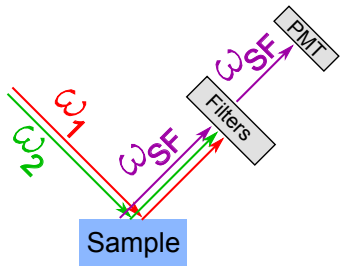

Figure 1.8 – Collinear geometry for SFG.

At the SFG setup, the ω_1 and ω_2 beams are separately focused to a point near the sample, with the focal length and focal point chosen to balance the desire for high signal strength with the need to avoid damaging the sample—or in the case of water, local boiling that would agitate the smooth surface. At the sample surface, the beam diameters are 300–600μm. Before reaching the surface, however, the beams are combined via a silicon Brewster plate into a collinear arrangement (see Fig. 1.8), with angle of incidence $\approx 45°$. This collinear arrangement is somewhat atypical for SFG, since it makes it more of a challenge to filter out the ω_1 and ω_2 beams from the 10^{15}-times-weaker ω_{SF} beam. Nevertheless, the collinear setup is used, because of the significant advantages for phase-sensitive measurements described below.

The "filters" shown in Fig. 1.8 consist of a series of glass and holographic filters, then a monochromator, then a photomultiplier tube (PMT) with low responsivity below ω_{SF}, and finally a boxcar integrator synchronized with the laser. The monochromator is particularly useful for filtering broadband fluorescence and scattering, while the boxcar is particularly important for eliminating room light. The effectiveness of this filtering system can be easily tested by separately blocking either ω_1 or ω_2, which reveals the background by eliminating

the $\omega_{\rm SF}$ signal. We found this way that we were, in fact, successful in achieving a very low background, typically one photon every fifty pulses with the room light on, or one photon every several hundred pulses with the room light off. This level was 10–100 times weaker than the water $\omega_{\rm SF}$ signal; nevertheless, to be safe, all water measurements were performed with the room light off. The quartz reference measurements were run at a much lower sensitivity, due to the stronger signal, and therefore the room lights could be left on.

Changing wavelength for SFG requires simultaneously rotating the crystals and grating in the OPG/OPA, the DFG crystal, and the monochromator knob, as well as a compensator after the DFG crystal to keep the beam location fixed. All this is achieved via calibrated computer-controlled stepper motors.

1.3.2 SFG measurement overview

The SFG intensity $|\overleftrightarrow{\chi}^{(2)}_{\rm eff}|$ and phase $\arg(\overleftrightarrow{\chi}^{(2)}_{\rm eff})$ were measured separately, on different days using different (albeit nominally-identical) samples, then combined mathematically into the complex $\overleftrightarrow{\chi}^{(2)}$. To ensure the robustness of this procedure, many measurements were repeated, often months apart, to ensure that the nominally-identical samples always did, in fact, have a consistent, repeatable SFG response.

The measurements primarily covered the range $3000\text{cm}^{-1} \leq \omega_1 \leq 3600\text{cm}^{-1}$. This range, called the "bonded OH" region (see Sec. 1.2.5), captures the data of interest in this study. For the intensity measurements, some more data out to 3800cm^{-1} was collected to capture the "dangling OH" peak for comparison with previous work [22–26]. The phase at these higher frequencies was not captured, as it is not controversial (expected to be a classic sharp resonance), and more importantly, not relevant for the study here. Finally, intensity data at lower frequencies $2700-3000\text{cm}^{-1}$ were occasionally taken in order to check for the presence of hydrocarbon contamination (which has a signal at the CH stretch frequency).

In the next sections we discuss, first the SFG intensity measurement procedure, then second, the SFG phase measurement procedure.

1.3.3 SFG intensity measurement

The intensity measurement uses the setup shown schematically in Fig. 1.8. The "sample" shown in the figure is first z-cut quartz, then the water sample, then z-cut quartz again. (The repeated measurement of quartz allows us to check that the laser intensity has not drifted over the course of the measurement.) For the quartz, the photomultiplier tube (PMT) was used in linear (current) mode with 800V bias; for water, it was used in photon-counting mode with 1200V bias. Under the reasonable assumption that $\overleftrightarrow{\chi}^{(2)}_{\rm quartz}$ is approximately constant over the wavelength range of interest (it is an off-resonance response), the value of $|\overleftrightarrow{\chi}^{(2)}_{\rm water\ sample}|^2$ is simply calculated as the ratio of the water sample measurement to the quartz sample measurement.

In photon-counting mode, it is always impossible to distinguish the arrival of one photon in a pulse from the arrival of two or more photons in a pulse (except insofar as the threshold for counting is not set properly). Therefore the standard correction is made based on Poissonian statistics:

$$\begin{pmatrix}\text{Average number of}\\ \text{photons per pulse}\end{pmatrix} = -\ln\left(1 - \begin{pmatrix}\text{Fraction of pulses with}\\ \text{at least one photon}\end{pmatrix}\right)$$

where "ln" is natural logarithm.

For each measurement, many wavelength sweeps are performed. The sweeps are superimposed to visually inspect for signs of laser instability or bad datapoints, then averaged for the final answer. Both the water and quartz curves are smoothed via a five-point rolling average, then their ratio is computed, and finally the Fresnel factors (Sec. 1.2.6) are divided out to get $(|\chi^{(2)}|^2)$ for the water sample.

One or two specific data-points, usually 3530cm^{-1} and/or 3540cm^{-1}, were consistently problematic outliers and had to be discarded. This artifact was discovered to arise from unintentional four-wave mixing at the sample, from light at frequency $(\omega_{\text{idler}} + \omega_2 - \omega_1)$. This should not occur because, first, the idler should have been blocked by the germanium plate from reaching the sample, and second, the four-wave-mixing light has the wrong wavelength to pass through the monochromator (except at $\omega_1 = 3133\text{cm}^{-1}$ when, by an unlucky coincidence, $\omega_{\text{idler}} + \omega_2 - \omega_1 = \omega_{\text{SF}}$). However, around 3530cm^{-1}, neither of these mechanisms was sufficiently effective to block the four-wave-mixing signal. (The monochromator slit had to be rather wide for technical reasons.) The discarding of these one or two data-points posed no problem for graphs or data interpretation, and other points were confirmed to have no such contamination.

1.3.4 SFG phase measurement

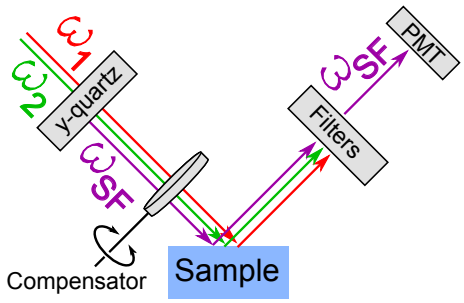

Figure 1.9 – Phase measurement setup.

The phase of $\overleftrightarrow{\chi}^{(2)}$ was measured by an interference method [27], with schematic shown in Fig. 1.9.

In this method, the SFG signal from a y-cut quartz reference plate is interfered with the SFG signal from the sample, which is either the real sample (water) or a reference sample (quartz). A glass compensator plate is placed between the two SFG sources. As it rotates, it shifts the relative phases of the two SFG sources, creating interference fringes. Comparing the interference fringe offset of the real sample (water) to the reference sample (quartz) enables the phase of $\overleftrightarrow{\chi}^{(2)}_{\text{water sample}}$ to be measured. In the following paragraphs we describe this in more detail.

The crucial quantity in this measurement is the combined phase difference between the three waves: $\phi_{\text{SF}} - \phi_1 - \phi_2$. At a given location, the quantity $\phi_{\text{SF}} - \phi_1 - \phi_2$ is fixed in time, due to the relation $\omega_{\text{SF}} - \omega_1 - \omega_2 = 0$. By similar logic, if the three waves are plane-waves traveling in the same direction through vacuum, the quantity $\phi_{\text{SF}} - \phi_1 - \phi_2$ is the same everywhere along the path. Finally, by Eq. (1.1), the quantity $\phi_{\text{SF}} - \phi_1 - \phi_2$ in an SFG-active sample equals the phase $\arg(\overleftrightarrow{\chi}^{(2)})$:

$$\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}}) = \arg(\vec{P}) - \arg(\vec{E}_1) - \arg(\vec{E}_2) = \phi_{\text{SF}} - \phi_1 - \phi_2.$$

By this principle, immediately after the quartz reference plate, the three waves—ω_1, ω_2, and the newly-generated ω_{SF}—have a fixed phase relation:

$$(\phi_{\text{SF}} - \phi_1 - \phi_2)_{\text{after y-quartz}} = \text{constant}.$$

The constant is solely a property of the y-quartz, and although its specific value is known in principle (either 0° or 180° depending on orientation), it is irrelevant to the measurement.

After the y-quartz, the three waves travel collinearly through air for several centimeters. As they do, their relative phase $(\phi_{\text{SF}} - \phi_1 - \phi_2)$ changes very gradually, due to the dispersion of air [28, 29]:

$$\frac{\Delta(\phi_{\text{SF}} - \phi_1 - \phi_2)}{\text{distance traveled through air}} = (n_{\text{SF,air}}\omega_{\text{SF}} - n_{1,\text{air}}\omega_1 - n_{2,\text{air}}\omega_2)/c \lesssim \frac{360^\circ}{10\text{cm}}. \tag{1.7}$$

Next, the three waves pass through the glass compensator plate, which is oriented at an angle θ from normal. As shown in Fig. 1.10, we can define ΔX as the lateral displacement of the beam, ΔY as the other orthogonal component of the distance traveled within the plate, d as the plate thickness, and α as the angle between the light's propagation direction within the plate and the plate's normal. From straightforward geometry and trigonometry, one can calculate:

$$\Delta X = d\,(\sin\theta - \cos\theta\tan\alpha)$$
$$\Delta Y = d\,(\cos\theta + \sin\theta\tan\alpha)$$

The phase shift, in radians, due of the plate, compared to propagating the same orthogonal distance in vacuum, is:

$$\Delta\phi = \frac{2\pi}{\lambda_{\text{vac}}}\left(n\sqrt{(\Delta X)^2 + (\Delta Y)^2} - \Delta Y\right) \tag{1.8}$$

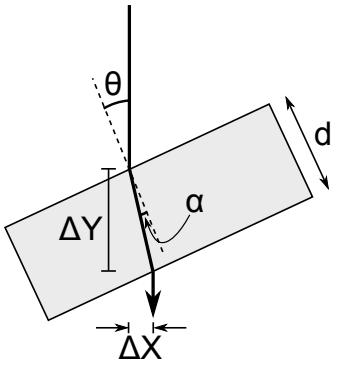

Figure 1.10 – Compensator plate (gray box) with light passing through (heavy line).

Doing a Taylor-expansion for small angles θ, substituting Snell's law, gives:

$$\Delta\phi \approx \frac{2\pi d}{\lambda_{\text{vac}}}\left[(n-1)+\frac{n-1}{2n}\theta^2\right] \tag{1.9}$$

The quantity of interest is the change in relative phase:

$$\Delta(\phi_{\text{SF}}-\phi_1-\phi_2) \approx (\text{constant}) - \pi d\left(\frac{1}{\lambda_{\text{SF,vac}}n_{\text{SF}}}-\frac{1}{\lambda_{1,\text{vac}}n_1}-\frac{1}{\lambda_{2,\text{vac}}n_2}\right)\theta^2$$

As seen below, the θ-dependence in this formula will be key to teasing out the sample's SFG phase.

As an example, we plug into this formula some typical refractive indices for a 1.5mm-thick fused silica plate ($n_1 \approx 1.42, n_2 \approx n_{\text{SF}} \approx 1.46$). This should be similar to the experimental parameters. The calculation result is:

$$\Delta(\phi_{\text{SF}}-\phi_1-\phi_2)_{1.5\text{mm SiO}_2} \approx (-36.7\times 360^\circ) + 0.96\times\theta^2$$

(in degrees). Therefore, when the plate is rotated $\approx 20^\circ$ off normal, the relative phases should shift by a full cycle, and by $\approx 30^\circ$, it should shift by a second cycle. This rough guess is remarkably consistent with the measured curves (Fig. 1.11); the measured value was 1.06 instead of 0.96. We also confirmed in this example that the second-order Taylor series (Eq. (1.9)) is indeed an excellent approximation for calculating $\Delta(\phi_{\text{SF}}-\phi_1-\phi_2)$ with $|\theta| < 30^\circ$. We also checked for this example whether rotating the plate might harm the beam overlap by shifting the different beams by different amounts ΔX. In fact, the relative shift is only about 15 μm at 30°, far less than the beam diameters, so this should not be a problem.

Finally, after exiting the compensator plate, the three beams travel together collinearly through air and arrive at the sample surface, now satisfying

$$\phi_{\text{SF}}-\phi_1-\phi_2 = C_1 + C_2\theta^2 \tag{1.10}$$

where C_1 and C_2 are constants. The measurement does *not* require that C_1 and C_2 be constant with respect to wavelength, only that they be the same (at a given wavelength) between consecutive measurements of the water sample and the quartz reference sample.

After the sample, there are two SF signals: The reflection of the y-quartz signal off the sample surface, with phase $\phi_{\text{SF,y-quartz}} = \phi_1 + \phi_2 + C_1' + C_2\theta^2$ (where $C_1' = C_1 + \pi$, due to the phase shift upon reflection); and the new SF signal from the sample, with phase $\phi_{\text{SF,sample}} = \phi_1 + \phi_2 + \arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}})$. The measured intensity of light is:

$$I_{\text{measured}} = \left|A_{\text{SF,y-quartz}} e^{i\phi_{\text{SF,y-quartz}}} + A_{\text{SF,sample}} e^{i\phi_{\text{SF,sample}}}\right|^2 \tag{1.11}$$

$$= K_1 + K_2 \cos(\phi_{\text{SF,y-quartz}} - \phi_{\text{SF,sample}}) \tag{1.12}$$

$$= K_1 + K_2 \cos(C_1' + C_2\theta^2 - \arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}})) \tag{1.13}$$

where K_1, K_2 are constants independent of θ.

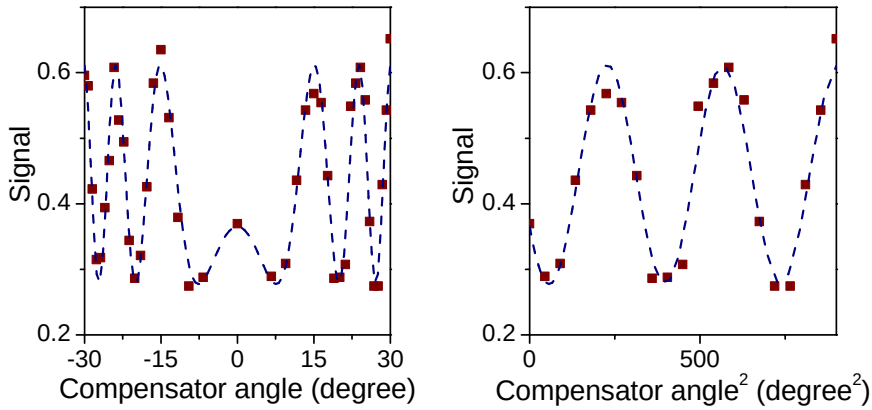

Figure 1.11 – Example of SFG signal as a function of compensator angle (left) and compensator angle squared (right). Points represent measurement data, and dashed line is the best fit.

Therefore, holding wavelength constant and varying θ, there should be interference fringes. An example is shown in Fig. 1.11. We do a least-squares fitting to the pattern and extract the phase at $\theta = 0$, which corresponds to $(C_1' - \arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}}))$. C_1' is separately measured by doing this procedure with the quartz reference sample, for which it is known that $\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}}) = \pm\pi/2$. (Quartz has an off-resonant, therefore real, bulk SFG susceptibility $\overleftrightarrow{\chi}^{(2)}_{B}$. This creates an equivalent surface susceptibility of $\overleftrightarrow{\chi}^{(2)}_{B}/(i\Delta k)$, which is purely imaginary.)

Finally, since C_1' is known, and $(C_1' - \arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}}))$ is measured for the water sample, $\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}})$ can be inferred.

To be more precise, $\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}})$ can be determined up to a possible offset of π. This offset comes from both uncertainty in the sign of C_2, and uncertainty in the sign of $\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff,z-quartz}}) =$

$\pm\pi/2$. Both could be resolved after some effort, but there is no need: A glance at the data makes the appropriate π phase-offset very clear.

Note that it is crucial that C_1' remain constant for the duration of the measurements of both the water sample and the quartz reference sample. This requirement is the key motivator of the collinear geometry: Even if the water level lowers slightly as it evaporates, or if the quartz and water are not at exactly the same height, it will have negligible effect on C_1', because of Eq. (1.7). (Any unintentional changes in the path length will be *much* less than 10cm!)

Another aspect of this measurement is the ideal orientation of the y-cut quartz plate. The symmetry of y-cut quartz means that, as the plate is rotated about its normal axis, its SFG signal oscillates between zero and very large. For this measurement, it was rotated to get the optimal visibility of the interference fringes, for both the weak water sample and the strong quartz reference sample. As in any interference measurement, if the intensity is too high, the fringes become drowned out by noise arising from laser fluctuations; if the intensity is too low, the fringes become drowned out by shot noise. After an appropriate orientation was found, it was kept fixed for the duration of all experiments.

1.3.5 Chemical

In a surface measurement, it is particularly important to avoid contamination, as even an extremely dilute impurity may become concentrated at the surface.

All glassware was cleaned with soap, acetone, and isopropanol, then soaked for at least a few days in 98% H_2SO_4 mixed with "Nochromix" (a proprietary glass cleaning agent). When removed, it was thoroughly rinsed in deionized water and blown dry with pure nitrogen gas passed through a filter.

Deionized water with resistivity 18.3 MΩ·cm was drawn from an EASYpure purifier system. Salts were purchased from Sigma-Aldrich; the list below shows the molar concentration into which they were mixed and their purity.

- 2M NaCl: 99.999% purity (trace metals basis)
- 2M KCl: 99.999% purity (trace metals basis)
- 2M $NaNO_3$: 99.995% purity (trace metals basis)
- 2M NH_4Cl: 99.998% purity (trace metals basis)
- 1M Na_2SO_4: 99.99% purity (trace metals basis)
- 1M $(NH_4)_2SO_4$: 99.999% purity (trace metals basis)

Prior to use, the samples of NaCl, KCl, and Na_2SO_4 were baked at 500°C to remove any residual organic contamination, while the ammonium salts (unstable at that temperature)

were treated in UV-ozone for 30–90 minutes for the same reason. The salts were mixed with water at the appropriate concentration in a glass petri dish. After it dissolved, it was transfered through a syringe filter into a second petri dish, in order to remove additional contaminant particles.

After the glass petri dish was placed into the light path, glass slides were positioned to mostly enclose the liquid, in order to minimize evaporation and contamination over the course of the ten-hours-or-so measurement. The light did not pass *through* the glass slides, but rather passed through a narrow slit between the slides.

1.4 Results and interpretations

Measurement results were taken over the course of six months, in the middle of which the entire laser system was thoroughly realigned. Therefore, it was encouraging that all results were consistently repeatable within the experimental error margins. A representative example is Fig. 1.12, three independent measurements of $\arg(\overleftrightarrow{\chi}^{(2)}_{\text{eff}})$ for neat water. These results are consistent with each other, and also with measurements by colleagues in prior years.

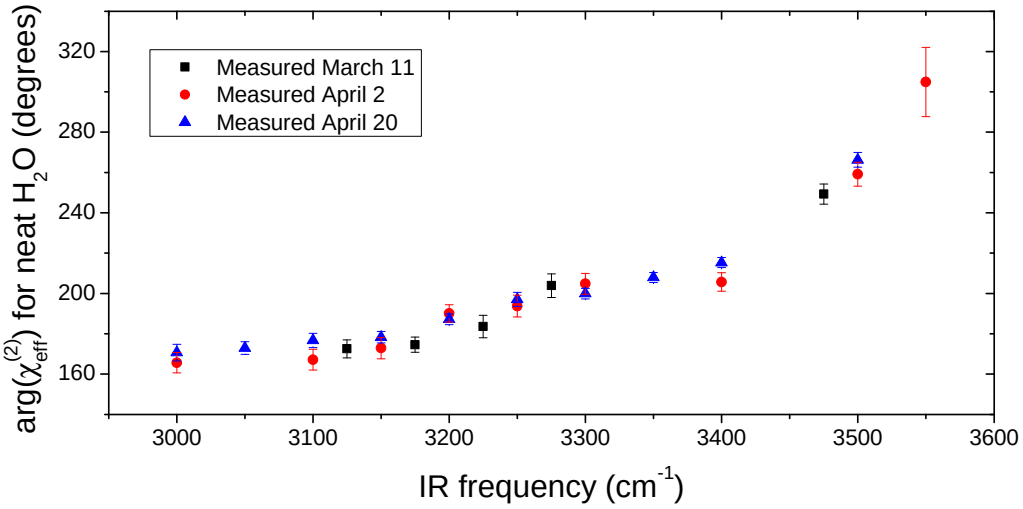

Figure 1.12 – Phase measurements showed good repeatability, as shown by these three separate phase measurements of neat water. These results were also consistent with measurements done in the lab in previous years (not shown). Error bars are uncertainty inferred from the least-squares fitting procedure.

All of the intensity measurements $|\chi^{(2)}|^2$ are shown in Fig. 1.13. All the raw phase measurements $\arg(\chi^{(2)}_{\text{eff}})$ are shown in Fig. 1.14. Combining these gives the $\text{Im}(\chi^{(2)}_{\text{eff}})$, Fig. 1.15. (NaI data from a previous study [30] is also plotted for comparison.)

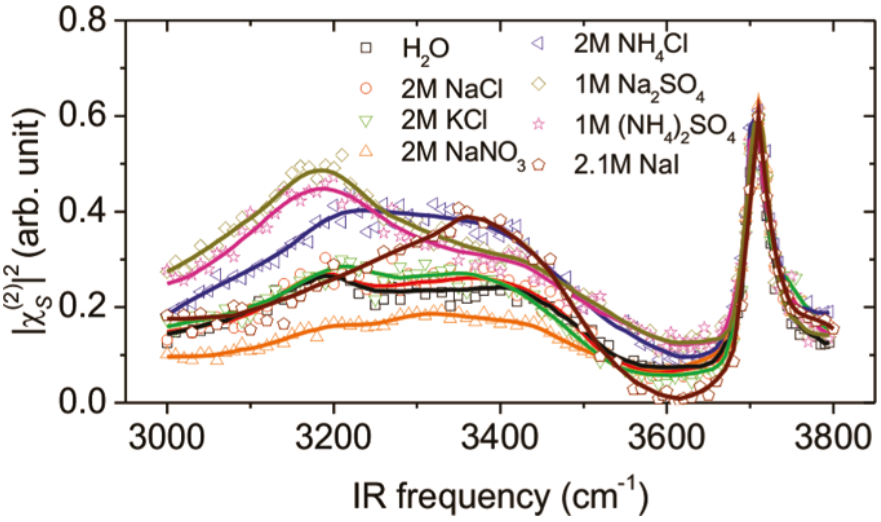

Figure 1.13 – All measured SFG intensities. Fresnel factor is already divided out.

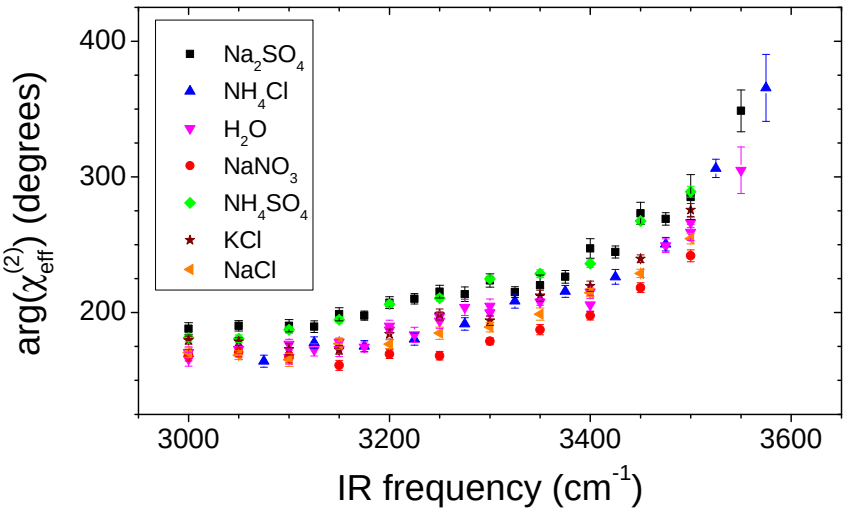

Figure 1.14 – All measured SFG phases.

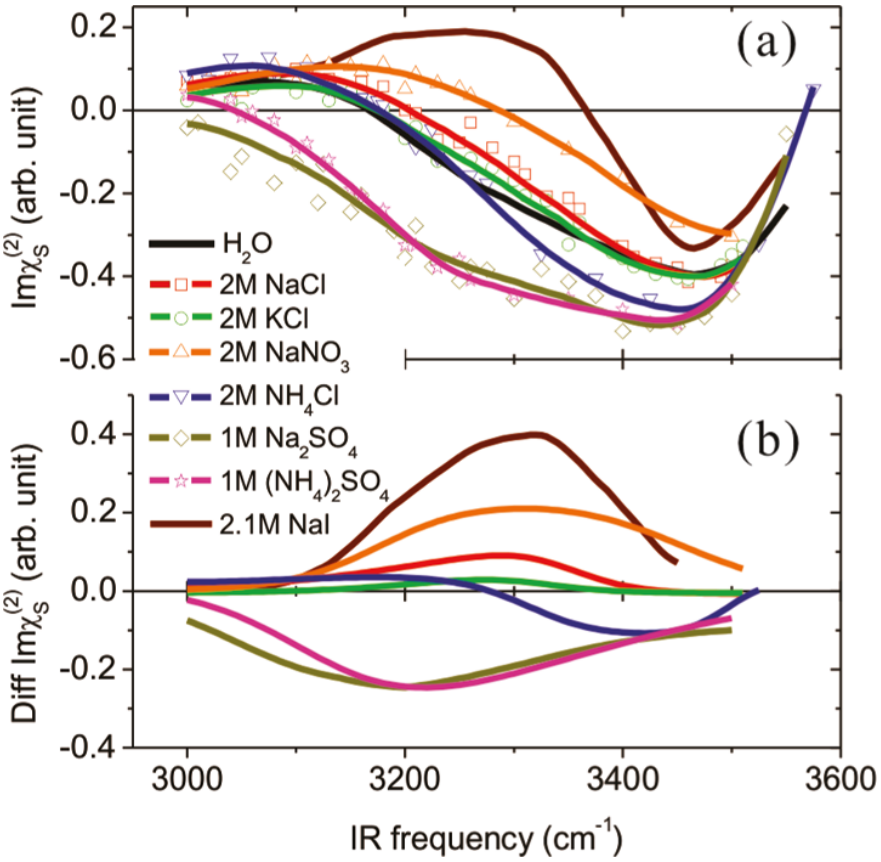

Figure 1.15 – Top: Imaginary part of SFG susceptibility. Bottom: Difference between each solution and neat water.

We interpret these data using the framework described in Sec. 1.2.5, where a change in $\mathrm{Im}\,\chi^{(2)}$ is assumed to be caused by the reorientation of water molecules due to the electric field from an ionic double-layer. However, before using this assumption, other possibilities should be ruled out. First, the ionic species might directly contribute to the $\mathrm{Im}\,\chi^{(2)}$ signal; however, this cannot occur because their vibrational modes are far off-resonance in this part of the spectrum. A possible exception is the umbrella-bending-rocking combination mode of NH_4 at $3060\mathrm{cm}^{-1}$ [31], which does not show clearly in the data, but cannot be ruled out either. Second, the ionic species may physically get in the way in the networking and bonding of water molecules at the topmost surface. However, this is unlikely to be a large effect because, first, the dangling-OH peak is not strongly effected by any of the solutes, and second, at the concentrations used, even the most surface-enhanced species measured, I^-, should have a surface concentration of only a few percent [32], much too small to account for the measured change in SFG signal. Therefore, we can confidently use the analysis framework described in Sec. 1.2.5.

For example, $\mathrm{Im}\,\chi^{(2)}$ is more positive with dissolved $NaNO_3$ than with neat water, and more negative with dissolved Na_2SO_4 than with neat water. Therefore, we infer that NO_3^- tends to have higher surface affinity than Na^+, which in turn tends to have higher surface affinity than SO_4^{2-}. Continuing in this manner, we get a rank ordering of surface affinity, from closest-to-the-surface to farthest-to-the-surface:

$$I^- > NO_3^- \gtrsim NH_4^+ > Cl^- \gtrsim K^+ \gtrsim Na^+ > SO_4^{2-} \tag{1.14}$$

The anion ordering in (1.14) agrees with the Hofmeister series, with later ions in the series being closer to the surface [33]. The cations were generally more similar to each other, but their trend appears reversed from the Hofmeister-series expectation. We emphasize that the ordering (1.14) conveys the average behavior, not every detailed aspect of the depth-dependent profile. Moreover, this type of analysis neglects the possibility of specific anion-cation interactions. In fact, there is some suggestion of these interactions: $(NH_4)_2SO_4$ is almost indistinguishable from $(Na)_2SO_4$, whereas NH_4Cl is clearly distinguished from NaCl. Nevertheless, we expect that the basic conclusions are robust. Next, we go through these species one-by-one to discuss the data in more detail, and compare with previous results and expectations.

The spectra of NaCl, KCl, and water are all quite similar. Previous measurements [22–26] could not confidently detect a difference at all, hence concluding that Na^+, K^+, and Cl^- have a similar propensity to accumulate at the topmost surface of water, and have similar concentration profiles as a function of depth. In this more refined $\mathrm{Im}\,\chi^{(2)}$ measurement, however, it is possible to detect a small increase in $\mathrm{Im}\,\chi^{(2)}$ resulting from the presence of NaCl or KCl. This is consistent with the molecular-dynamics (MD) simulation expectation that Na^+ and K^+ are depleted at the interface, while Cl^- is neither depleted nor enhanced.

Next we consider the SO_4^{2-} ion. Simulations have suggested that the ion should be strongly repelled from the surface, because its doubled charge gives it a quadrupled electric repulsion from the interface in the Onsager model [22]. Previous non-phase-resolved SFG measure-

ments have seemed to support this [25]. Our measurements confirm this more definitively: We found SO_4^{2-} to have the strongest surface repulsion of any species measured.

Next we consider the NH_4^+ ion, by comparing NH_4Cl to NaCl, and also by comparing $(NH_4)_2SO_4$ and Na_2SO_4. Consistent with expectations from Raman and IR intensities [22], as well as mode symmetry considerations, we see no clear evidence of a direct SFG signal from NH_4^+ ion vibrational modes, although we cannot rule out a small contribution from the umbrella-bending-rocking combination mode at 3060cm^{-1} [22]. Therefore we attribute the signal, as usual, to the OH bonds in water molecules reorienting due to ionic double-layer fields. Below 3250cm^{-1}, there is no experimentally-resolved difference between NH_4Cl and NaCl, or between $(NH_4)_2SO_4$ and Na_2SO_4. Between 3250–3500cm^{-1}, however, the NH_4Cl solution has significantly more negative $\mathrm{Im}\,\chi^{(2)}$ than NaCl. The sign of this change corresponds to NH_4^+ being on average closer to the surface than Na^+. This trend matches expectations from MD [22], but while MD finds a Cl^- depth profile more similar to NH_4^+ than Na^+, the experimental results are the opposite [1, 22]. Compared with the chloride salts, the sulfate salts $(NH_4)_2SO_4$ and Na_2SO_4 show a more similar spectrum, with the ammonium salt spectrum only slightly more negative than the sodium one. This may be an anion-cation interaction, where perhaps the attraction between the deeper sulfate anion and shallower cation limits the separation between the two. Such anion-cation interactions are seen in MD simulations of these solutions [22], but not in macroscopic surface-tension measurements [9].

Finally, we consider the NO_3^- ion. From the $NaNO_3$ measurement, the NO_3^- ions appear closer to the surface than the Na^+ counter-ions, but not as close as I^-. This result is a helpful contribution to the literature, as the surface affinity of nitrate has been quite controversial. Earlier MD simulations predicted a surface excess of NO_3^- [34], but more recent results, with a different model for polarizability, predicted a surface repulsion [31, 35, 36]. Experimental results have also been somewhat inconsistent. X-ray photoemission spectroscopy found not much surface excess of NO_3^- [12] in a 3M $NaNO_3$ solution, while UV-SHG found that the surface excess of NO_3^- in a 2M $NaNO_3$ solution was appreciable but not as strong as the surface excess of I^- [33] (consistent with the results reported here). On the other hand, macroscopic surface-tension results suggest that NO_3^- had neither a surface excess nor deficit [9].

In summary, we have measured the complex $\chi^{(2)}$ spectrum of water with various salts, and shown that $\mathrm{Im}\,\chi^{(2)}$ gives valuable semi-quantitative information about the surface fields and ionic surface affinities. This data can help refine models of atmospheric chemistry, and constrain and validate molecular simulations of the surfaces. In the next chapter, we will discuss how SFG predictions are extracted from these molecular simulations, an area which is crucial for enabling better simulation-based interpretations of experimental spectra, and at the same time, better experiment-based validation of simulation models, leading ultimately to a better understanding of the water surface.

1.5 References

[1] Jungwirth, P., Finlayson-Pitts, B. J., and Tobias, D. J. *Chem. Rev.* **106**, 1137–1139 (2006).

[2] Netz, R. R. and Horinek, D. *Annu. Rev. Phys. Chem.* **63**, 401–418 (2012).

[3] Knipping, E. M., Lakin, M. J., Foster, K. L., Jungwirth, P., Tobias, D. J., Gerber, R. B., Dabdub, D., and Finlayson-Pitts, B. J. *Science* **288**, 301 (2000).

[4] Finlayson-Pitts, B. J. *Chem. Rev.* **103**, 4801 (2003).

[5] Wingen, L. M., Moskun, A. C., Johnson, S. N., Thomas, J. L., Roeselova, M., Tobias, D. J., Kleinman, M. T., and Finlayson-Pitts, B. J. *Phys. Chem. Chem. Phys.* **10**, 5668 (2008).

[6] Onsager, L. and Samaras, N. N. T. *J. Chem. Phys.* **2**, 528–536 (1934).

[7] Petersen, P. B. and Saykally, R. J. *Annu. Rev. Phys. Chem.* **57**, 333 (2006).

[8] Noah-Vanhoucke, J. and Geissler, P. L. *Proc. Natl. Acad. Sci. USA* **106**, 15125–15130 (2009).

[9] Pegram, L. M. and Record, M. T. *Proc. Natl. Acad. Sci. USA* **103**, 14278 (2006).

[10] Yan, E. C. Y., Liu, Y., and Eisenthal, K. B. *J. Phys. Chem. B* **102**, 6331–6336 (1998).

[11] Creux, P., Lachaise, J., Graciaa, A., Beattie, J. K., and Djerdjev, A. M. *J. Phys. Chem. B* **113**, 14146–14150 (2009).

[12] Brown, M. A., Winter, B., Faubel, M., and Hemminger, J. C. *J. Am. Chem. Soc.* **131**, 8354 (2009).

[13] Shen, Y. R. *The Principles of Nonlinear Optics*. John Wiley & Sons, Inc., Hoboken, (1984).

[14] Guyot-Sionnest, P. and Shen, Y. R. *Phys. Rev. B* **35**, 4420 (1987).

[15] Wei, X., Hong, S., Lvovsky, A. I., Held, H., and Shen, Y. R. *J. Phys. Chem. B* **104**, 3349 (2000).

[16] Hirose, C., Yamamoto, H., Akamatsu, N., and Domen, K. *J. Phys. Chem.* **97**, 10064–10069 (1993).

[17] Tian, C. S. and Shen, Y. R. *Chem. Phys. Lett.* **470**, 1 (2009).

[18] Busson, B. and Tadjeddine, A. *J. Phys. Chem. C* **113**, 21895 (2009).

[19] Wei, X. *Sum-Frequency Spectroscopic Studies: I. Surface Melting of Ice, II. Surface Alignment of Polymers*. PhD thesis, University of California, Berkeley, (2000).

[20] Zhuang, X., Miranda, P. B., Kim, D., and Shen, Y. R. *Phys. Rev. B* **59**, 12632 (1999).

[21] Zhang, J. Y., Huang, J. Y., Shen, Y. R., and Chen, C. *J. Opt. Soc. Am. B* **10**, 1758–1764 (1993).

[22] Gopalakrishnan, S., Jungwirth, P., Tobias, D. J., and Allen, H. C. *J. Phys. Chem. B* **109**, 8861 (2005).

[23] Raymond, E. A. and Richmond, G. L. *J. Phys. Chem. B* **108**, 5051 (2004).

[24] Mucha, M., Frigato, T., Levering, L. M., Allen, H. C., Tobias, D. J., Dang, L. X., and Jungwirth, P. *J. Phys. Chem. B* **109**, 7617 (2005).

[25] Schnitzer, C., Baldelli, S., and Shultz, M. J. *J. Phys. Chem. B* **104**, 585 (2000).

[26] Liu, D., Ma, G., Levering, L. M., and Allen, H. C. *J. Phys. Chem. B* **108**, 2252 (2004).

[27] Ji, N., Ostroverkhov, V., Chen, C. Y., and Shen, Y. R. *J. Am. Chem. Soc.* **129**, 10056 (2007).

[28] Ciddor, P. E. *Appl. Optics* **35**, 1566–1573 (1996).

[29] Mathar, R. J. *J. Opt. A - Pure Appl. Op.* **9**, 470–476 (2007).

[30] Tian, C., Ji, N., Waychunas, G. A., and Shen, Y. R. *J. Am. Chem. Soc.* **130**, 13033 (2008).

[31] Miller, Y., Thomas, J. L., Kemp, D. D., Finlayson-Pitts, B. J., Gordon, M. S., Tobias, D. J., and Gerber, R. B. *J. Phys. Chem. A* **113**, 12805 (2009).

[32] Ishiyama, T. and Morita, A. *J. Phys. Chem. C* **111**, 721–737 (2006).

[33] Otten, D. E., Petersen, P. B., and Saykally, R. J. *Chem. Phys. Lett.* **449**, 261 (2007).

[34] Salvador, P., Curtis, J. E., Tobias, D. J., and Jungwirth, P. *Phys. Chem. Chem. Phys.* **5**, 3752 (2003).

[35] Dang, L. X., Chang, T., Roeselova, M., Garrett, B. C., and Tobias, D. J. *J. Chem. Phys.* **124**, 066101 (2006).

[36] Thomas, J. L., Roeselova, M., Dang, L. X., and Tobias, D. J. *J. Phys. Chem. A* **111**, 3091 (2007).

2 Addressing ambiguities in sum-frequency-generation predictions from molecular simulations

2.1 Introduction

2.1.1 Background

As described in Chapter 1, sum-frequency generation (SFG), including its special case of second-harmonic generation (SHG), has been established as a powerful technique for studying surfaces and interfaces because of its ability to provide surface-specific electronic and vibrational spectra [1,2]. The technique has already been widely used to probe systems with fundamental importance in physics, chemistry, biology, and geology. In some cases, however, interpretation of the spectra can be difficult. This is particularly true for surface vibrational spectra of liquids, such as water, where the diversity of molecular arrangements leads to a broad, but featured, spectrum (Fig. 1.4). To understand such a spectrum and hence be able to deduce structural information about the surface or interface, we would need a theoretical calculation that can reproduce the experimental spectrum. Agreement between theory and experiment would lend weight to calculations which can then be further extended to predict new properties or phenomena about the surface. So far, molecular simulations have been the only theoretical technique used to calculate SF vibrational spectra of liquid interfaces, in particular the water interfaces because of their importance. However, their success in reproducing experimental spectra has been limited. Even in the case of the neat water/vapor interface, the calculated spectrum in the OH stretch range often does not fully agree with the experimental one, especially on the low-frequency side [3]. (However, more recent studies claim to have reconciled the low-frequency discrepancies [4, 5].) Different groups have also reported somewhat different calculated spectra [3].

Recently, Noah-Vanhoucke et al. found that different seemingly-valid ways to perform the calculation could yield very different spectra [6]. Therefore it is appropriate to carefully reexamine the approach and assumptions used in the simulations. In this article, we show existing deficiencies in current computational approaches and suggest ways to minimize them. We shall focus on the neat water/vapor interface as a representative example, but our discussion is generally applicable to all interfaces.

Let us first briefly review the basic theory of reflected SFG from an interface (with more details presented in Sec. 2.3). Consider two input beams at frequencies ω_1 and ω_2 overlapping on the interface that generates a SF output at frequency $\omega_{\mathrm{SF}} \equiv \omega_1 + \omega_2$ in the reflected direction (Fig. 2.1). The SF signal can be written as [1]:

$$S(\hat{e}_{\mathrm{SF}}, \hat{e}_1, \hat{e}_2) \propto \left| \hat{e}_{\mathrm{SF}} \cdot \overleftrightarrow{\chi}^{(2)}_{S,\mathrm{eff}} \vdots \hat{e}_1 \hat{e}_2 \right|^2, \quad \overleftrightarrow{\chi}^{(2)}_{S,\mathrm{eff}} \equiv \overleftrightarrow{\chi}^{(2)}_{S} + \frac{\overleftrightarrow{\chi}^{(2)}_{B}}{-i \left| \Delta \vec{k} \right|} \tag{2.1}$$

where $\overleftrightarrow{\chi}^{(2)}_{S}$ and $\overleftrightarrow{\chi}^{(2)}_{B}$ are defined as the second-order surface and bulk nonlinear susceptibil-

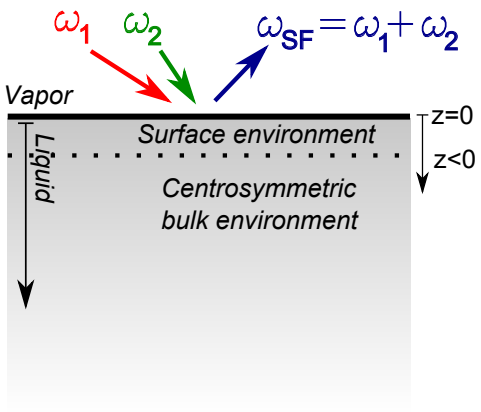

Figure 2.1 – A schematic of the water/vapor interface probed by an SFG experiment. The water is depicted as infinitely deep, which corresponds to an experiment where light attenuation or spatial filtering suppresses the signal of the opposite (bottom) interface.

ities, respectively, $\hat{e}_i$ is the polarization unit vector of the ith field, $\Delta\vec{k} \equiv \vec{k}_{\mathrm{SF}} - \vec{k}_1 - \vec{k}_2$ is the wave vector mismatch between input and output beams, and $\overleftrightarrow{\chi}^{(2)}_{S,\mathrm{eff}}$ is the net "effective" surface susceptibility (note that it is defined differently than in Chapter 1). The boundary between "surface" and "bulk" in this context is wherever the perturbing effects of the surface can no longer be felt (Fig. 2.1), for example a few monolayers for neat water, or a screening length for ionic solutions. In centrosymmetric media, $\overleftrightarrow{\chi}^{(2)}_B$ vanishes under the electric-dipole approximation, and is therefore dominated by the electric-quadrupole and magnetic-dipole response. On the other hand, $\overleftrightarrow{\chi}^{(2)}_S$ is dominated by the electric-dipole contribution because of the broken inversion symmetry at the interface. Its resonant spectrum, especially the vibrational one, can provide information about the interfacial structure as well as the orientation of the interfacial molecules. In some cases, the bulk term may be negligible compared with the surface term. In general, however, this is not necessarily true [1], and there is no simple theory that can be used to predict whether the bulk term can be neglected or not for a given interface. An order-of-magnitude estimate yields that the ratio of the bulk to the surface term is equal to the ratio of the geometric dimension of the induced electric-quadrupole on individual molecules to the average distance between molecules along the surface normal [7]. Thus, roughly speaking, if molecules are polar, electric-quadrupoles are well localized on the molecules and the surface layer thickness is large compared to a chromophore, the bulk term can be significantly smaller than the surface term. Experimentally, it is generally impossible to separate surface and bulk contributions [1, 8–10]. We usually resort to the observed sensitivity of the spectrum to surface perturbations, for example by adsorbed molecules, to judge whether the surface term dominates or not. (However, as discussed in Ref. [2], the experimental separation is possible if "surface" and "bulk" are defined in a more specific way, going beyond just Eq. (2.1).)

One contribution to the SFG signal comes from the electric-quadrupole (and magnetic-

dipole) contribution of interfacial molecules due to the strong field gradient at the interface between two media of different refractive indices [9]. Because this contribution is not relevant to the study here, we shall neglect it in our discussion.

2.2 Overview of molecular simulation ambiguities (Fundamental cause)

Consider now the molecular simulation of SFG from surfaces. Conventional approaches were recently discovered to be ambiguous and ill-defined [6]. We will show that this problem arises from the neglect of electric-quadrupole and magnetic-dipole contributions, i.e., the calculations implicitly assume $\overleftrightarrow{\chi}^{(2)}_B = 0$ and $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}} = \overleftrightarrow{\chi}^{(2)}_S$. However, it is deeply problematic to ignore $\overleftrightarrow{\chi}^{(2)}_B$, not merely because it may lead to inaccuracies, but more fundamentally because it makes the whole calculation ill-defined. While $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ of Eq. (2.1) is well-defined, $\overleftrightarrow{\chi}^{(2)}_S$ on its own and $\overleftrightarrow{\chi}^{(2)}_B$ on its own are not. Therefore, any attempt to calculate $\overleftrightarrow{\chi}^{(2)}_S$ while ignoring $\overleftrightarrow{\chi}^{(2)}_B$ will necessarily yield an incorrect and ambiguous result. The ambiguities in defining $\overleftrightarrow{\chi}^{(2)}_S$ and $\overleftrightarrow{\chi}^{(2)}_B$ are the main subject of this paper, and in Sec. 2.3, we will explain and quantify their origins and their consequences.

These ambiguities stem from the usual ambiguity in multipole-expansion: the electromagnetic response of a system can be correctly described using different forms of multipole expansion [8, 11–13]. The ambiguity in multipole expansions appears in many areas of physics—perhaps the most famous example is the fact that the dipole moment of a charged molecule is coordinate-system-dependent. Similarly, the electric-quadrupole moment of a polar molecule is also coordinate-system-dependent. In the case of SFG, different multipole expansions give different weights to the dipole response $\overleftrightarrow{\chi}^{(2)}_S$ versus the quadrupole response $\overleftrightarrow{\chi}^{(2)}_B$, but their combined response $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ is constant, as we shall show later.

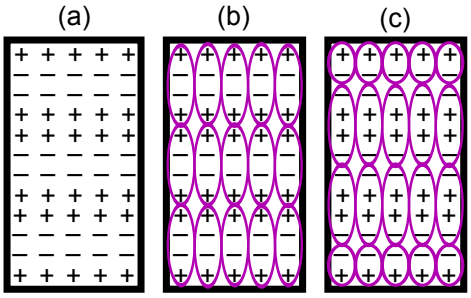

Figure 2.2 – An example of the well-known ambiguity in multipole expansions: (a) A distribution of point charges in a box. (b) By grouping these charges as shown, the box appears to have zero surface dipole and positive bulk quadrupole. (c) By grouping these charges differently, the box now appears to have positive surface dipole and negative bulk quadrupole. (After Ref. [8].)

In Fig. 2.2, we illustrate this classical ambiguity in multipole expansions with a simple example [8]: The same charge-distribution of a charge-neutral system (Fig. 2.2a) can be described as having zero surface dipole and positive bulk quadrupole density (Fig. 2.2b), or positive surface dipole and negative bulk quadrupole density (Fig. 2.2c). These are two equally-valid descriptions of the same system [11–13].

(As shown in Ref. [2], if the calculation is set up in a slightly different way, there arises a way to split the surface and bulk which is preferred as being the most natural, canonical, and physically-motivated split. However, for the purpose of this chapter, the important point is that the surface and bulk *can* be split in infinitely many different ways, all of which will add up to the same correct total signal.)

2.2.1 Overview of molecular simulation ambiguities (Specific cause)

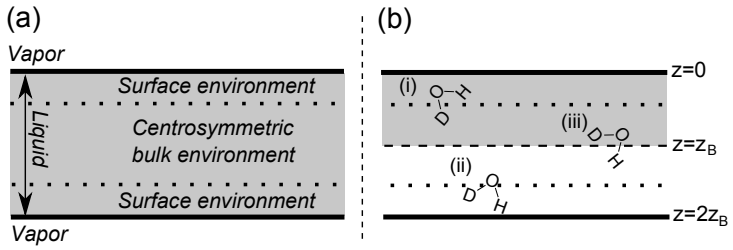

Figure 2.3 – (a) A schematic of a simulated liquid slab, with vapor on both sides. (b) The SFG signal is typically calculated by neglecting molecules below the artificial boundary (long dashed line). Therefore the signal from molecule (i) would be included in the total, (ii) would not, and (iii) might or might not, depending on the molecular center used.

As explained above, from a fundamental and general perspective, previous molecular-simulation calculations of SFG have been ambiguous because they calculated the ill-defined quantity $\overleftrightarrow{\chi}^{(2)}_S$, rather than the unambiguous quantity $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$. Different multipole expansion schemes (e.g., Fig. 2.2) in simulation yield different values of $\overleftrightarrow{\chi}^{(2)}_S$. Such arbitrariness is often hidden deep within subtle details of the algorithms. We focus on the common case of molecular simulations in a slab geometry: Here, the arbitrary part of the algorithm was recently discovered by Noah-Vanhoucke et al. [6].

In the slab geometry, the real system is approximated by a thin slab with periodic boundary conditions (Fig. 2.3a). In calculating many surface properties, such as surface tension, contributions from the two surfaces of the slab can be added and averaged, but this step is clearly not appropriate for SFG. The two surfaces, being opposite in orientation, will have their electric-dipole contributions canceled by each other. Therefore, in molecular simulations of SFG, it is necessary to artificially break the symmetry of the slab. Typically, an artificial boundary surface is set up at the middle of the slab, and only electric-dipole responses of molecules between the top surface of the slab and the artificial boundary surface

are assumed to contribute to SFG (Fig. 2.3b). This way, it was thought, the SFG signal from just the top surface would be calculated. (The artificial boundary surface is not the only symmetry-breaking method, but it is the simplest, and other methods [14–16] are only superficially different as discussed further below.)

Since molecules residing around the artificial boundary surface are randomly oriented in an isotropic bulk-like environment, the detailed implementation of the artificial separation might seem unimportant. Unfortunately, this expectation is not true, as was highlighted by Noah-Vanhoucke et al. [6]. The problem arises because of an ambiguity in assigning molecules that straddle the artificial boundary to one interface or the other. In the simplest scheme, the assignment is based on whether a predetermined point within the molecule is above or below the artificial boundary. In using a single point to represent the molecule's position, one thus establishes a "molecular center". In the case of an HOD molecule, this point could be placed on the H, O, or D atom, or any other fixed site in the molecular frame (Fig. 2.3b). Surprisingly, this choice can have a significant effect on the calculation. (A superficially similar ambiguity arises in DC electrostatic potential calculations [17, 18].)

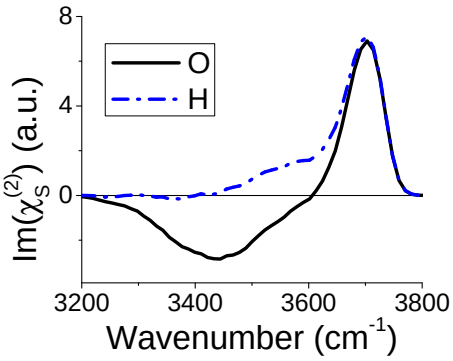

Figure 2.4 – Slab-based MD-calculated SFG spectra of water, with oxygen (O) or hydrogen (H) as the molecular center. Light polarization is SSP.

To illustrate the importance of this issue, we use the neat water/vapor interface as an example. Fig. 2.4 shows two spectra of $\mathrm{Im}\,\overleftrightarrow{\chi}_S^{(2)}$, calculated using two different choices of molecular center (these simulations are discussed in more detail below). Taking the O atom as center yields a qualitatively different result than taking instead the H as the center. It is not clear which one (if either) gives the "correct" spectrum, i.e. the one that should be compared with experiment. (The sharp peak at $\approx 3700\mathrm{cm}^{-1}$ is the same in the two cases, because it is associated with a surface mode—dangling OH vibration—that does not exist near the artificial boundary surface [3].)

Actually, the correct spectrum must come from calculation of $\overleftrightarrow{\chi}_{S,\mathrm{eff}}^{(2)}$, which includes both $\overleftrightarrow{\chi}_S^{(2)}$ and $\overleftrightarrow{\chi}_B^{(2)}$. As we shall show in Sec. 2.3, both the $\overleftrightarrow{\chi}_S^{(2)}$ and $\overleftrightarrow{\chi}_B^{(2)}$ spectra depend on the choice of the molecular center, but that of $\overleftrightarrow{\chi}_{S,\mathrm{eff}}^{(2)}$ does not. With current simulation approaches, it can be difficult to calculate $\overleftrightarrow{\chi}_B^{(2)}$. One therefore hopes that the $\overleftrightarrow{\chi}_B^{(2)}$ contribution can be

negligible in comparison with $\overleftrightarrow{\chi}_S^{(2)}$. For this to be true, we must have a sufficiently large $\overleftrightarrow{\chi}_S^{(2)}$, such as in the cases of a surface layer of polar molecules with a significant net polar orientation, and quite importantly, a choice of molecular center that minimizes the absolute value of $\overleftrightarrow{\chi}_B^{(2)}$. For our example in Sec. 2.5, an SFG reflection measurement of the air/water interface, we shall suggest that the most appropriate molecular center is at the oxygen atom for SSP polarization (denoting S-, S-, and P-polarizations for the fields at $\omega_{\rm SF}$, visible ω_2, and infrared ω_1 respectively) and SPS polarization, but slightly displaced towards the hydrogen atom for PSS polarization.

The paper is organized as follows: Sec. 2.3 describes the basics of sum-frequency generation. We first discuss SF response from individual molecules in terms of multipole expansion and show that the division into terms of electric-dipole, electric-quadrupole, and so on is not unique, but depends on the choice of molecular center. We then show the same ambiguity arises in describing nonlinear susceptibilities, $\overleftrightarrow{\chi}_S^{(2)}$ and $\overleftrightarrow{\chi}_B^{(2)}$, but that the effective surface susceptibility, $\overleftrightarrow{\chi}_{S,\rm eff}^{(2)}$, which takes into account both surface and bulk contributions, is independent of the molecular center. Sec. 2.4 discusses how such ambiguities will lead to different SF vibrational spectra calculated by molecular simulation using different choices of molecular center. Section 2.5 presents molecular dynamics simulations of the air/water interface as an example, and shows how the effect of $\overleftrightarrow{\chi}_B^{(2)}$ can be minimized by a proper choice of molecular center.

2.3 Basics of Sum-Frequency Generation

We describe in this section the basics of SF response of an interfacial system and show that $\overleftrightarrow{\chi}_{S,\rm eff}^{(2)}$ of Eq. (2.1) does not depend on the choice of molecular center, but $\overleftrightarrow{\chi}_S^{(2)}$ and $\overleftrightarrow{\chi}_B^{(2)}$ do. We start from SF responses of individual molecules, and show that the sum of the electric-dipole and electric-quadrupole responses of an individual molecule is independent of the choice of molecular center, as any observable property should be.

2.3.1 SF response from individual molecules

The SF response of an individual molecule in the incoming fields $\vec{E}_1(\omega_1, \vec{k}_1)$ and $\vec{E}_2(\omega_2, \vec{k}_2)$ can be characterized by its effective electric dipole $\vec{p}_{\rm eff}^{(2)}$, defined as the electric-dipole moment which, by itself, would give the same electromagnetic signal in the direction $\vec{k}_{\rm SF}$ as the combined effects of its true electric-dipole and higher-order moments together. The effective dipole is given in the form of a multipole expansion (i.e., power series in the wavevectors $\vec{k}_1$,

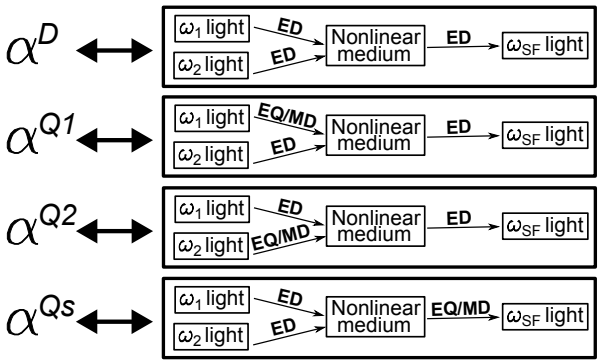

Figure 2.5 – The four lowest-order contributions to SFG susceptibility in the multipole expansion. The light-matter interaction is specified by "ED" for "Electric Dipole", or "EQ/MD" for "Electric Quadrupole or Magnetic Dipole".

$\vec{k}_2$, $\vec{k}_{\rm SF}$) as [1, 2]:

$$\begin{aligned}
\vec{p}^{(2)}_{\rm eff}(\omega_{\rm SF},\vec{k}_{\rm SF}) &= \vec{p}^{(2)}(\omega_{\rm SF},\vec{k}_{\rm SF}) - i\vec{k}_{\rm SF}\cdot\overleftrightarrow{q}^{(2)}(\omega_{\rm SF},\vec{k}_{\rm SF}) - \frac{c}{\omega_{\rm SF}}\vec{k}_{\rm SF}\times\vec{\mu}^{(2)}(\omega_{\rm SF},\vec{k}_{\rm SF}) + \cdots \\
\vec{p}^{(2)} &= \overleftrightarrow{\alpha}^{D}\!:\vec{E}_1\vec{E}_2 + \overleftrightarrow{\alpha}^{Q1,\rm EQ}\!:(\nabla\vec{E}_1)\vec{E}_2 + \overleftrightarrow{\alpha}^{Q2,\rm EQ}\!:\vec{E}_1(\nabla\vec{E}_2) + \\
&\quad + \overleftrightarrow{\alpha}^{Q1,\rm MD}\!:\vec{B}_1\vec{E}_2 + \overleftrightarrow{\alpha}^{Q2,\rm MD}\!:\vec{E}_1\vec{B}_2 \\
&= \overleftrightarrow{\alpha}^{D}\!:\vec{E}_1\vec{E}_2 + i\overleftrightarrow{\alpha}^{Q1}\!:\vec{E}_1\vec{E}_2\vec{k}_1 + i\overleftrightarrow{\alpha}^{Q2}\!:\vec{E}_1\vec{E}_2\vec{k}_2 \qquad (2.2) \\
\overleftrightarrow{q}^{(2)} &= \overleftrightarrow{\alpha}^{Qs,\rm EQ}\!:\vec{E}_1\vec{E}_2 \\
\vec{\mu}^{(2)} &= \overleftrightarrow{\alpha}^{Qs,\rm MD}\!:\vec{E}_1\vec{E}_2
\end{aligned}$$

where $\vec{p}^{(2)}$, $\overleftrightarrow{q}^{(2)}$, and $\vec{\mu}^{(2)}$ are the induced electric dipole, electric quadrupole, and magnetic dipole at the sum frequency, respectively; $\overleftrightarrow{\alpha}^{D}$, $\overleftrightarrow{\alpha}^{Qi,\rm EQ}$, $\overleftrightarrow{\alpha}^{Qi,\rm MD}$ are corresponding nonlinear electric-dipole, electric-quadrupole, and magnetic-dipole polarizabilities (see Fig. 2.5); and $\overleftrightarrow{\alpha}^{Qi}$ is a linear combination of $\overleftrightarrow{\alpha}^{Qi,\rm EQ}$ and $\overleftrightarrow{\alpha}^{Qi,\rm MD}$ (see Appendix 2.A for details). For simplicity, we shall henceforth use the term "quadrupole" to refer to both the electric-quadrupole and magnetic-dipole responses together, as described by the tensor $\overleftrightarrow{\alpha}^{Qi}$ which combines their effects.

Since $(a/\lambda) \ll 1$, where a is a molecular dimension and λ is a light wavelength, we are well justified in ignoring the higher-order multipoles beyond those shown in Eq. (2.2). More specifically, for a given molecule, each higher order in the multipole expansion is suppressed by an additional factor of order (a/λ), as usual. On the other hand, when averaged over many molecules in a centrosymmetric bulk, the signal associated with odd-rank susceptibility tensors (like $\overleftrightarrow{\alpha}^{D}$) vanishes, while even-rank contributions (like $\overleftrightarrow{\alpha}^{Qi}$) do not. Since there are many more molecules in the bulk than the surface, the effect of even-rank tensors is boosted by a factor of order $(a/\lambda)^{-1}$ compared to odd-rank. From these two considerations, $\overleftrightarrow{\alpha}^{D}$ and

$\overleftrightarrow{\alpha}^{Qi}$ *together* comprise the lowest-order term of an $(a/\lambda) \ll 1$ expansion, justifying the choice of terms shown in Eq. (2.2) [1].

Before using Eqs. (2.2), a specific position which we call the "molecular center" $\vec{O}$ must be given. (Usually, one would choose $\vec{O}$ to be at a point on the molecule.) This position serves several functions. First, it is the point where the fields and their derivatives are sampled, i.e.

$$\vec{E}_1 \equiv \vec{E}_1(\vec{O}) = \vec{E}_1^0 e^{i\vec{k}_1\cdot\vec{O}}, \quad \vec{E}_2 \equiv \vec{E}_2(\vec{O}) = \vec{E}_2^0 e^{i\vec{k}_2\cdot\vec{O}}. \tag{2.3}$$

(Similarly for $(\nabla\vec{E}_i)$ and $\vec{B}_i$.) Second, it is the point where the dipole and quadrupole radiation fields from $\vec{p}^{(2)}$, $\overleftrightarrow{q}^{(2)}$, $\vec{\mu}^{(2)}$ are assumed to be centered. This is relevant in calculating the phase delay in traveling to the detector: The measured signal amplitude is proportional to

$$\vec{\pi}^{(2)}_{\rm eff} \equiv \vec{p}^{(2)}_{\rm eff} e^{-i\vec{k}_{\rm SF}\cdot\vec{O}}. \tag{2.4}$$

Formally, the molecular center $\vec{O}$ is the origin about which the multipole expansion is performed. This theoretical parameter $\vec{O}$ cannot, of course, alter a measurable quantity such as $\vec{\pi}^{(2)}_{\rm eff}$. It does, however, alter other relevant but non-measurable parameters. To see this, we formally rewrite Eq. (2.2) in a way that distinguishes purely dipolar coupling between the molecule and input fields (yielding a response $\vec{p}^{(2)}_D$) from those involving a field gradient ($\vec{p}^{(2)}_Q$). We will soon see, however, that the division is not unique, but generally depends on $\vec{O}$.

$$\begin{aligned}
\vec{p}^{(2)}_{{\rm eff},\vec{O}}(\omega_{\rm SF},\vec{k}_{\rm SF}) &= \vec{p}^{(2)}_{D,\vec{O}}(\omega_{\rm SF},\vec{k}_{\rm SF}) + \vec{p}^{(2)}_{Q,\vec{O}}(\omega_{\rm SF},\vec{k}_{\rm SF}) \\
\vec{p}^{(2)}_{D,\vec{O}} &= \overleftrightarrow{\alpha}^{D}\!:\vec{E}_{1,\vec{O}}\vec{E}_{2,\vec{O}} \\
\vec{p}^{(2)}_{Q,\vec{O}} &= i\left[\overleftrightarrow{\alpha}^{Q1}_{\vec{O}}\!:\vec{E}_{1,\vec{O}}\vec{E}_{2,\vec{O}}\vec{k}_1 + \overleftrightarrow{\alpha}^{Q2}_{\vec{O}}\!:\vec{E}_{1,\vec{O}}\vec{E}_{2,\vec{O}}\vec{k}_2 - \overleftrightarrow{\alpha}^{Qs}_{\vec{O}}\!:\vec{E}_{1,\vec{O}}\vec{E}_{2,\vec{O}}\vec{k}_{\rm SF}\right]
\end{aligned} \tag{2.5}$$

(Quantities in Eq. (2.5) which vary depending on the molecular center are marked with a subscript $\vec{O}$). As we prove in Appendix 2.A, $\overleftrightarrow{\alpha}^{D}$ is independent of the molecular center $\vec{O}$, but $\overleftrightarrow{\alpha}^{Qi}$ depends on $\vec{O}$: If $\vec{O}$ and $\vec{O}' \equiv \vec{O} - \Delta\vec{O}$ are two molecular centers,

$$\overleftrightarrow{\alpha}^{Qi}_{\vec{O}'} = \overleftrightarrow{\alpha}^{Qi}_{\vec{O}} + \overleftrightarrow{\alpha}^{D}\Delta\vec{O} \tag{2.6}$$

To confirm that $\vec{\pi}^{(2)}_{{\rm eff},\vec{O}}$ is independent of $\vec{O}$, we combine Eqs. (2.2)-(2.5) to get

$$\vec{\pi}^{(2)}_{{\rm eff},\vec{O}} = [\vec{p}^{(2)}_{D,\vec{O}} + \vec{p}^{(2)}_{Q,\vec{O}}]e^{-i\vec{k}_{\rm SF}\cdot\vec{O}} \tag{2.7}$$

$$\vec{p}^{(2)}_{D,\vec{O}'}e^{-i\vec{k}_{\rm SF}\cdot\vec{O}'} - \vec{p}^{(2)}_{D,\vec{O}}e^{-i\vec{k}_{\rm SF}\cdot\vec{O}} = \left(\overleftrightarrow{\alpha}^{D}\!:\vec{E}_1^0\vec{E}_2^0\right)\left(e^{-i\Delta\vec{k}\cdot\vec{O}'} - e^{-i\Delta\vec{k}\cdot\vec{O}'}\right) \tag{2.8}$$

$$= \left(\overleftrightarrow{\alpha}^{D}\!:\vec{E}_1^0\vec{E}_2^0\right)\left(i\Delta\vec{k}\cdot\Delta\vec{O} + \cdots\right)e^{-i\Delta\vec{k}\cdot\vec{O}} \tag{2.9}$$

$$\vec{p}^{(2)}_{Q,\vec{O}'}e^{-i\vec{k}_{\rm SF}\cdot\vec{O}'} - \vec{p}^{(2)}_{Q,\vec{O}}e^{-i\vec{k}_{\rm SF}\cdot\vec{O}} = -i\left(\overleftrightarrow{\alpha}^{D}\!:\vec{E}_1^0\vec{E}_2^0\right)\left(\Delta\vec{k}\cdot\vec{O}\right)e^{-i\Delta\vec{k}\cdot\vec{O}} + \cdots \tag{2.10}$$

We then see, in the limit of neglecting higher-order terms of $|\Delta\vec{k}\cdot(\vec{O}-\vec{O}')|$ (i.e., in the spirit of multipole expansion, neglecting responses of higher order than quadrupole), that the molecular-center dependences of $\vec{p}_D^{(2)}$ and $\vec{p}_Q^{(2)}$ cancel each other:

$$\vec{\pi}^{(2)}_{\text{eff},\vec{O}} = [\vec{p}^{(2)}_{D,\vec{O}} + \vec{p}^{(2)}_{Q,\vec{O}}]e^{-i\vec{k}_{\text{SF}}\cdot\vec{O}} = [\vec{p}^{(2)}_{D,\vec{O}'} + \vec{p}^{(2)}_{Q,\vec{O}'}]e^{-i\vec{k}_{\text{SF}}\cdot\vec{O}'} = \vec{\pi}^{(2)}_{\text{eff},\vec{O}'} \tag{2.11}$$

This confirms the earlier statement that $\vec{\pi}^{(2)}_{\text{eff}}$ does not depend on the choice of the molecular center $\vec{O}$ of the coordinate system, but $\vec{p}_D^{(2)}$ and $\vec{p}_Q^{(2)}$ do. The latter often leads to ambiguity in distinguishing electric-dipole and quadrupole responses of a molecule.

2.3.2 Surface and Bulk SF Susceptibities

The bulk SF susceptibility of a system is usually defined as

$$\overleftrightarrow{\chi}^{(2)}_B(\omega_{\text{SF}},\vec{k}_{\text{SF}}) \equiv \frac{1}{V}\sum_j [\vec{p}^{(2)}_{\text{eff},j}/\vec{E}_{1j}\vec{E}_{2j}], \tag{2.12}$$

where the sum is over individual molecules (j) in a mesoscopic bulk volume V with dimensions large compared to a molecule, and $\vec{E}_{1j} = \vec{E}_1^0 e^{i\vec{k}_1\cdot\vec{O}_j}$, $\vec{E}_{2j} = \vec{E}_2^0 e^{i\vec{k}_2\cdot\vec{O}_j}$. (The vector division in (2.12) is defined in the obvious way: For the $(i,m,n)^{\text{th}}$ Cartesian component of $\overleftrightarrow{\chi}^{(2)}_B$, use the ith component of $\vec{p}^{(2)}_{\text{eff},j}$ m^{th} of $\vec{E}_{1j}$, and n^{th} of $\vec{E}_{2j}$.) For a medium with inversion symmetry, the electric-dipole part, $\vec{p}_D^{(2)}$, of $\vec{p}^{(2)}_{\text{eff}}$ vanishes in summation, but the quadrupole part, $\vec{p}_Q^{(2)}$, survives; as we showed in the preceding section, $\vec{p}_Q^{(2)}$ depends on the choice of molecular center. Hence, we expect that $\overleftrightarrow{\chi}^{(2)}_B$ must also depend on the choice of molecular center. From the expression for $\vec{p}_Q^{(2)}$ in Eq. (2.5), we obtain, for the choice of molecular center of the j^{th} molecule at $\vec{O}_j$,

$$\overleftrightarrow{\chi}^{(2)}_{B,\vec{O}} = \frac{1}{V}\sum_j i\left(\overleftrightarrow{\alpha}^{Q1}_{j,\vec{O}_j}\cdot\vec{k}_1 + \overleftrightarrow{\alpha}^{Q2}_{j,\vec{O}_j}\cdot\vec{k}_2 - \overleftrightarrow{\alpha}^{Qs}_{j,\vec{O}_j}\cdot\vec{k}_s\right). \tag{2.13}$$

If the molecular center is shifted from $\vec{O}_j$ to $\vec{O}'_j = \vec{O}_j - \Delta\vec{O}_j$, then from Eq. (2.6), we have

$$\overleftrightarrow{\chi}^{(2)}_{B,\vec{O}'} - \overleftrightarrow{\chi}^{(2)}_{B,\vec{O}} = \frac{-i}{V}\sum_j(\overleftrightarrow{\alpha}^D_j)(\vec{O}_j\cdot\Delta\vec{k}) = -in\left\langle\overleftrightarrow{\alpha}^D(\Delta\vec{O}\cdot\Delta\vec{k})\right\rangle = -in\left\langle\overleftrightarrow{\alpha}^D\Delta O_z\right\rangle\left|\Delta\vec{k}\right| \tag{2.14}$$

where n denotes the density of molecules, and angular brackets indicate an average over molecular orientations and arrangements in the bulk environment.

It was already known in the early development of second-harmonic generation and SFG for surface studies that the surface and bulk terms in $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ of Eq. (2.1) are not separable, either in theory or in measurement [9, 10]. On the other hand, as a measurable physical quantity,

$\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ naturally is independent of the choice of molecular center. Thus if $\overleftrightarrow{\chi}^{(2)}_{B}$ depends on the choice of molecular center, so must $\overleftrightarrow{\chi}^{(2)}_{S}$, but the sum of $\overleftrightarrow{\chi}^{(2)}_{S}$ and $\overleftrightarrow{\chi}^{(2)}_{B}/(-i\Delta k)$ must not. The explicit proof is briefly outlined here, with more details given in Appendix 2.C.

Similar to the bulk case, the surface SF susceptibility is generally defined as

$$\overleftrightarrow{\chi}^{(2)}_{S}(\omega_{\text{SF}}, \vec{k}_{\text{SF}}) \equiv \frac{1}{A} \sum_{\substack{\text{surface layer} \\ \text{molecules } j}} [\vec{p}^{(2)}_{\text{eff},j}/\vec{E}_{1j}\vec{E}_{2j}] \tag{2.15}$$

where the summation is on molecules in a surface layer over a surface area A. The surface layer is a thin region at the interface that is structurally different from the bulk. For second-order nonlinear optical response of an iostropic liquid like water, for example, the surface layer is the layer that has broken inversion symmetry.

Although this definition of "surface layer" is appropriate as a general guideline, it is too vague to uniquely specify exactly which molecules belong in the sum (2.15). The transition between anisotropic surface and isotropic bulk is gradual at an atomic level, not a sharp line; and even if it were a sharp line, there would be molecules straddling the boundary. Instead, it is shown in Appendix 2.B that the sum (2.15) should be modified to the following more specific definition of "surface layer":

$$\overleftrightarrow{\chi}^{(2)}_{S,\vec{O}}(\omega_{\text{SF}}, \vec{k}_{\text{SF}}) \equiv \frac{1}{A} \sum_{O_{j,z}>z_B} [\vec{p}^{(2)}_{\text{eff},j}/\vec{E}_{1j}\vec{E}_{2j}]. \tag{2.16}$$

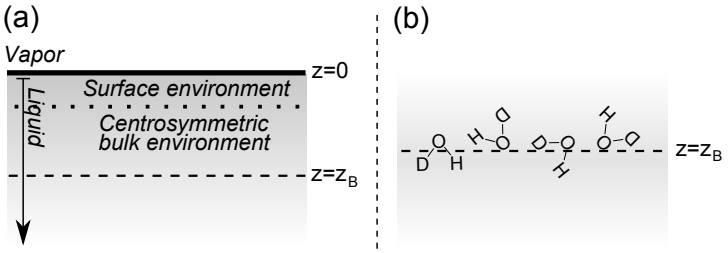

Figure 2.6 – (a) An illustration for Eq. (2.17). The system is the same as the one described in Fig. 2.1, with the plane $z = z_B$ indicated by a dashed line. The notional dividing plane $z = z_B$ could be placed at any arbitrary depth within the isotropic bulk environment. (b) A close-up view of the group of molecules with oxygen atoms immediately above the plane $z = z_B$. If oxygen is chosen as the molecular center, all of these molecules are included in Eq. (2.17); if hydrogen is chosen instead, only half are.

Here, as sketched in Fig. 2.6a, the interface is at $z = 0$, the semi-infinite bulk medium under discussion is at $z < 0$, and $z = z_B\,(< 0)$ is a plane sufficiently deep to be in the bulk environment, but separated from the interface only by a microscopically small distance. The molecules j with molecular center $\vec{O}_j$ above the plane $z = z_B$ are the ones included in the

sum (2.16). In practice, the surface layer is microscopically thin, so the quadrupole part, $\vec{p}_Q^{(2)}$, of $\vec{p}_{\rm eff}^{(2)}$ can be neglected in the summation restricted to the surface layer:

$$\overleftrightarrow{\chi}^{(2)}_{S,\vec{O}} = \frac{1}{A}\sum_{O_{j,z}>z_B} \overleftrightarrow{\alpha}^{D}_{j}\,. \tag{2.17}$$

Here, with $\overleftrightarrow{\alpha}^{D}_{j,\vec{O}_j}$ being independent of $\vec{O}_j$, the dependence on the choice of molecular center comes as through counting of the set of molecules included in the summation. Changing the molecular center from $\vec{O}_j$ to $\vec{O}'_j = \vec{O}_j - \Delta\vec{O}_j$, but keeping z_B unchanged for simplicity, will add some molecules near z_B to the sum in Eq. (2.17), and remove others. For illustration, we show in Fig. 2.6b a system of HDO molecules, with the plane $z = z_B$ denoted by the dashed line. There are a number of randomly oriented HDO molecules with their oxygen (O) just above the plane (Fig. 2.6b). If the molecular center of HDO is taken to be on O, these molecules will all be counted in the summation of Eq. (2.17). However, if the molecular center is taken to be on the hydrogen (H), then half of these molecules will no longer be counted in the sum, yielding a different value for $\overleftrightarrow{\chi}^{(2)}_S$. Since the effect of changing the molecular center on $\overleftrightarrow{\chi}^{(2)}_S$ is roughly from molecules within a layer of $|\Delta\vec{O}|$ thick, the change of $\overleftrightarrow{\chi}^{(2)}_S$ is expected to be of the order of $|\overleftrightarrow{\alpha}^{D}\Delta\vec{O}|$. In Appendix 2.C, we derive the net change of $\overleftrightarrow{\chi}^{(2)}_S$ due to the change from $\vec{O}_j$ to $\vec{O}'_j$ precisely as:

$$\overleftrightarrow{\chi}^{(2)}_{S,\vec{O}'} - \overleftrightarrow{\chi}^{(2)}_{S,\vec{O}} = -n\left\langle \overleftrightarrow{\alpha}^{D}\Delta O_z \right\rangle \tag{2.18}$$

(consistent with Ref. [6]). Since $\left\langle \overleftrightarrow{\alpha}^{D}\Delta O_z \right\rangle \neq 0$ in general, this explicitly shows that a different choice of $\vec{O}$ will yield a different $\overleftrightarrow{\chi}^{(2)}_S$.

From Eqs. (2.14) and (2.18), we see readily

$$\overleftrightarrow{\chi}^{(2)}_{S,\vec{O}} + \frac{\overleftrightarrow{\chi}^{(2)}_{B,\vec{O}}}{-i\left|\Delta\vec{k}\right|} = \overleftrightarrow{\chi}^{(2)}_{S,\vec{O}'} + \frac{\overleftrightarrow{\chi}^{(2)}_{B,\vec{O}'}}{-i\left|\Delta\vec{k}\right|} \tag{2.19}$$

As a result, $\overleftrightarrow{\chi}^{(2)}_{S,\rm eff}$ is independent of the choice of molecular center $\vec{O}$.

2.4 Ambiguities in molecular-dynamics calculations

Discussion in the previous section provides the correct framework for calculation to compare with experiment of SFG, emphasizing the need to calculate $\overleftrightarrow{\chi}^{(2)}_{S,\rm eff}$. Unfortunately, this is not the usual practice reported in the literature. Instead, it is simply assumed that $\overleftrightarrow{\chi}^{(2)}_B$ would vanish and just $\overleftrightarrow{\chi}^{(2)}_S$ is calculated, ignoring the fact that $\overleftrightarrow{\chi}^{(2)}_S$ depends on the choice of molecular center. As noticed recently by Noah-Vanhoucke et al. [6], different choices of

molecular center can yield significantly different SF surface spectra. We shall elucidate this by an example in a later section.

In practice, the ambiguity in $\overset{\leftrightarrow}{\chi}_S^{(2)}$ is typically resolved by arbitrarily choosing a molecular center to arrive at a definite answer. However, the arbitrary step (which may be subtle and unintentional) depends on the simulation approach. Therefore we focus on molecular simulations using a slab model, as described in the Introduction. For surface SFG, these simulations often impose an artificial boundary surface at the middle of the slab to break the inherent inversion symmetry of the slab [6]. We can relate this procedure to Eq. (2.17) by assuming the artificial boundary surface is at $z = z_B$. The molecules with molecular centers above the artificial boundary surface are summed up, with the other molecules ignored. This procedure is consistent with Eq. (2.17), so it is a legitimate way to calculate $\overset{\leftrightarrow}{\chi}_S^{(2)}$. However, it also inherits the pathology that $\overset{\leftrightarrow}{\chi}_S^{(2)}$ is an ambiguous quantity whose value depends on the choice of molecular center. Consequently, as noted in Ref. [6], the SF spectrum calculated from $\overset{\leftrightarrow}{\chi}_S^{(2)}$ by molecular simulation depends on the choice of molecular center. Clearly, the correct SF spectrum must come from calculating $\overset{\leftrightarrow}{\chi}_{S,\text{eff}}^{(2)} \equiv \overset{\leftrightarrow}{\chi}_S^{(2)} + \overset{\leftrightarrow}{\chi}_B^{(2)}/(-i\Delta k)$, which is independent of the choice of molecular center.

The higher-order bulk nonlinear susceptibility, $\overset{\leftrightarrow}{\chi}_B^{(2)}$, given by Eq. (2.13) (with further details in Appendix 2.A), is generally more difficult to calculate. Although detailed calculation procedures can be found in Refs. [19–22], the facts that $\overset{\leftrightarrow}{\chi}_B^{(2)}$ is an ambiguous quantity and that its value depends on the choice of molecular center have not been discussed in the literature. As shown in Appendix 2.A, a proper quantum-mechanical calculation of $\overset{\leftrightarrow}{\chi}_{B,\vec{O}}^{(2)}$ should use the "relative" position operators $\vec{r}^{(\vec{O})} \equiv \vec{r} - \vec{O}$ in the expressions for the moment operators $\vec{p}$, $\overset{\leftrightarrow}{q}$, $\vec{\mu}$. Therefore, the result will in general depend on the choice of $\vec{O}$.

In a molecular-simulation calculation of SFG, one would ideally determine $\overset{\leftrightarrow}{\chi}_{S,\text{eff}}^{(2)}$ by calculating both $\overset{\leftrightarrow}{\chi}_S^{(2)}$ and $\overset{\leftrightarrow}{\chi}_B^{(2)}$ with a consistent choice of molecular center. The difficulty of computing $\overset{\leftrightarrow}{\chi}_B^{(2)}$, however, may render this approach impractical. A simpler approach is to choose a molecular center that will minimize $\overset{\leftrightarrow}{\chi}_B^{(2)}$, and therefore maximize $\overset{\leftrightarrow}{\chi}_S^{(2)}$, for the frequency-range and other parameters under investigation. Then, we may be able to argue that $\left|\overset{\leftrightarrow}{\chi}_{B,\vec{O}}^{(2)}/\Delta\vec{k}\right| \ll \left|\overset{\leftrightarrow}{\chi}_{S,\vec{O}}^{(2)}\right|$, and hence $\overset{\leftrightarrow}{\chi}_{S,\text{eff}}^{(2)} \approx \overset{\leftrightarrow}{\chi}_{S,\vec{O}}^{(2)}$, allowing us to obtain a fairly accurate SF surface spectrum from calculating only $\overset{\leftrightarrow}{\chi}_S^{(2)}$. This is likely the case for isotropic liquids composed of small molecules, or small functional groups on larger molecules. In the following section, we use water as an example to illustrate these points.

2.5 Water as an Example

The air/neat-water interface has been studied extensively both in theory and in experiment. There are quite a few molecular simulations of the SF surface spectrum of the system reported

in the literature, but to our knowledge, none of them includes $\overleftrightarrow{\chi}_B^{(2)}$ in the calculation. Here, we show that $\overleftrightarrow{\chi}_B^{(2)}$ is not negligible if the molecular center is not judiciously chosen. We begin by analyzing $\overleftrightarrow{\chi}_B^{(2)}$ more closely. The contribution of $\overleftrightarrow{\chi}_B^{(2)}$ to the experimental signal can be calculated by appropriately combining its tensor elements, assuming an isotropic bulk. This contribution is proportional to [8]:

$$P_B^{(2)} \propto \frac{(\vec{k}_2\cdot\hat{e}_1)\chi^{Q2}_{xyxy} - (\vec{k}_s\cdot\hat{e}_1)\chi^{Qs}_{xyxy}}{\Delta k}(\hat{e}_2\cdot\hat{e}_s) + \\ + \frac{(\vec{k}_1\cdot\hat{e}_2)\chi^{Q1}_{xxyy} - (\vec{k}_s\cdot\hat{e}_2)\chi^{Qs}_{xxyy}}{\Delta k}(\hat{e}_1\cdot\hat{e}_s) + \frac{(\vec{k}_1\cdot\hat{e}_s)\chi^{Q1}_{xyyx} + (\vec{k}_2\cdot\hat{e}_s)\chi^{Q2}_{xyyx}}{\Delta k}(\hat{e}_1\cdot\hat{e}_2) \quad (2.20)$$

where $\hat{e}_1, \hat{e}_2, \hat{e}_s$ are the polarizations of the three waves, and following Eq. (2.13), we define $\overleftrightarrow{\chi}^{Qi} = (1/V)\sum_t \overleftrightarrow{\alpha}_t^{Qi}$, a sum over a representative volume V in the bulk.

We now consider, for simplicity, an isotopically diluted HDO:D_2O (in the infinite dilution limit) system and focus on SFG with one incoming wave ω_1 resonant with the OH stretch mode. In the OH stretch vibrational mode of HDO, the hydrogen atom oscillates along the OH bond, while the heavier oxygen atom is relatively stationary and the electronic wavefunction is not substantially perturbed. To the extent that one-dimensional hydrogen vibration dominates charge motion in this mode, we expect the corresponding transition quadrupole to be very small, $\overleftrightarrow{\alpha}^{Q1}_{\vec{r}_\mathrm{H}}$, provided the molecular center is placed at the H atoms equilibrium position $\vec{r}_\mathrm{H}$. On the other hand, the response of an HOD molecule at visible frequencies ω_2 and ω_SF arises mainly from valence electron fluctuations, which are governed by wave functions centered nearly at the oxygen atom. Assuming these transitions do not have a strong intrinsic quadrupole character—for example, if the electron cloud shifts back and forth without much distortion in shape—we expect that $\overleftrightarrow{\alpha}^{Q2}_{\vec{r}_\mathrm{O}} \approx 0$ and $\overleftrightarrow{\alpha}^{Qs}_{\vec{r}_\mathrm{O}} \approx 0$, where $\vec{r}_\mathrm{O}$ is the oxygen atom position. If we now choose the molecular center at an arbitrary point $\vec{O}$, then we find from Eq. (2.6):

$$(\alpha^{Q1}_{\vec{O}})_{j\ell mn} \approx \alpha^D_{j\ell m}\cdot\left(\vec{r}_\mathrm{H} - \vec{O}\right)_n,\ (\alpha^{Q2}_{\vec{O}})_{j\ell mn} \approx \alpha^D_{j\ell m}\cdot\left(\vec{r}_\mathrm{O} - \vec{O}\right)_n,\ (\alpha^{Qs}_{\vec{O}})_{j\ell mn} \approx \alpha^D_{j\ell m}\cdot\left(\vec{r}_\mathrm{O} - \vec{O}\right)_n \quad (2.21)$$

Substituting these idealized expressions for quadrupole susceptibilities into Eq. (2.20), assuming a typical experimental setup (45° angle of incidence, $\lambda_1 \approx 3\mu$m, $\lambda_2 = 512$nm [23]), we have

$$(\chi^{(2)}_{B,\vec{O}})_{SSP} \propto (\vec{k}_2\cdot\hat{e}_1)\chi^{Q2}_{\vec{O},xyxy} - (\vec{k}_s\cdot\hat{e}_1)\chi^{Qs}_{\vec{O},xyxy} \propto \left|\vec{r}_\mathrm{O} - \vec{O}\right| \quad (2.22)$$

$$(\chi^{(2)}_{B,\vec{O}})_{SPS} \propto (\vec{k}_1\cdot\hat{e}_2)\chi^{Q1}_{\vec{O},xxyy} - (\vec{k}_s\cdot\hat{e}_2)\chi^{Qs}_{\vec{O},xxyy} \propto \left|(0.995\vec{r}_\mathrm{O} + 0.005\vec{r}_\mathrm{H}) - \vec{O}\right| \quad (2.23)$$

$$(\chi^{(2)}_{B,\vec{O}})_{PSS} \propto (\vec{k}_1\cdot\hat{e}_s)\chi^{Q1}_{\vec{O},xyyx} + (\vec{k}_2\cdot\hat{e}_s)\chi^{Q2}_{\vec{O},xyyx} \propto \left|(0.85\vec{r}_\mathrm{O} + 0.15\vec{r}_\mathrm{H}) - \vec{O}\right| \quad (2.24)$$

where the abbreviations "SSP", "SPS", "PSS" each represent the polarizations for the SF, visible, and IR light, respectively. (For the various reflection-geometry experimental setups

in the literature, the decimals in Eqs. (2.23)–(2.24) change only slightly, although they are dramatically different for transmissive measurements.) The above equations suggest that $\overset{\leftrightarrow}{\chi}{}^{(2)}_B$ is nearly vanishing and $\overset{\leftrightarrow}{\chi}{}^{(2)}_{S,\text{eff}} \approx \overset{\leftrightarrow}{\chi}{}^{(2)}_S$ if the molecular center is chosen to be:

- At the O of HDO for SFG with SSP or SPS polarizations,
- Near the O of HDO, but displaced 15% of the way towards H along the OH bond, for PSS polarization.

Thus, with the proper choice of molecular center, we may only need to find $\overset{\leftrightarrow}{\chi}{}^{(2)}_S$ to obtain an approximately correct OH stretch spectrum for the air/HDO interface. This conclusion is supported by an MD calculation, as described below. We emphasize again, however, that this conclusion is contingent on the assumptions about HOD molecular transitions discussed above Eq. (2.21).

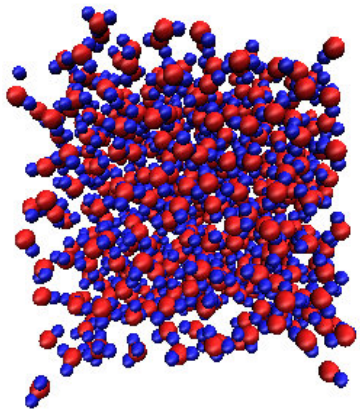

Figure 2.7 – An instantaneous water configuration from the MD simulation. Red and blue circles represent hydrogen and oxygen respectively. The top and bottom are the air interfaces, while the left, right, front, and back are periodic boundary conditions.

We have carried out MD simulations to calculate $\overset{\leftrightarrow}{\chi}{}^{(2)}_S$ for the air/HDO:D_2O interface with SSP polarization, using a slab model with different choices of molecular center. We closely followed the method of Ref. [6]. Specifically, we simulate an isotopically-diluted water system, comprising one HDO molecule together with 511 D_2O molecules. These SPC/E molecules were put in a $6 \times 6 \times 6$ nm^3 box, with 4 nm of vacuum separating the 2nm-thick slab from its periodic replicas. After Nosé-Hoover equilibration at 298K, we integrated Newton's equations of motion for 100 ps with LAMMPS software [24]. The OH stretch frequencies were inferred from the local electric fields, which in turn were calculated in LAMMPS using slab-corrected Ewald summation [25]. The SFG signal of each bond was calculated from the full HOD hyperpolarizability tensor, using the vacuum values tabulated in Ref. [6]. Im $\overset{\leftrightarrow}{\chi}{}^{(2)}_S$ was calculated from the simulations, and Re $\overset{\leftrightarrow}{\chi}{}^{(2)}_S$ inferred from Kramers-Kronig relations to obtain

$\left|\overleftrightarrow{\chi}_S^{(2)}\right|^2$. As discussed in Ref. [6], these simplified calculations neglect a host of effects including homogeneous broadening, motional narrowing, and intermolecular coupling. In addition to SSP polarization, an SPS spectrum (not shown) was also calculated, but its correspondence with the experimental measurements [26] is too weak to draw useful conclusions on the merits of different molecular centers. This poor agreement may be due to an important motional effect [26] not included in our static simulations.

The center-of-mass plane of the slab was used as the "artificial boundary" at each time-step: Molecules with their molecular centers above the plane were included in the top interface region, while molecules with molecular centers below the plane were ignored (but reused for a separate calculation for the bottom interface that was averaged into the final results).

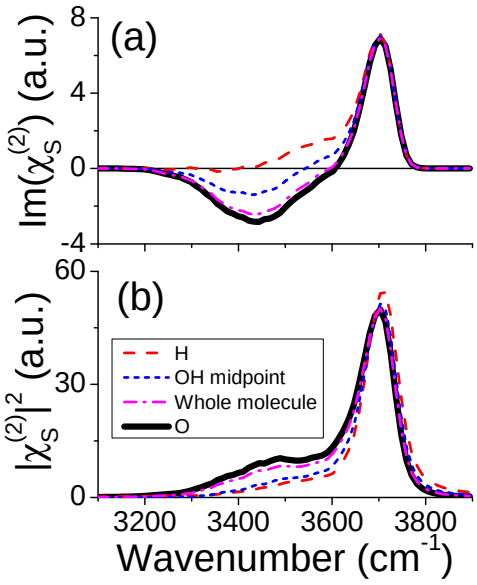

Figure 2.8 – Calculated SFG spectra for HOD:D_2O with different molecular centers. Light polarization is SSP. See text for descriptions.

Simulation results corresponding to different choices of molecular center are shown in Fig. 2.8. The calculations performed with oxygen as molecular center are plotted with a thicker line in the graphs, highlighting the fact that this choice is suggested to be optimal for SSP polarization in the analysis above. Note that, as expected, the peak at 3700 cm^{-1} is independent of molecular center: It is due to dangling OH bonds at the surface, and therefore is insensitive to how molecules in the bulk environment near the artificial boundary surface are counted.

The plots in Fig. 2.8 show the consequences of choosing O, H, or OH-midpoint as the molecular center. Also shown is the"whole molecule" result, where the contribution from HDO was included only if all three atoms of HDO were above the slab center-of-mass plane.

We calculated as well the spectrum for which the center of mass of HDO served as the molecular center [27]; it is almost indistinguishable from the result for O as the molecular center.

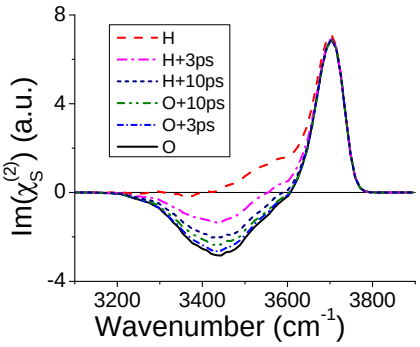

Figure 2.9 – Effect of time-delayed molecular center. Each molecule is counted as above or below the slab center-of-mass using hydrogen or oxygen as the molecular center. Following equilibrium dynamics over 0, 3, or 10ps, its contribution to the SFG signal is subsequently calculated. Polarization is SSP.

Others have used a different approach: Instead of choosing molecules nearest the interface in each configuration, the molecules are assigned to the top or bottom interface at the start of a short trajectory. The same molecules are counted as contributors throughout the time duration as they diffuse around [14–16]. In our formalism, we would say that the molecular center $\vec{O}$ for a molecule is chosen based on where the molecule *was*, instead of where it *is*. We illustrate the effect of this method in Fig. 2.9, where molecules are associated to an interface at one time-step, then allowed to freely time-evolve for a certain "delay", before their signal is calculated.

It is clear from Figs. 2.8–2.9 that the differences between $\mathrm{Im}\overset{\leftrightarrow(2)}{\chi}_S$ spectra calculated with different molecular centers can be comparable in magnitude to the spectra themselves. Therefore, we conclude that if the molecular center $\vec{O}$ is not carefully chosen, then $\left|\overset{\leftrightarrow(2)}{\chi}_{B,\vec{O}}\right|$ can be as large as $\overset{\leftrightarrow(2)}{\chi}_{S,\vec{O}}$. On the other hand, if $\vec{O}$ is judiciously chosen, it is possible that $\left|\overset{\leftrightarrow(2)}{\chi}_{B,\vec{O}}/\Delta\vec{k}\right| \ll \left|\overset{\leftrightarrow(2)}{\chi}_{S,\vec{O}}\right|$.

Finally, we compare the calculated spectra with experimental measurements. For SSP, the $\mathrm{Im}\overset{\leftrightarrow(2)}{\chi}_{S,\mathrm{eff}}$ spectrum for dilute HDO/D_2O isotopic mixtures has been measured directly [23], and shows a strong negative peak in the 3300-3600 cm^{-1} range. This feature is most consistent with our calculation that takes the oxygen molecular center. Therefore, our *a priori* reasoning about optimal choices for molecular centers appears well-founded. We note our calculation, similar to many others, is incapable of reproducing the experimentally observed positive band below 3300 cm^{-1} [3].

2.6 Conclusion

We have examined the formalism of SFVS in terms of multipole expansion, emphasizing that $\overleftrightarrow{\chi}^{(2)}_S$ and $\overleftrightarrow{\chi}^{(2)}_B$ are ambiguous, but their combination $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ is unambiguous. Accordingly, a calculation of $\overleftrightarrow{\chi}^{(2)}_S$ alone to describe the SF response should yield ambiguous results, as is the case in practice. This analysis explains the recently-discovered troubling ambiguity in slab-based molecular simulations [6].

Ambiguity in separation of surface and bulk contributions in second-harmonic and SF reflection from a surface or interface is a famously problematic issue [1, 8–10]. However, we show in this paper that we can still discuss from a *physical* perspective how to optimally choose the molecular center to minimize $\overleftrightarrow{\chi}^{(2)}_B$ in favor of $\overleftrightarrow{\chi}^{(2)}_S$ in the division of $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}}$ into $\overleftrightarrow{\chi}^{(2)}_S$ and $\overleftrightarrow{\chi}^{(2)}_B$. In many cases, we can then argue from physical reasoning that $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}} \approx \overleftrightarrow{\chi}^{(2)}_S$, so that the spectrum calculated from $\overleftrightarrow{\chi}^{(2)}_S$ alone should compare well with experiment.

We have examined in particular the case of air/water interface, where choosing different molecular centers and calculating only $\overleftrightarrow{\chi}^{(2)}_S$ in SF response can result in very different SF vibrational spectra [6]. The problem was resolved, at least approximately, by consideration of the charge motion within the water molecule. Such considerations allow us to predict *a priori* which molecular center should be chosen to minimize the contribution of $\overleftrightarrow{\chi}^{(2)}_B$. We accordingly suggest that the O atom is an optimal molecular center for reflection-geometry SSP and SPS polarizations, while the point 0.15Å from O towards H is optimal for PSS. Comparison between experimentally measured and computed spectra supports this expectation.

As a final note, we have argued that the surface and bulk can be mathematically split in many different ways. When combined, any split will give the same correct answer. In this sense, all splits are "equally valid". But this is a narrow statement; from other points of view, some splits are better than others. One example was discussed above: For simulation purposes, one might prefer the split that accomplishes the goal of $\overleftrightarrow{\chi}^{(2)}_{S,\text{eff}} \approx \overleftrightarrow{\chi}^{(2)}_S$. A different but very important example, not discussed in this thesis, is the goal of finding a *physically meaningful* split: One where $\overleftrightarrow{\chi}^{(2)}_S$ is helpful for *physically* understanding the surface, and $\overleftrightarrow{\chi}^{(2)}_B$ is helpful for *physically* understand the bulk. There is indeed a "most physically meaningful" split (called $\overleftrightarrow{\chi}^{(2)}_{SS}$ and $\overleftrightarrow{\chi}^{(2)}_{BB}$), and it is both theoretically unique and experimentally measurable. See Ref. [2] for a thorough discussion.

2.7 References

[1] Shen, Y. R. *Appl. Phys. B - Lasers O.* **68**, 295 (1999).

[2] Shen, Y. R. *J. Phys. Chem. C* **116**, 15505–15509 (2012).

[3] Tian, C. S. and Shen, Y. R. *Chem. Phys. Lett.* **470**, 1 (2009).

[4] Ishiyama, T. and Morita, A. *J. Chem. Phys.* **131**, 244714 (2009).

[5] Pieniazek, P. A., Tainter, C. J., and Skinner, J. L. *J. Chem. Phys.* **135**, 044701 (2011).

[6] Noah-Vanhoucke, J., Smith, J. D., and Geissler, P. L. *J. Phys. Chem. B* **113**, 4065 (2009).

[7] Shen, Y. R. *The Principles of Nonlinear Optics*. John Wiley & Sons, Inc., Hoboken, (1984).

[8] Held, H., Lvovsky, A. I., Wei, X., and Shen, Y. R. *Phys. Rev. B* **66**, 205110 (2002).

[9] Guyot-Sionnest, P. and Shen, Y. R. *Phys. Rev. B* **38**, 7985 (1988).

[10] Sipe, J. E., Mizrahi, V., and Stegeman, G. I. *Phys. Rev. B* **35**, 9091 (1987).

[11] Landau, L. D. and Lifshitz, E. M. *Electrodynamics of continuous media*. Butterworth-Heinemann, Oxford, 2nd edition, (1960).

[12] Pershan, P. S. *Phys. Rev.* **130**, 919 (1963).

[13] Raab, R. and De Lange, O. *Multipole Theory in Electromagnetism*. Clarendon Press, Oxford, (2005).

[14] Auer, B. M. and Skinner, J. L. *J. Chem. Phys.* **129**, 214705 (2008).

[15] Brown, E. C., Mucha, M., Jungwirth, P., and Tobias, D. J. *J. Phys. Chem. B* **109**, 7934 (2005).

[16] Perry, A., Ahlborn, H., Space, B., and Moore, P. B. *J. Chem. Phys.* **118**, 8411 (2003).

[17] Kastenholz, M. A. and Hunenberger, P. H. *J. Chem. Phys.* **124**, 124106 (2006).

[18] Hüneberger, P. and Reif, M. *Single-Ion solvation : Experimental and Theoretical Approaches to Elusive Thermodynamic Quantities*. Royal Society of Chemistry, Cambridge, (2010).

[19] Morita, A. *Chem. Phys. Lett.* **398**, 361 (2004).

[20] Neipert, C., Space, B., and Roney, A. B. *J. Phys. Chem. C* **111**, 8749 (2007).

[21] Munn, R. *Mol. Phys.* **89**, 555 (1996).

[22] Zhu, X. D. and Wong, A. *Phys. Rev. B* **46**, 2540 (1992).

[23] Tian, C. and Shen, Y. R. *J. Am. Chem. Soc.* **131**, 2790 (2009).

[24] Plimpton, S. *J. Comput. Phys.* **117**, 1 (1995).

[25] Yeh, I. and Berkowitz, M. L. *J. Chem. Phys.* **111**, 3155 (1999).

[26] Wei, X. and Shen, Y. R. *Phys. Rev. Lett.* **86**, 4799 (2001).

[27] Morita, A. *J. Phys. Chem. B* **110**, 3158 (2006).

[28] Barron, L. D. and Gray, C. G. *J. Phys. A: Math. Nucl. Gen.* **6**, 59 (1973).

[29] Guyot-Sionnest, P., Chen, W., and Shen, Y. R. *Phys. Rev. B* **33**, 8254 (1986).

[30] Eaves, J. D., Tokmakoff, A., and Geissler, P. L. *J. Phys. Chem. A* **109**, 9424–9436 (2005).

2.A Appendix: Microscopic expressions for electric-dipole, electric-quadrupole, and magnetic-dipole polarizabilities, and their relations to the choice of molecular center.

We show here, from the microscopic expressions of the electric-dipole and quadrupole SF polarizabilities of a molecule, $\overleftrightarrow{\alpha}^{D}$ and $\overleftrightarrow{\alpha}^{Qi}$, that $\overleftrightarrow{\alpha}^{D}$ does not depend on the choice of molecular center, but $\overleftrightarrow{\alpha}^{Qi}$ does. (The molecular-center dependence of $\overleftrightarrow{\chi}_S^{(2)}$ comes from the summation over $\overleftrightarrow{\alpha}^{D}$ in Eq. (2.17), not from the individual $\overleftrightarrow{\alpha}^{D}$s themselves.)

The basis of this analysis is the light-matter interaction Hamiltonian for a plane wave, written in the form of a multipole expansion about the point $\vec{O}$ [13, 28]:

$$H = -p_j E_j - q_{j\ell}\partial_j E_\ell - \mu_j B_j + \cdots \tag{2.25}$$

with $p_j \equiv \sum_\beta e^{(\beta)} r_j^{(\beta),(\vec{O})}$, $q_{j\ell} \equiv \frac{1}{2}\sum_\beta e^{(\beta)} r_j^{(\beta),(\vec{O})} r_\ell^{(\beta),(\vec{O})}$, $\mu_j \equiv \sum_\beta \frac{e^{(\beta)}}{2m^{(\beta)}} (\vec{r}^{(\beta),(\vec{O})} \times \vec{\pi}^{(\beta)})_j$, with $e^{(\beta)}$, $r^{(\beta)}$, $\pi^{(\beta)}$, and $m^{(\beta)}$ denoting the charge, position, momentum, and mass of the electron or ion labeled β, and where j and ℓ are Cartesian indices. In this expression, all EM fields and derivatives are evaluated at the point $\vec{O}$ (i.e., $\vec{E} \equiv \vec{E}(\vec{O})$, $\vec{B} \equiv \vec{B}(\vec{O})$, etc.), and all position-operators are defined relative to $\vec{O}$ (i.e. $\vec{r}^{(\beta),(\vec{O})} \equiv \vec{r}^{(\beta)} - \vec{O}$).

A multipole expansion about $\vec{O}$ of the outgoing sum-frequency radiation field is characterized by the induced moments $p_j^{(2)}(\omega_{\rm SF})$, $q_{j\ell}^{(2)}(\omega_{\rm SF})$, $\mu_j^{(2)}(\omega_{\rm SF})$ at the sum frequency are given by

$$p_j^{(2)}(\omega_{\rm SF}) = \alpha^{D}_{j\ell n} E_{1\ell} E_{2n} + \alpha^{Q1,{\rm EQ}}_{j\ell nt} (\partial_t E_{1\ell}) E_{2n} + \alpha^{Q2,{\rm EQ}}_{j\ell nt} E_{1\ell} (\partial_t E_{2n}) + \\ + \alpha^{Q1,{\rm MD}}_{j\ell n} B_{1\ell} E_{2n} + \alpha^{Q2,{\rm MD}}_{j\ell n} E_{1\ell} B_{2n} \tag{2.26}$$

$$q_{jt}^{(2)}(\omega_{\rm SF}) = \alpha^{Qs,{\rm EQ}}_{j\ell nt} E_{1\ell} E_{2n} \tag{2.27}$$

$$\mu_j^{(2)}(\omega_{\rm SF}) = \alpha^{Qs,{\rm MD}}_{j\ell n} E_{1\ell} E_{2n} \tag{2.28}$$

with

$$\alpha^{D}_{j\ell n} = F(p_j, p_\ell, p_n) \tag{2.29}$$

$$\alpha^{Qs,\mathrm{EQ}}_{j\ell nt} = F(q_{jt}, p_\ell, p_n) \tag{2.30}$$

$$\alpha^{Qs,\mathrm{MD}}_{j\ell nt} = F(\mu_j, p_\ell, p_n) \tag{2.31}$$

$$\alpha^{Q1,\mathrm{EQ}}_{j\ell nt} = F(p_j, q_{\ell t}, p_n) \tag{2.32}$$

$$\alpha^{Q1,\mathrm{MD}}_{j\ell n} = F(p_j, \mu_\ell, p_n) \tag{2.33}$$

$$\alpha^{Q2,\mathrm{EQ}}_{j\ell nt} = F(p_j, p_\ell, q_{nt}) \tag{2.34}$$

$$\alpha^{Q2,\mathrm{MD}}_{j\ell n} = F(p_j, p_\ell, \mu_n) \tag{2.35}$$

$$F(A,B,C) = \frac{-1}{\hbar^2}\sum_{g,s,t}\left[\rho^{(0)}_{gg}\frac{(A)_{gs}(B)_{st}(C)_{tg}}{(\omega_2 - \omega_{tg} + i\Gamma_{tg})(\omega_{\mathrm{SF}} - \omega_{sg} + i\Gamma_{sg})} + \cdots\right] \tag{2.36}$$

where "$\cdots$" denotes five other similar terms, ω_{sg} and Γ_{sg} and the frequency and damping constant of the resonance between states s and g, and A, B, and C represent operators involved in the light-matter interaction (Eq. (2.25)) at frequencies ω_{SF}, ω_1, and ω_2, respectively. (The conclusions would not be altered if the Γs were replaced with a more sophisticated treatment of dephasing.) The "effective" electric-dipole moment, $\vec{p}^{(2)}_{\mathrm{eff}}(\omega_{\mathrm{SF}}, \vec{k}_{\mathrm{SF}})$ is defined as the oscillating dipole moment which, by itself, would produce the same SF radiation in the direction $\vec{k}_{\mathrm{SF}}$ as is due in reality to $\vec{p}^{(2)}$, $\overleftrightarrow{q}^{(2)}$, and $\vec{\mu}^{(2)}$ *together*. This can be written as:

$$\begin{aligned} p^{(2)}_{\mathrm{eff},j}(\omega_{\mathrm{SF}}, \vec{k}_{\mathrm{SF}}) = p^{(2)}_j - (i\vec{k}_{\mathrm{SF}} \cdot \overleftrightarrow{q}^{(2)})_j - \frac{c}{\omega_{\mathrm{SF}}}(\vec{k}_{\mathrm{SF}} \times \vec{\mu}^{(2)}))_j = \\ = \left(\alpha^{D}_{j\ell n} + i\left(\alpha^{Q1}_{j\ell nt}k_{1t} + \alpha^{Q2}_{j\ell nt}k_{2t} - \alpha^{Qs}_{j\ell nt}k_{\mathrm{SF},t}\right)\right)E_{1\ell}E_{2n} \end{aligned} \tag{2.37}$$

in which $\overleftrightarrow{\alpha}^{Qi}$ includes both electric-quadrupole and magnetic-dipole parts:

$$\alpha^{Qs}_{j\ell nt} = \alpha^{Qs,\mathrm{EQ}}_{j\ell nt} + \frac{i}{\omega_{\mathrm{SF}}}\alpha^{Qs,\mathrm{MD}}_{b\ell n}\epsilon_{btj} \tag{2.38}$$

$$\alpha^{Q1}_{j\ell nt} = \alpha^{Q1,\mathrm{EQ}}_{j\ell nt} - \frac{i}{\omega_1}\alpha^{Q1,\mathrm{MD}}_{jbn}\epsilon_{bt\ell} \tag{2.39}$$

$$\alpha^{Q2}_{j\ell nt} = \alpha^{Q2,\mathrm{EQ}}_{j\ell nt} - \frac{i}{\omega_2}\alpha^{Q2,\mathrm{MD}}_{j\ell b}\epsilon_{btn} \tag{2.40}$$

where $\epsilon_{j\ell n}$ is the Levi-Civita tensor.

Next, we wish to show that $\overleftrightarrow{\alpha}^{D}$ does not depend on the choice of molecular center, while $\overleftrightarrow{\alpha}^{Qi}$ does. To proceed, we state a few mathematical identities which are straightforward but tedious to prove. First,

$$0 = F(c,X,Y) = F(X,c,Y) = F(X,Y,c) \tag{2.41}$$

for any operators X, Y and any c-number c. Second,

$$\sum_\beta \frac{e^{(\beta)}}{m^{(\beta)}} F(\pi_j^{(\beta)}, \mu_\ell, \mu_n) = -i\omega_{\rm SF} F(\mu_j, \mu_\ell, \mu_n) \tag{2.42}$$

$$\sum_\beta \frac{e^{(\beta)}}{m^{(\beta)}} F(\mu_j, \pi_\ell^{(\beta)}, \mu_n) = i\omega_1 F(\mu_j, \mu_\ell, \mu_n) \tag{2.43}$$

$$\sum_\beta \frac{e^{(\beta)}}{m^{(\beta)}} F(\mu_j, \mu_\ell, \pi_n^{(\beta)}) = i\omega_2 F(\mu_j, \mu_\ell, \mu_n) \tag{2.44}$$

These three equations are derived from $\vec{\pi}^{(\beta)} = (mi/\hbar)[H_0, r^{(\beta)}]$, where H_0 is the unperturbed Hamiltonian, $H_0 = V(\vec{r}^{(1)}, \vec{r}^{(2)}, \ldots) + \sum_\beta |\vec{\pi}^{(\beta)}|^2/(2m^{(\beta)})$. Third, for different molecular centers $\vec{O}$ and $\vec{O}'$,

$$p_j^{(\vec{O}')} = p_j^{(\vec{O})} + (\vec{O} - \vec{O}')_j Q \tag{2.45}$$

$$q_{\ell t}^{(\vec{O}')} = q_{\ell t}^{(\vec{O})} + \frac{1}{2}(\vec{O} - \vec{O}')_t p_\ell^{(\vec{O})} + \frac{1}{2}(\vec{O} - \vec{O}')_\ell p_t^{(\vec{O}')}, \tag{2.46}$$

$$\mu^{(\vec{O}')} = \mu^{(\vec{O}')} + (\vec{O} - \vec{O}') \times \sum_\beta \frac{e^{(\beta)}}{2m^{(\beta)}} \vec{\pi}^{(\beta)}. \tag{2.47}$$

where $Q \equiv \sum_\beta e^{(\beta)}$ is the total charge.

Using Eqs. (2.41) and (2.45), and the fact that $(\vec{O} - \vec{O}')Q$ is a constant (c-number) vector, we can prove that α^D is independent of $\vec{O}$. However, $\overleftrightarrow{\alpha}^{Qi}$ *does* have molecular-center dependence: Using Eqs. (2.43),(2.46),(2.47):

$$\alpha_{j\ell nt}^{Q1,{\rm EQ},(\vec{O}')} = \alpha_{j\ell nt}^{Q1,{\rm EQ},(\vec{O})} + \frac{1}{2}\alpha_{j\ell n}^D (\vec{O} - \vec{O}')_t + \frac{1}{2}\alpha_{jtn}^D (\vec{O} - \vec{O}')_\ell \tag{2.48}$$

$$\alpha_{j\ell n}^{Q1,{\rm MD},(\vec{O}')} = \alpha^{Q1,{\rm MD},(\vec{O})} + \frac{i\omega_1}{2}\epsilon_{\ell tb}\alpha_{jbn}^D (\vec{O} - \vec{O}')_t \tag{2.49}$$

and therefore, by Eq. (2.39),

$$\alpha_{j\ell nt}^{Q1,(\vec{O}')} - \alpha_{j\ell nt}^{Q1,(\vec{O})} = \frac{1}{2}\alpha_{jtn}^D (\vec{O} - \vec{O}')_\ell + \frac{1}{2}\alpha_{j\ell n}^D (\vec{O} - \vec{O}')_t + \frac{1}{2}\epsilon_{bt\ell}\epsilon_{bdf}\alpha_{jfn}^D (\vec{O} - \vec{O}')_t = \alpha_{j\ell n}^D (\vec{O} - \vec{O}')_t. \tag{2.50}$$

Similarly for all $\overleftrightarrow{\alpha}^{Qi}$,

$$\alpha_{j\ell nt}^{Qi,(\vec{O}')} - \alpha_{j\ell nt}^{Qi,(\vec{O})} = \alpha_{j\ell n}^D (\vec{O} - \vec{O}')_t. \tag{2.51}$$

Thus, $\overleftrightarrow{\alpha}^{Qi}$ depends on the choice of molecular center in a way that incorporates some part of the dipole susceptibility.

2.B Appendix: Effective surface susceptibility

The SF signals in transmission and reflection from a semi-infinite medium are proportional to $\left|\overset{\leftrightarrow}{\chi}^{(2)}_{S,\text{eff}}\right|^2$, where the effective surface nonlinear susceptibility $\overset{\leftrightarrow}{\chi}^{(2)}_{S,\text{eff}}$ is related to the surface and bulk nonlinear susceptibilities $\overset{\leftrightarrow}{\chi}^{(2)}_{S,\vec{O}}$ and $\overset{\leftrightarrow}{\chi}^{(2)}_{B,\vec{O}}$ by

$$\overset{\leftrightarrow}{\chi}^{(2)}_{S,\text{eff}} \equiv \overset{\leftrightarrow}{\chi}^{(2)}_{S,\vec{O}} + \frac{\overset{\leftrightarrow}{\chi}^{(2)}_{B,\vec{O}}}{-i\Delta k} \tag{2.52}$$

Macroscopically, this relation is usually derived from a three-layer model [29]. However, as discussed in the text, there is no unique sharp line splitting surface and bulk, so it is not necessarily obvious how to define $\overset{\leftrightarrow}{\chi}^{(2)}_{S}$ in terms of microscopic variables, nor how the definition should depend on the molecular center $\vec{O}$. We restrict attention to the simple case we have been considering, where the index of refraction is constant everywhere. The purpose of this section is to show explicitly that Eq. (2.52) is consistent with the microscopic definitions Eqs. (2.13) and (2.17), which we copy here for convenience:

$$\overset{\leftrightarrow}{\chi}^{(2)}_{B,\vec{O}} = \frac{1}{V}\sum_j i\left(\overset{\leftrightarrow}{\alpha}^{Q1}_{j,\vec{O}_j}\cdot\vec{k}_1 + \overset{\leftrightarrow}{\alpha}^{Q2}_{j,\vec{O}_j}\cdot\vec{k}_2 - \overset{\leftrightarrow}{\alpha}^{Qs}_{j,\vec{O}_j}\cdot\vec{k}_s\right), \tag{2.13}$$

$$\overset{\leftrightarrow}{\chi}^{(2)}_{S,\vec{O}} = \frac{1}{A}\sum_{O_{j,z}>z_B}\overset{\leftrightarrow}{\alpha}^{D}_{j}. \tag{2.17}$$

The SF output field from the surface and the bulk of the medium can be considered as generated by an equivalent surface sheet of nonlinear polarization, $\vec{P}^{(2)}_{S,\text{eff}}$, which contains contributions from all molecules:

$$\vec{P}^{(2)}_{\text{eff}} = \frac{1}{A}\sum_j \vec{\pi}^{(2)}_{\text{eff},j} = \frac{1}{A}\sum_{(\vec{O}_j)_z>z_B}\vec{\pi}^{(2)}_{\text{eff},j} + \frac{1}{A}\sum_{(\vec{O}_j)_z<z_B}\vec{\pi}^{(2)}_{\text{eff},j} \tag{2.53}$$

where $\vec{\pi}^{(2)}_{\text{eff},j}$ is the contribution of the j^{th} molecule to the signal amplitude at the detector, Eq. (2.4). The above equation splits the sum of $\vec{\pi}^{(2)}_{\text{eff}}$ over all molecules into a sum of molecules with molecular center above the $z = z_B$ plane, and a sum of molecules with molecular centers below. Here, the coordinate system is defined so that the medium spans $z < 0$, and $|z_B|$ is the thickness of the surface layer (Fig. 2.6a), assumed to be much less than a wavelength but thick enough that $z = z_B$ is indistinguishable from the bulk environment.

Starting with the bulk part,

$$\frac{1}{A}\sum_{(\vec{O}_j)_z<z_B}\vec{\pi}^{(2)}_{\text{eff},j} = \frac{1}{A}\int_{-\infty}^{z_B}\left(\sum_{(\vec{O}_j)_z=z}\vec{p}^{(2)}_{\text{eff},j,\vec{O}_j}\,e^{-i\vec{k}_{\text{SF}}\cdot\vec{O}_j}\right)dz. \tag{2.54}$$

Here, the sum is split into an integral, grouping together molecules with molecular center at a given height $z = z_B$. For each $z \leq z_B$, the contributions from $\vec{p}_D^{(2)}$ (Eq. 2.54) cancel *exactly*, due to the centrosymmetry of each group of molecules. More specifically, with the incoming fields given by $\vec{E}_1(\vec{r}) = \vec{E}_1^0 e^{i\vec{k}_1\cdot\vec{r}}$ and $\vec{E}_2(\vec{r}) = \vec{E}_2^0 e^{i\vec{k}_2\cdot\vec{r}}$, we can write

$$\sum_{(\vec{O}_j)_z=z} \vec{p}^{(2)}_{D,j,\vec{O}_j} e^{-i\vec{k}_{\rm SF}\cdot\vec{O}_j} = \left(\sum_{(\vec{O}_j)_z=z} \overset{\leftrightarrow}{\alpha}{}^D_j e^{-i\Delta\vec{k}\cdot\vec{O}_j}\right) \vdots\, \vec{E}_1^0\vec{E}_2^0 \tag{2.55}$$

For any group of bulk molecules with centrosymmetric distribution of orientations, the sum $\sum_j \overset{\leftrightarrow}{\alpha}{}^D_j$ vanishes exactly. However, the sum above also involves an exponential phase factor, and with that factor, the sum from such a group of molecules need *not* vanish exactly, but only to lowest order in the exponential. In this analysis, we are grouping molecules by the height of their molecular center to ensure that the sum (2.55) vanishes exactly (not just to lowest order) because the molecules in these groups have not only centrosymmetric orientation distributions, but also constant exponential factor.

Therefore, we can drop the contributions from $\vec{p}_D^{(2)}$, and write the bulk part as

$$\frac{1}{A}\sum_{(\vec{O}_j)_z>z_B} \vec{\pi}^{(2)}_{{\rm eff},j} = \left(\frac{1}{A}\int_{-\infty}^{z_B}\left(\sum_{(\vec{O}_j)_z=z} e^{-i\Delta\vec{k}\cdot\vec{O}_j} i\left(\overset{\leftrightarrow}{\alpha}{}^{Q1}_{j,\vec{O}_j}\cdot\vec{k}_1 + \overset{\leftrightarrow}{\alpha}{}^{Q2}_{j,\vec{O}_j}\cdot\vec{k}_2 - \overset{\leftrightarrow}{\alpha}{}^{Qs}_{j,\vec{O}_j}\cdot\vec{k}_s\right)\right) dz\right) \vdots\, \vec{E}_1^0\vec{E}_2^0$$

$$= \left(\overset{\leftrightarrow}{\chi}{}^{(2)}_{B,\vec{O}} \vdots\, \vec{E}_1^0\vec{E}_2^0\right)\int_{-\infty}^{z_B} e^{-i|\Delta k|z}dz \approx \left(\overset{\leftrightarrow}{\chi}{}^{(2)}_{B,\vec{O}} \vdots\, \vec{E}_1^0\vec{E}_2^0\right)\int_{-\infty}^{0} e^{-i|\Delta k|z}dz = \frac{\overset{\leftrightarrow}{\chi}{}^{(2)}_{B,\vec{O}}}{-i|\Delta k|} \vdots\, \vec{E}_1^0\vec{E}_2^0$$

(using Eq. 2.13).

For the surface part, since we are neglecting the possibility of field discontinuities across the interface, the quadrupole contribution of surface molecules is negligible compared to their dipole contribution. (We are assuming the surface layer is much thinner than a wavelength.) Therefore we have:

$$\frac{1}{A}\sum_{(\vec{O}_j)_z>z_B} \vec{\pi}^{(2)}_{{\rm eff},j} \approx \left(\frac{1}{A}\sum_{(\vec{O}_j)_z>z_B} \overset{\leftrightarrow}{\alpha}{}^D_j e^{-i\Delta\vec{k}\cdot\vec{O}_j}\right) \vdots\, \vec{E}_1^0\vec{E}_2^0 \approx \left(\frac{1}{A}\sum_{(\vec{O}_j)_z>z_B} \overset{\leftrightarrow}{\alpha}{}^D_j\right) \vdots\, \vec{E}_1^0\vec{E}_2^0 = \overset{\leftrightarrow}{\chi}{}^{(2)}_{S,\vec{O}} \vdots\, \vec{E}_1^0\vec{E}_2^0$$

(using Eq. 2.17).

This proves that the microscopic definitions of $\overset{\leftrightarrow}{\chi}{}^{(2)}_{B,\vec{O}}$ and $\overset{\leftrightarrow}{\chi}{}^{(2)}_{S,\vec{O}}$ in Eqs. (2.13) and (2.17) are indeed appropriate to get results consistent with macroscopic continuum models.

2.C Appendix: Dependence of $\chi_S^{(2)}$ on the choice of molecular center

To calculate the dependence of $\overset{\leftrightarrow}{\chi}_S^{(2)}$ on $\vec{O}$, we start with Eq. (2.17) from above:

$$\overset{\leftrightarrow}{\chi}_{S,\vec{O}}^{(2)} = \frac{1}{A} \sum_{O_{j,z}>z_B} \overset{\leftrightarrow}{\alpha}_j^D . \tag{2.17}$$

We assume that the molecular center $\vec{O}$ for each molecule is fixed in molecular coordinates. Changing the molecular center from $\vec{O}$ to $\vec{O}' = \vec{O} - \Delta\vec{O}$, but keeping z_B unchanged, will remove from the sum those molecules with $O_z > z_B$ but $O'_z < z_B$, i.e. molecules originally in the region $z_B < O_z < z_B + \Delta O_z$. Likewise, it will add to the sum molecules with $O_z < z_B$ but $O'_z > z_B$, i.e. molecules originally in the region $z_B > O_z > z_B + \Delta O_z$.

Consider an ensemble of molecules with a given orientation $\overset{\leftrightarrow}{\Omega}$. For this ensemble, DO_z has a constant value $\Delta O_z^{(\overset{\leftrightarrow}{\Omega})}$. If $\Delta O_z^{(\overset{\leftrightarrow}{\Omega})} > 0$, some molecules from this ensemble will be eliminated from the sum, namely those in the region $z_B < O_z < z_B + \Delta O_z^{(\overset{\leftrightarrow}{\Omega})}$. The volume of this region is $A\Delta O_z^{(\overset{\leftrightarrow}{\Omega})}$, so the change in $\overset{\leftrightarrow}{\chi}_S^{(2)}$ due to these eliminated molecules is $-n^{(\overset{\leftrightarrow}{\Omega})}\langle\overset{\leftrightarrow}{\alpha}^D\rangle_{\overset{\leftrightarrow}{\Omega}}\Delta O_z^{(\overset{\leftrightarrow}{\Omega})} d\overset{\leftrightarrow}{\Omega}$, where $n^{(\overset{\leftrightarrow}{\Omega})} d\overset{\leftrightarrow}{\Omega}$ is the differential number density of molecules with the given orientation $\overset{\leftrightarrow}{\Omega}$. Likewise, if $\Delta O_z^{(\overset{\leftrightarrow}{\Omega})} < 0$, some molecules from this ensemble will be *added into* the sum (2.17); namely those in the region $z_B > O_z > z_B + \Delta O_z^{(\overset{\leftrightarrow}{\Omega})}$. The volume of this region is $\left|A\Delta O_z^{(\overset{\leftrightarrow}{\Omega})}\right| = -A\Delta O_z^{(\overset{\leftrightarrow}{\Omega})}$, so the change in $\overset{\leftrightarrow}{\chi}_S^{(2)}$ due to these eliminated molecules is $-n^{(\overset{\leftrightarrow}{\Omega})}\langle\overset{\leftrightarrow}{\alpha}^D\rangle_{\overset{\leftrightarrow}{\Omega}}\Delta O_z^{(\overset{\leftrightarrow}{\Omega})} d\overset{\leftrightarrow}{\Omega}$, the same expression as before.

Altogether,

$$\overset{\leftrightarrow}{\chi}_{S,\vec{O}'}^{(2)} - \overset{\leftrightarrow}{\chi}_{S,\vec{O}}^{(2)} = \int d\overset{\leftrightarrow}{\Omega} \left(-n^{(\overset{\leftrightarrow}{\Omega})} \langle \overset{\leftrightarrow}{\alpha}_{\overset{\leftrightarrow}{\Omega}}^D \Delta O_z^{(\overset{\leftrightarrow}{\Omega})} \right) = -n\langle \overset{\leftrightarrow}{\alpha}^D \Delta O_z \rangle,$$

where the angle-brackets are a statistical average over molecules' orientations and local environments. This formula is consistent with that derived in Noah-Vanhoucke et al. [6].

2.D Appendix: Modification of LAMMPS source-code to output electric forces

As described above, the process for inferring an SFG spectrum from a molecular configuration followed the simplified approach of Ref. [6]. An important step is that the instantaneous OH vibrational frequency is inferred from the electric field on the hydrogen atom, following a presumed linear dependence established in prior studies [30].

Molecular simulations were performed using the LAMMPS software [24]. This software, of course, internally calculates electric fields in the course of the simulation, as the electric force is an important component of the total force on the atoms (other components include the Lennard-Jones force, the bond stiffness force, etc.). However, even though LAMMPS knows the electric field internally, it is not possible to output it. In this section, I describe how to modify the LAMMPS source code to allow the electric field data to be exported and saved.

I was using the following configuration:

- "cut/coul/long" pair-style
- "full" atom-style
- "verlet" (not rRESPA) integration
- Ewald (not pppm) field calculation
- 15 January 2010 version of the LAMMPS software

In the following, I will go through the source code file-by-file, saying what lines need to be modified and how.

`atom.h`

Add into "`public`" area:

```
double **fcoul;  // my code
```

"`atom->fcoul`" will be the Coulomb component of the force felt by the atom. This is modeled on "`atom->f`", the total force on the atom, which is stored and accessed in the same way.

`atom.cpp`

Add a line into the initialization (i.e. `Atom::Atom(LAMMPS *lmp) : Pointers(lmp)`):

```
x = v = f = NULL;
fcoul = NULL; // my code
```

Add a line into the destruction (i.e.`Atom::~Atom`):

```
memory->destroy_2d_double_array(f);
memory->destroy_2d_double_array(fcoul); // my code
```

`atom_vec_full.h`

Add into "`private`" area (after "`double **x,**v,**f;`"):

```
double **fcoul; // my code
```

`atom_vec_full.cpp`

In "`void AtomVecFull::grow(int n)`", add an extra line:

```
f = atom->f = memory->grow_2d_double_array(atom->f,nmax,3,"atom:f");
fcoul = atom->fcoul = memory->
           grow_2d_double_array(atom->fcoul,nmax,3,"atom:fcoul"); // my code
```

In "`double AtomVecFull::memory_usage()`", add an extra line:

```
if (atom->memcheck("f")) bytes += nmax*3 * sizeof(double);
if (atom->memcheck("fcoul")) bytes += nmax*3 * sizeof(double); // my code
```

In `void AtomVecFull::grow_reset()`, add an extra line:

```
x = atom->x; v = atom->v; f = atom->f;
fcoul = atom->fcoul; // my code
```

There are two more places where "`f`" is used in `atom_vec_full.cpp`: `pack_reverse` and `unpack_reverse`. These should not get called, as explained below.

`verlet.cpp`

In `void Verlet::force_clear()` (which resets the force on atoms to zero), wherever it resets a force, it should also reset fcoul. There should be three places in the program where you modify it to look like:

```
f[i][0] = 0.0;
f[i][1] = 0.0;
f[i][2] = 0.0;
// my code
atom->fcoul[i][0] = 0.0;
atom->fcoul[i][1] = 0.0;
atom->fcoul[i][2] = 0.0;
// end my code
```

(f was defined earlier in the program as atom->f).

Note on Coulomb force calculation:

If you look at verlet.cpp, the Coulomb force is calculated in two places: The short-range part is calculated in the command pair->compute, and the long-range part in the command kspace->compute. In both cases, we will "intercept" the Coulomb force data as it gets written into atom->f, so that we can also write it into atom->fcoul.

---

pair_lj_cut_coul_long.cpp

Here is where the short-range part of the Coulomb force is calculated.

There are four main functions here, compute, compute_inner, compute_middle, compute_outer. The last three are for rRESPA and do not get used in Velocity Verlet, so ignore them. You only need to modify compute (more specifically,
void PairLJCutCoulLong::compute(int eflag, int vflag)).

You will see the following code to store the total short-range force:

```
fpair = (forcecoul + factor_lj*forcelj) * r2inv;

f[i][0] += delx*fpair;
f[i][1] += dely*fpair;
f[i][2] += delz*fpair;
if (newton_pair || j < nlocal) {
  f[j][0] -= delx*fpair;
  f[j][1] -= dely*fpair;
  f[j][2] -= delz*fpair;
}
```

(f was defined earlier in the program as atom->f). There are three things to notice. First, the program makes it clear how fpair is the sum of a Coulomb force and an LJ force. Therefore it is easy to just extract the Coulomb part. Second, Newton's third law is being used: The force on j is minus the force on i. In LAMMPS, Newton's third law is always invoked if j is on the same processor as i (i.e., j<nlocal), and it is also always invoked if the LAMMPS script has the "newton" flag turned on (which is the LAMMPS default). Finally, when forcecoul is calculated (just earlier in the program), it uses factor_coul which comes from special_coul which comes from the LAMMPS special_bonds command. The default for special_bonds is that atoms directly or indirectly bonded do not have their Coulomb forces included. (So if factor_coul=0, the pairwise Coulomb force should be zero, and in fact forcecoul is set so that this short-range force cancels out the contribution to the Ewald-sum long-range force.) If you instead want the *total* Coulomb force, including Coulomb's law applied to bonded atoms, you need to change the pair_special setting in

your LAMMPS script, or else edit the `forcecoul` calculation just above to have it store extra information. (`factor_lj` is a similar thing.)

As mentioned, just drop out the LJ part and keep the Coulomb part, and you should be able to store the short-range Coulomb force. So put in the following code:

```
// my code: fcoul stores the Coulomb contribution to force
atom->fcoul[i][0] += delx * forcecoul * r2inv;
atom->fcoul[i][1] += dely * forcecoul * r2inv;
atom->fcoul[i][2] += delz * forcecoul * r2inv;
if (newton_pair || j < nlocal) {
  atom->fcoul[j][0] -= delx * forcecoul * r2inv;
  atom->fcoul[j][1] -= dely * forcecoul * r2inv;
  atom->fcoul[j][2] -= delz * forcecoul * r2inv;
}
// end my code
```

Also, in your LAMMPS input script, you must turn off "newton":

```
newton off
```

If you want to leave it on, you need to figure out how to modify `void Comm::reverse_comm()` in `comm.cpp`, including its subroutines in `atom_vec_full.cpp` (i.e., `AtomVecFull::unpack_reverse` and `AtomVecFull::pack_reverse`). This function accounts for the fact that part of the force on some of the atoms was calculated by the wrong processor when newton is turned on, by making the processors communicate their calculations with each other. Therefore if you use newton, the function needs to be modified to also correctly communicate `fcoul`. But I did not bother to figure out the details: I am using serial mode anyway.

`ewald.cpp`

In `void Ewald::compute(int eflag, int vflag)`, here are the lines where the Coulomb force is being stored:

```
f[i][0] += qqrd2e*q[i]*ek[i][0];
f[i][1] += qqrd2e*q[i]*ek[i][1];
f[i][2] += qqrd2e*q[i]*ek[i][2];
```

(`f` was defined earlier in the program as `atom->f`. `qqrd2e` is a unit conversion factor, also incorporating the dielectric constant, which is the Coulomb energy between two charges q at distance r: 332.06371 in "real" units with dielectric constant 1. [See `force.h` and `force.cpp` and `update.cpp`.]) Modify it to:

```
f[i][0] += qqrd2e*q[i]*ek[i][0];
f[i][1] += qqrd2e*q[i]*ek[i][1];
f[i][2] += qqrd2e*q[i]*ek[i][2];
// my code: fcoul stores the Coulomb contribution to force
atom->fcoul[i][0] += qqrd2e*q[i]*ek[i][0];
atom->fcoul[i][1] += qqrd2e*q[i]*ek[i][1];
atom->fcoul[i][2] += qqrd2e*q[i]*ek[i][2];
// end my code
```

Next, if you want to use slab correction (`kspace_modify slab` command in LAMMPS), you do the same thing in `void Ewald::slabcorr(int eflag)`:

```
for (int i = 0; i < nlocal; i++) f[i][2] += qqrd2e*q[i]*ffact;
// my code
for (int i = 0; i < nlocal; i++) atom->fcoul[i][2] += qqrd2e*q[i]*ffact;
// end my code
```

`dump_custom.h`

I just wanted to output the electric field, so the dump custom command is a good way to do it. We want to add `efieldx, efieldy, efieldz` to the list of atom attributes that you can dump. In this file, you need to add to the big list of `pack` (for example, after `void pack_fz(int);`):

```
//my code
void pack_efieldx(int);
void pack_efieldy(int);
void pack_efieldz(int);
//end my code
```

`dump_custom.cpp`

In `void DumpCustom::parse_fields(int narg, char **arg)`, there is a big list reading the input and calling the appropriate pack function, and you just add the new dump commands into it:

```
  ...
} else if (strcmp(arg[iarg],"fz") == 0) {
  pack_choice[i] = &DumpCustom::pack_fz;
  vtype[i] = DOUBLE;

  // my code
} else if (strcmp(arg[iarg],"efieldx") == 0) {
  pack_choice[i] = &DumpCustom::pack_efieldx;
```

```
  vtype[i] = DOUBLE;
} else if (strcmp(arg[iarg],"efieldy") == 0) {
  pack_choice[i] = &DumpCustom::pack_efieldy;
  vtype[i] = DOUBLE;
} else if (strcmp(arg[iarg],"efieldz") == 0) {
  pack_choice[i] = &DumpCustom::pack_efieldz;
  vtype[i] = DOUBLE;
  // end my code
```

Next, you need to add your new pack commands. Here is one of the ones in the original file...

```
void DumpCustom::pack_z(int n)
{
  double **x = atom->x;
  int nlocal = atom->nlocal;

  for (int i = 0; i < nlocal; i++)
    if (choose[i]) {
      buf[n] = x[i][2];
      n += size_one;
    }
}
```

We basically copy this format. So the new ones to add are:

```
// my code

void DumpCustom::pack_efieldx(int n)
{
  double **fcoul = atom->fcoul;
  int nlocal = atom->nlocal;

  for (int i = 0; i < nlocal; i++)
    if (choose[i]) {
      buf[n] = fcoul[i][0] / atom->q[i] / force->qe2f;
      n += size_one;
    }
}

void DumpCustom::pack_efieldy(int n)
{
  double **fcoul = atom->fcoul;
  int nlocal = atom->nlocal;

  for (int i = 0; i < nlocal; i++)
```

```
    if (choose[i]) {
      buf[n] = fcoul[i][1] / atom->q[i] / force->qe2f;
      n += size_one;
    }
}

void DumpCustom::pack_efieldz(int n)
{
  double **fcoul = atom->fcoul;
  int nlocal = atom->nlocal;

  for (int i = 0; i < nlocal; i++)
    if (choose[i]) {
      buf[n] = fcoul[i][2] / atom->q[i] / force->qe2f;
      n += size_one;
    }
}

// end my code
```

Note that `force->qe2f` is a unit-conversion factor: 23.060549 in "real" units (see `update.cpp`). It converts from a charge ("q") times an electric field ("e") to ("2") a force ("f"). Therefore this code will output an electric field in whatever unit system you are using – volts per angstrom in the case of "real" units.

When compiling:

I found that my modified code had a "heisenbug": When I compiled it normally, the program crashed, but when I compiled it in debugging mode, it worked fine. I do not know what the bug is. Instead, I just used the debugging-mode program for my calculations.

In your LAMMPS script file:

The code that goes into the LAMMPS script is something like:

```
dump SteveDump all custom 10 stevedumpfile.xyze &
                    id type x y z efieldx efieldy efieldz
#10 means every 10 timesteps
#Note that atoms are outputted in random and inconsistent order.
```

Again, remember to turn off newton:

```
newton off
```

and remember that if you want to include the electric field due to directly-bonded atoms, you need

```
special_bonds coul 1.0 1.0 1.0
```

or you need to edit `pair_lj_cut_coul_long.cpp` differently. Note that the `special_bonds` command can be overwritten by other pair commands if you are not careful.

3 Phase plate for nonlinear chirp compensation

3.1 Overview

Mode-locked lasers are widely used to produce ultrashort light pulses (in the femtosecond range), for use in science and industry [1]. To get the pulse as short as possible, it is important to perform "dispersion compensation" [1,2]. The goal of dispersion compensation is to make the different light-frequencies in the pulse all have the same phase. Poor dispersion compensation (called "chirp") increases the pulse length. Perfect dispersion compensation gives the "transform-limited" pulse; i.e., the shortest possible pulse for a given spectral intensity profile.

Dispersion compensation is especially important in mode-locked *fiber* lasers, which suffer much larger dispersion than other types of mode-locked lasers (e.g., Ti:Sapphire lasers), primarily because of self-phase modulation in the fiber. It is also very important in few-cycle pulse generation, where dispersion must be canceled very accurately, and in chirped-pulse-amplification, where an enormous amount of chirp must be accurately canceled. It is also important in fiber-optic telecommunication—Uncompensated nonlinear chirp can limit the data-transmission rate [3,4]

The dispersion of the pulse is characterized by the spectral phase function (phase as a function of frequency). If this function is constant or linear, the light has the ideal (transform-limited) pulse length. In reality, unfortunately, the function will not be linear, but instead it will have a quadratic component, a cubic component, a quartic component, etc. In general these components get smaller and smaller at higher and higher orders. Therefore, the most important aspect of dispersion compensation is to eliminate the quadratic component, which is also called "linear chirp" or "group velocity dispersion"; the second-most-important aspect is to eliminate the cubic component ("quadratic chirp"); the third-most-important is to eliminate the quartic component ("cubic chirp"); etc.

Compensating for linear chirp is a well-developed field. (Such techniques include Treacy grating pairs, prism pairs, fiber Bragg gratings, chirped mirrors, dispersion-compensating fibers, etc. [1,2].) Compensating for quadratic and higher-order chirp, however, is much less developed, and there is no solution that is inexpensive, effective, and easy to adjust and optimize.

In this study, we investigate a new optical component that promises to be an inexpensive and convenient method to simultaneously and independently compensate linear, quadratic, and cubic chirp, and thus to allow shorter pulses from mode-locked lasers, especially those using chirped-pulse amplification. The rest of the chapter will proceed as follows. In Sec. 3.2, we will give background about nonlinear chirp compensation, and how this device relates and compares to other chirp compensation systems. In Sec. 3.3, we will discuss the plate design and the motivating theoretical analysis. In Sec. 3.4, we will discuss how the plate was made. In Sec. 3.5 we will discuss how Frequency-Resolved Optical Gating (FROG) was used to characterize the ultrafast pulses, and hence to measure the plate's effectiveness. Finally,

in Sec. 3.6 we discuss the results of this characterization, and in Sec. 3.7 we discuss future work.

3.2 Background on pulses, dispersion, and chirp

The electric field of a light-pulse can be written [1]:

$$E(t) = \int d\nu A(\nu) \,\mathrm{Re}\left[e^{-2\pi i\nu t} e^{i\phi(\nu)}\right] \tag{3.1}$$

where ν is frequency, $A(\nu) > 0$ is the real spectral amplitude (i.e., the square root of spectral intensity), and $\phi(\nu)$ is the light-phase as a function of frequency.

The uncertainty principle states that [1]

$$(\Delta t)_{\mathrm{rms}} \times (\Delta\nu)_{\mathrm{rms}} \geq \frac{1}{4\pi} \tag{3.2}$$

where $(\Delta t)_{\mathrm{rms}}$ is the RMS width of the time-domain pulse $I(t) \equiv |E(t)|^2$, and $(\Delta\nu)_{\mathrm{rms}}$ is the RMS spread of the frequency-domain spectral intensity $A(\nu)^2$. Therefore, there are two aspects to achieving a short pulse of light: Broaden the spectrum, and make sure that the uncertainty product (Eq. (3.2)) is not too far above its minimum $1/4\pi$. We focus on the second aspect, taking the spectrum $A(\nu)$ to be fixed by external constraints (such as the gain bandwidth of the laser).

The phase $\phi(\nu)$ in Eq. (3.1) significantly affects the pulse length, even without changing $A(\nu)$. If ϕ is constant, then the pulse has a peak at $t = 0$, and is moreover this pulse is *transform-limited*, i.e. the highest possible intensity and narrowest width for the given $A(\nu)$.

If the phase $\phi(\nu)$ is linearly varying, the pulse is still transform-limited, but peaks at a time $t \neq 0$. This is because a shift of the time coordinate has the effect of adding a linear offset to the phase function, $\phi(\nu) \to \phi(\nu) + 2\pi\nu\Delta t$. However, if $\phi(\nu)$ varies quadratically with ν, the pulse has *linear chirp*, and lasts longer, with a lower peak intensity, then the transform-limited pulse. If $\phi(\nu)$ varies as a cubic function of ν, the pulse has *quadratic chirp*; if $\phi(\nu)$ is a quartic function, then the pulse has *cubic chirp*, and so forth. The origin of this strange-seeming terminology is that, in nth-order chirp, the *arrival time* of a given frequency component of the wave is an nth-order function of frequency. (Arrival time is the first derivative of phase.) (Warning: The terminology in this field is not universally consistent [2].) The quadratic, cubic, etc. components of chirp are referred to generically as *nonlinear chirp*.

Linear chirp is often corrected via a Treacy prism-pair or grating-pair (Fig. 3.1). In this configuration, the lower-frequency light travels farther, and therefore gets delayed relative to higher-frequency light. This increases the "anomalous dispersion" or decreases the "normal dispersion" of the light. To alter the linear chirp in the opposite manner—i.e., decrease the anomalous dispersion or increase the normal dispersion—one can simply pass the light

through glass, or any other transparent medium, of appropriate thickness. Alternatively, a Treacy grating-pair can be altered to provide normal dispersion by appropriately inserting one or more lenses in the optical path. [1]

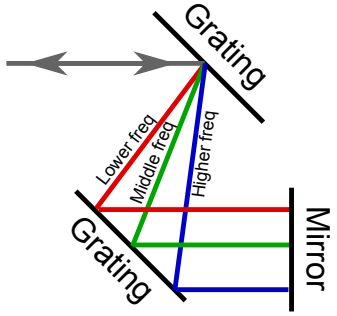

Figure 3.1 – Schematic of Treacy grating pair.

Unlike linear chirp, however, nonlinear chirp is quite difficult to correct. Ideally, a nonlinear chirp correction system would have the following desirable properties:

- Low cost
- High throughput (low loss by scattering, absorption, etc.)
- High damage threshold
- Easy alignment
- Easy tunability (As the laser is running, the pulse character may slightly change over time, for example if the laser power is adjusted, or if there is uncontrolled environmental instabilities; therefore it should be easy to adjust the chirp correction.)

Nonlinear-chirp-correction systems available to date satisfy some of these criteria to various extents, but none is ideal. We will next describe the systems available today, before discussing the new phase plate.

Among the most powerful and common nonlinear chirp correction systems is the programmable spatial light modulator, particularly liquid-crystal (LC) and acousto-optic (AOM) modulators [1,5]. Both are incorporated into a prism or grating structure, in a position where different light frequencies pass through different parts of the device (cf. Fig. 3.2(b)). The phase and/or amplitude of each frequency component is controlled separately, via computer input. This allows pulse compression (cancellation of linear and nonlinear chirp) [6], and also flexible pulse shaping [5]. These have simple alignment and tunability, but their disadvantages are high price, low throughput (especially for AOM), and low damage threshold (especially for LC).

Deformable mirrors [5] can also function as spatial light modulators and pulse compressors, with high throughput and high damage threshold. These have simple alignment, reasonable tunability, high throughput, and high damage threshold, but quite high cost.

Transparent plates with controlled thicknesses have been used to compensate quadratic chirp [3] and arbitrary chirp [7], functioning like a non-adjustable version of a programmable spatial light modulator. These have low cost, high throughput and damage threshold, and easy alignment. Its only shortcoming is the lack of tunability. The new phase plate described in this work is a variant of this design, but designed to be tunable.

As an alternative nonlinear chirp correction system, the stretcher and compressor of a chirped-pulse amplifier can be modified in many ways to alter nonlinear chirp, including putting in extra adjustable lenses, other optical elements, rotating the grating angles or using non-uniform grating grooves, using an "Offner triplet stretcher" with multiple adjustable mirrors, and so on [2]. In all cases the disadvantages are (1) More optical components (especially adjustable optical components) mean more complexity, more difficulty in alignment, more cost, and usually more optical loss; (2) Adjustment to alter the chirp may require realignment.

Another proposal along these lines is the "grism" (prism with grating etched into it), which can compensate linear and quadratic chirp [2,8,9]. The same disadvantages mentioned above apply here; additionally, in this design, the pulse must travel through prism glass after being completely compressed (spatially and temporally), which limits high-peak-power applications [9].

There are many other ways to control nonlinear chirp, that are effective (at least over small bandwidths [10]) but not tunable. They include thin-film structures (chirped mirrors, Gires-Tournois interferometers [1, 2] and similar multilayer-thin-film structures [11]) and fiber-optic techniques (using specially-constructed fibers [2] or fiber Bragg gratings [12, 13]). These lack tunability, and therefore are not very effective in cases where the dispersion is not known beforehand, or varies based on operating conditions. If they have any tunability at all, it is very limited—typically involving just one degree of freedom, such as a mirror spacing [14], a temperature, or a strain [15]. This compares unfavorably to the three degrees of freedom of the device described in this work.

3.3 Plate design and theoretical analysis

The plate is made from fused silica, with cross-section shown in Fig. 3.2(a), for the case of three step-heights. (The plate can also have two, four, or any number of step heights.) The plate is placed into a Treacy grating pair (or a similar setup) such that different light-frequencies pass through different parts of the plate (Fig. 3.2(b)).

The motivation is as shown in Fig. 3.3. The three step heights add relative phases of 0, $2\pi/3$, and $-2\pi/3$, with cubed-root-spaced step widths. This combination can approximately cancel out a quadratic-chirp phase profile, leaving the phase everywhere close to zero, corre-

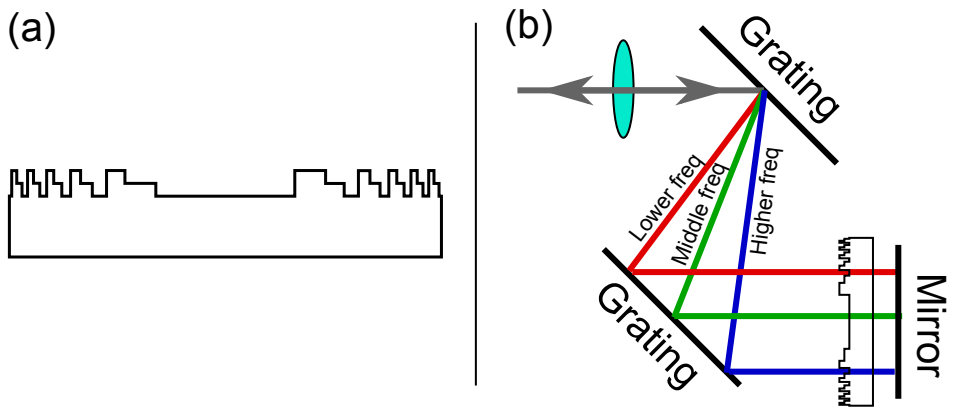

Figure 3.2 – (a) Schematic cross-section of the phase plate. (b) Phase-plate incorprotad into a Treacy grating pair. The cylindrical lens (top left), with focal point at the mirror, helps avoid spatial distortion of the pulse, as described in the text.

sponding to a sharply-peaked pulse. A simulation is shown in Fig. 3.4. Assuming the original pulse has a quadratic chirp that roughly doubles its FWHM, the plate can bring the pulse to near its transform-limited intensity and length. In this simulation, it is assumed that the pulse has center wavelength 1030nm and Gaussian spectral intensity profile with FWHM 7.8nm (corresponding to a transform-limited pulse FWHM of 200fs). It is assumed to have 10^7fs^3 of quadratic chirp, which increases the FWHM to about 500fs. Notice in Fig. 3.4 that the area under the curve of the corrected pulse is somewhat lower than the original. That extra intensity was pushed into a long, but quite weak, tail, not shown in Fig. 3.4.

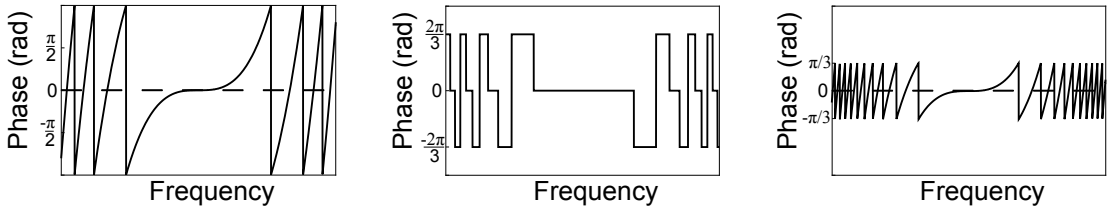

Figure 3.3 – Left: Phase profile of original pulse, assumed to have quadratic chirp. Center: The extra phase added on by the plate, on account of different frequencies passing through different thicknesses of SiO_2. Right, the final phase of the pulse stays close to zero, corresponding to a pulse with a sharp peak at time zero.

In its cross-section, the plate is somewhat similar to static phase masks discussed in the literature [3, 7]. Equally important, however, is the top view of the mask, designed to make the plate tunable, as shown and explained in Figs. 3.5–3.6. In theory, it should be possible to independently tune the compensation of linear, quadratic, and cubic chirp, simply by moving the plate, which would be mounted on a translation-and-rotation stage.

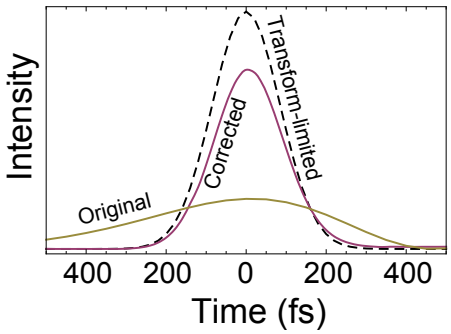

Figure 3.4 – Simulation of pulse shortening using the phase plate. See text for detailed parameters of simulation.

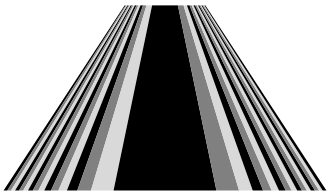

Figure 3.5 – Top view of plate. Each color represents a different thickness.

3.4 Methodology for making plate

The plate was made by evaporating SiO_2 or Al_2O_3 onto a 500μm-thick fused silica plate. In this work, all plates used the three-step-height design, which required two evaporations through two different masks. These metal evaporation masks were custom-ordered from Photo Etch Technology, and were fixed to the plate with wax and aligned by hand in a microscope. The deposition thickness was calibrated and checked by either scanning-contact-probe profilometry or transmission spectrum measurements. (The latter was only possible for Al_2O_3 depositions.)

Two sets of masks were made, each using the pattern in Fig. 3.5. The first was 9 cm along the bottom, 2.25 cm along the top, and 6 cm tall; the second was scaled down by a factor of three. The second (smaller) phase plate was used for the work described below, as its size was a better match for the spread of light within the Treacy grating pair of the laser under test.

3.5 Frequency-resolved optical gating

The phase plate is designed to reduce the duration, and increase the peak power, of the pulse. To test its effectiveness, we used the Frequency Resolved Optical Gating (FROG)

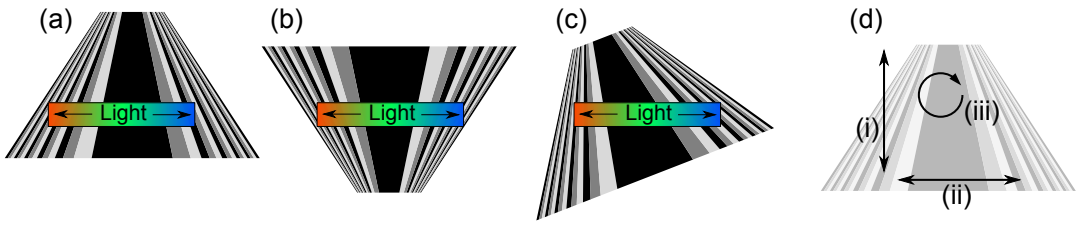

Figure 3.6 – Three possible ways of moving the plate relative to the grating-dispersed light: (a) Small positive quadratic chirp; (b) Large negative quadratic chirp; (c) Intermediate positive quadratic chirp, plus positive cubic chirp. (d) More generally, by moving the plate as shown, one can separately compensate (i) quadratic, (ii) linear, and (iii) cubic chirp.

technique [16, 17] to characterize the pulse, with and without the phase plate.

3.5.1 FROG overview

The Second-Harmonic-Generation Frequency Resolved Optical Gating (SHG-FROG) technique [1, 16, 17] is one of the standard methods for completely characterizing ultrafast light-pulses. The basic setup is as shown in Fig. 3.7. It involves splitting the pulse into two copies using a Michelson interferometer, then passing the combined pulse through an SHG crystal, and measuring the result with a spectrometer. The resulting dataset is two-dimensional: Intensity as a function of both frequency and mirror position (i.e., delay between the pulses).

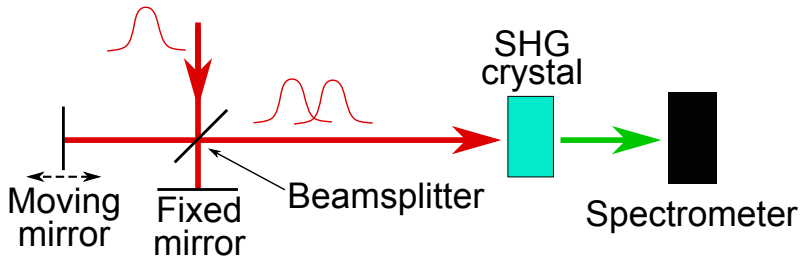

Figure 3.7 – Simplest FROG setup

If the (complex) electric field of the pulse is $E(t)$, then the FROG data is given by:

$$I_{\mathrm{FROG}}(\nu, \tau) = \left| \int_{-\infty}^{\infty} E(t) E(t-\tau) e^{-2\pi i \nu t} dt \right|^2 \tag{3.3}$$

where ν is the SHG light frequency and τ is the delay between the two copies of the pulse. If the delay τ is significantly longer than the length of the pulse, then the two copies of the pulse will not overlap, so there is no SHG between them, and I_{FROG} approaches zero. (The

SHG from within a single copy of the pulse is suppressed in the measurement, as discussed below.)

In the setup shown schematicaly in Fig. 3.7, the FROG measurement is a many-shot measurement: One measurement for each delay as the moving mirror is scanned. However, the same data can be obtained in a one-shot or few-shot way as described below.

3.5.2 Initial FROG setup (GRENOUILLE)

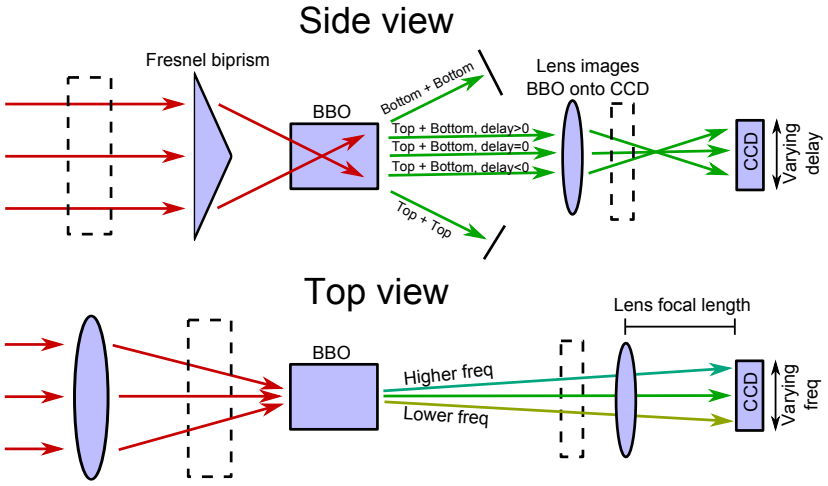

Figure 3.8 – Overview of GRENOUILLE setup. All lenses are cylindrical.

Initially, the measurement was set up using GRENOUILLE (GRating-Eliminated Nonsense Observation of Ultrafast Incident Laser Light E-fields) [18, 19], as shown schematically in Fig. 3.8. Although the experiment was eventually switched to a different configuration, we will nevertheless explain how GRENOUILLE works, as it is helpful for understanding the final setup. GRENOUILLE sets up SHG-FROG as a one-shot measurement, with different delays along one axis and different frequencies along the other.

The different delays are set up by passing the light through a Fresnel biprism (glass prism with large apex angle). This splits the pulse into downward-traveling and upward-traveling sub-pulses. At the top of the BBO, the upward-traveling pulse had to travel a longer distance, so is delayed relative to the downward-traveling pulse; at the bottom of the BBO, it is the opposite. Therefore, the SHG at different delays is spread to different vertical parts of the crystal. A cylindrical lens focuses the different vertical parts of the crystal onto different pixel-rows of the CCD, so that this can function as the delay axis. Conveniently, the same setup ensures that the SHG-FROG measurement is "background-free", i.e. the signal is zero when the delay is longer than the pulse-width. This is because the collected signal comes only from SHG generated by one upward-traveling and one downward-traveling

photon, and not from two downward-traveling or two upward-traveling photons, thanks to spatial filtering as shown in Fig. 3.8.

The different frequencies are resolved by taking advantage of phase-matching, which causes light of different SHG frequencies to come out more strongly at different angles. A lens ensures that the light enters at a spread of different angles; in order for all the different possible SHG frequencies to be visible, the lens is chosen to ensure a spread of incoming angles at least as wide as the spread of SHG output angles. A second lens converts the different angles into different positions on the CCD, thereby creating the frequency axis. Note that this sort of frequency spread due to phase-matching only occurs in the top view, not the side view, of Fig. 3.8. This is because of the relationship between the crystal's orientation and the light propagation direction. In the top-view, there is an approximately linear dependence of phase-matching angle and frequency. In the side-view, the angle varies quadratically with frequency, so in practice there is a negligible angle spread.

Initially, it was attempted to characterize pulses using GRENOUILLE, but the technique was unsuccessful as described in the next section.

3.5.3 Modified FROG setup

There were two problems that were discovered upon running pulses through the GRENOUILLE: Poor frequency resolution and insufficient temporal range. There is a recently-developed variant of GRENOUILLE tailored for large temporal range (up to 15ps) and correspondingly fine frequency resolution [20]; however, it could not be used due to time and budget constraints. Instead, the traditional GRENOUILLE described above was modified by incorporating elements of FROG. In this section, we describe these modifications.

First, we discuss the poor frequency resolution. This is illustrated in Fig. 3.9, showing the GRENOUILLE measurement on the left. (Poor frequency resolution corresponds to blurring in the left-right direction.) This problem was addressed by using a diffraction grating to spread out the SHG frequencies better; the results are shown on the right side of Fig. 3.9.

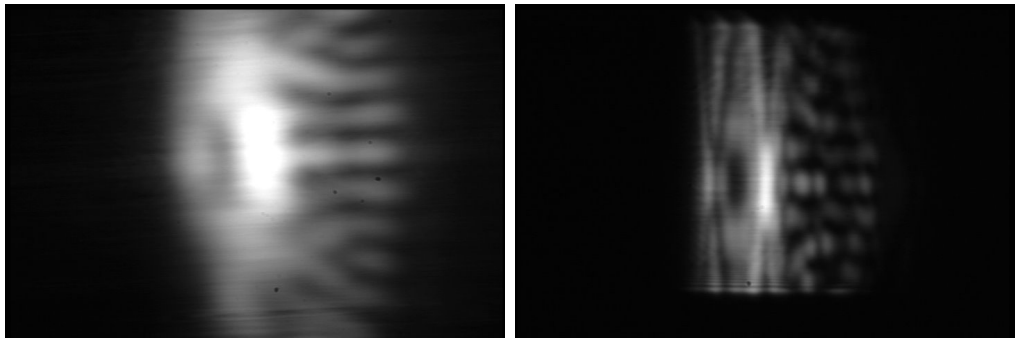

Figure 3.9 – Example image showing improvement in frequency resolution by using a grating (right) instead of the GRENOUILLE technique (left).

The cause of the poor resolution is the inadequate group velocity mismatch of the 8mm-thick BBO crystal used. As shown in Fig. 3.10, at the angle where 515nm light is maximally emitted (perfectly phase-matched), all the light in the range 514.5–515.5nm is also significantly emitted. As it turned out, this degree of frequency separation was inadequate for the specific laser tested.

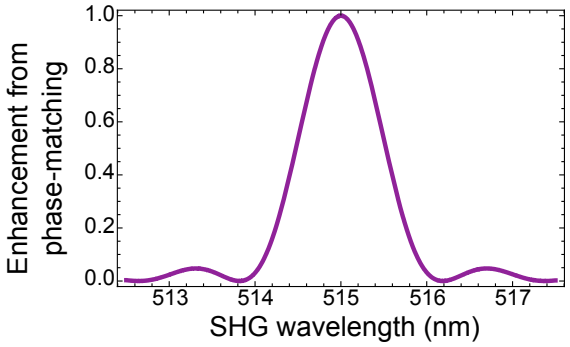

Figure 3.10 – Calculation illustrating the frequency separation when running GRENOUILLE with an 8mm-thick BBO SHG crystal. At the angle where 515nm light is perfectly phase-matched, this plot indicates that light at 514.5nm and 515.5nm is still reasonably well phase-matched.

The grating to improve frequency resolution was installed as shown in Fig. 3.11. A normal (round not cylindrical) lens was used to image the BBO onto the CCD, thereby throwing out the inadequately-resolved frequency spread due to phase matching. Instead, the frequency spread came entirely from the reflective grating. Other configurations were attempted, where frequency spread could come (in varying degrees) from *both* phase-matching *and* the grating, but it was found that the configuration of Fig. 3.11 gave the sharpest resolution.

The second problem discovered during GRENOUILLE testing was an insufficient range of temporal delay—for example, note the vertical cropping in Fig. 3.9. The GRENOUILLE was designed to cover a delay range from −1.3ps to +1.3ps, based on the presumed length of the pulse. It turned out that this was enough to cover the main peak, but not the tail, which was measureable up to a delay of ≈5ps. (For proper pulse reconstruction with FROG, it is important to get the whole FROG trace, including the tails, and not just the central peak [17].)

The temporal delay problem was solved by switching to a sort of FROG-GRENOUILLE hybrid. While GRENOUILLE is a one-shot measurement, and FROG is hundreds of shots, the technique pursued here required ten or so shots for a full pulse measurement. As shown in Fig. 3.11, a Micheleson interferometer was used to create two well-separated copies of the pulse. However, the interferometer was deliberately misaligned, to center one copy of the pulse in the bottom half of the biprism and the other in the top half. One or more thick glass plates delayed the lower half so that the two copies would arrive *approximately*

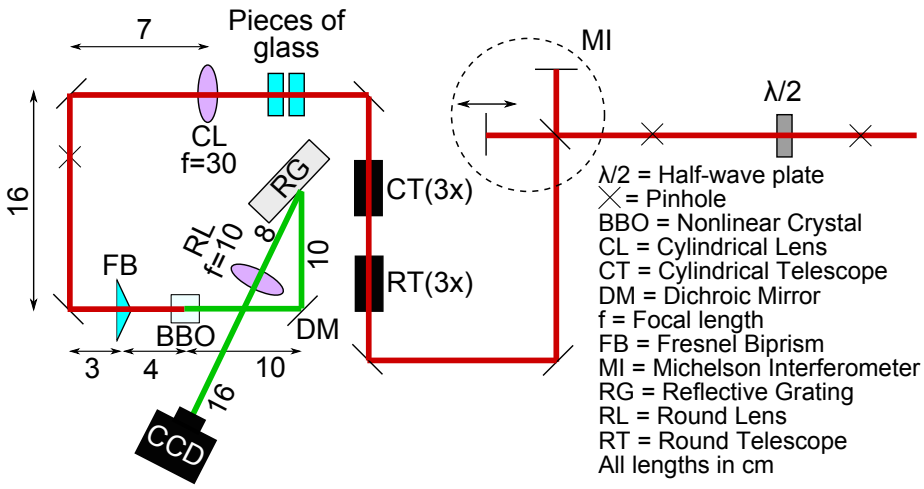

Figure 3.11 – Final FROG setup. All lengths in cm. The pieces of glass covered half the beam and were switched in and out to cover different parts of the delay range. The cylindrical telescope spread out the light vertically, while the cylindrical lens compressed the light horizontally (from left or right towards center). Each telescope consisted of an $f = -5$cm diverging lens spaced 10cm from an $f = +15$cm converging lens.

the same time, but by adjusting the interferometer, the exact delay could be varied. Just as in GRENOUILLE, each CCD image captured not just a single delay, but a range of various delays spanning ≈2.5ps. Adjusting the interferometer changed the central delay in the CCD image, so by stitching together a few images, the full range of FROG signal could be measured. In practice, however, many more CCD images than necessary were taken—as many as 35 images per pulse. This way, it was possible to average over the overlapping areas, and ignore the areas near the edges of the CCD. This made the images higher-quality and more uniform. Conveniently, it was only necessary to measure positive delays, not negative, because of the symmetry in the FROG image. An example composite FROG trace image, made by appropriately gluing and averaging a large number of individual images, is shown in Fig. 3.12. This is not quite the final image used in the FROG analysis; the last step (not shown) is to replace the cropped half with a mirror image of the higher-quality half.

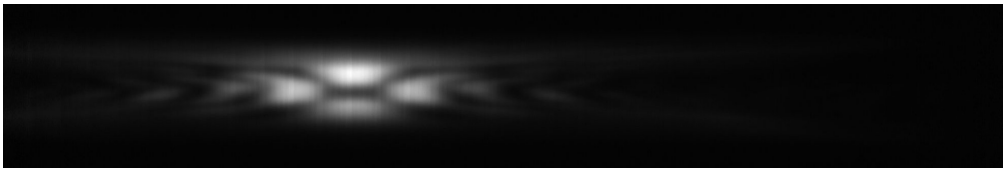

Figure 3.12 – An example FROG image (rotated 90°): Different delays correspond to different horizontal positions and different frequencies correspond to different vertical positions.

It may seem like unnecessary effort to make two very-well-separated pulses at the inter-

ferometer, then use glass plates to make them arrive at the same time. However, this was essential, as the pulse heading towards the top half of the biprism also unavoidable partly passed through the bottom half and vice-versa. The scheme used here prevented that effect from altering the data.

Another potential concern is pulse distortion upon passing through glass. However, a calculation using the typical dispersion of glass indicated that distortion would be negligible even for many-centimeters-thick glass, given the duration, bandwidth, and frequency of the pulse.

An important aspect of the experiment is calibration of the delay and frequency axis. The delay axis calibration is straightforward in this setup, and actually requires no extra work besides the data processing already required. Since the Michelson interferometer is on a calibrated translation stage, it was known how much the delay changed when the interferometer setting was changed. For example, in Fig. 3.12, images were captured each time the mirror was translated 50μm, which corresponds to a 0.333ps shift in delay between successive CCD images. Meanwhile, the image-gluing routines involved calculating the shift (in pixels) between successive CCD images, which was about 57 pixels in the case of Fig. 3.12. Therefore the calibration was that each pixel corresponded to 5.85fs delay.

For the frequency axis, the glass plates in Fig. 3.11 were removed while the two arms of the interferometer were set to various path differences rather close to zero. First the zero-path-difference point was measured; this is where the FROG image fluctuates between very bright and completely dark. Next, the arms were moved farther and farther apart. Since the interferometer acts as a sinusoidal frequency filter, there are stripes in the FROG image along the frequency axis, as shown in Fig. 3.13. In this example, the dark horizontal stripes are 650GHz apart, thanks to the interferometer mirrors being separated by 230μm. Computer analysis finds a 33-pixel separation between stripes. Therefore, the calibration in this case is that each pixel corresponds to a 20GHz change in SH frequency. This procedure was always repeated for two or three different interferometer positions as a consistency check. The system was re-calibrated each day, although it remained quite stable.

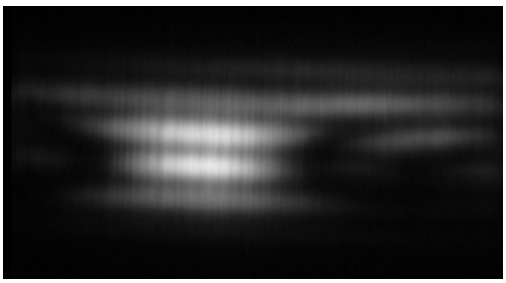

Figure 3.13 – An example image (rotated 90°) during a FROG frequency calibration. The dark horizontal stripes are 650GHz apart, thanks to the interferometer mirrors being separated by 230μm.

3.5.4 FROG algorithm

Since there is a nonlinear relation between a pulse and its FROG signal, a complex numerical algorithm is required to infer a pulse from its FROG signal. The algorithms used here were variants of the "power method principal-components generalized-projection algorithm" [21]. All code used is available online [22].

Before attempting to analyze the FROG signal, it was first pre-processed in a number of ways: It was smoothed by applying a top-hat filter in Fourier space [17], the background was subtracted off (where the background level was inferred from the lowest-average-intensity 8×8 block of pixels), and the image was downsampled from the original high-resolution CCD image to usually 256×256 pixels. The downsampling was performed by taking the delay- and frequency-calibration into account, so that in the final downsampled image, the time- and frequency intervals had a fast-Fourier-transform (FFT) relationship:

$$\Delta\tau \times \Delta\nu = \frac{1}{N}, \tag{3.4}$$

where $\Delta\tau$ is the delay-difference between pixels bordering each other vertically, $\Delta\nu$ is the optical frequency difference between pixels bordering each other horizontally, and $N \times N$ is the downsampled image size. This relationship is a prerequisite for the FROG analysis. The pixel count, usually 256×256, was chosen based on trial-and-error with the data: Lower resolution often had problems with image cut-off and detail loss, while higher resolution was both slow and prone to getting stuck in local minima. The exact value 256 was chosen because the FROG algorithm runs somewhat faster when N is a power of two, since then the FFTs are more efficient.

Following the pre-processing, the $N \times N$-pixel image is put into the main FROG algorithm. We will first describe the "power method principal-components generalized-projection algorithm" as described in Ref. [21], then below we will describe the variations used in this work.

We rewrite Eq. (3.3) for FROG intensity as follows:

$$E_{\mathrm{FROG}}(\nu,\tau) = \int_{-\infty}^{\infty} E(t)E(t-\tau)e^{-2\pi i\nu t}dt \tag{3.5}$$

$$I_{\mathrm{FROG}}(\nu,\tau) = |E_{\mathrm{FROG}}(\nu,\tau)|^2$$

where I_{FROG} is the actual FROG signal and E_{FROG} is the (complex) "FROG amplitude".

Broadly, the algorithm is as follows:

1. Start with a guess for the pulse field $E(t)$.
2. Plug $E(t)$ into Eq. (3.5) to get the guess for E_{FROG}.
3. Improve the guess for E_{FROG} by keeping the complex phase of each entry unchanged, but correcting the absolute value to agree with the experimentally-measured $\sqrt{I_{\mathrm{FROG}}}$.

4. Use the "power method" (described below) to generate a new pulse field $E(t)$ that agrees as closely as possible as possible with this guess of E_{FROG}.

The more detailed procedure for Step 2, going from $E(t)$ to its E_{FROG}, is as follows. Write the discretized $E(t)$ as $(E_1, E_2, \ldots, E_N)$. Then take the outer product of $E(t)$ with itself, rearrange the entries, and do a Fourier transform:

$$\begin{pmatrix} E_1E_1 & E_1E_2 & \cdots \\ E_2E_1 & E_2E_2 & \\ \vdots & & \ddots \end{pmatrix} \xrightarrow[\text{entries}]{\text{rearrange}} \begin{pmatrix} E_1E_1 & E_1E_2 & \cdots \\ E_2E_2 & E_2E_3 & \\ \vdots & & \ddots \end{pmatrix}$$

$$\xrightarrow[\text{column}]{\text{FFT each}} \left(\begin{array}{c|c|c} \uparrow & \uparrow & \\ E_{\mathrm{FROG}}(\nu, \tau = 0) & E_{\mathrm{FROG}}(\nu, \tau = \Delta t) & \cdots \\ \downarrow & \downarrow & \end{array} \right) \quad (3.6)$$

For Step 4, going from E_{FROG} to $E(t)$, one starts by doing the reverse of Eq. (3.6), and then the problem becomes to approximate an $N \times N$ matrix M, as accurately as possible, as the outer product of an N-entry vector v with itself: $M \approx v^T v$.

A related but more famous mathematical problem is "rank-one matrix approximation" (or more generally, low-rank matrix approximation), a problem with applications image-compression and elsewhere. This problem involves approximating an $N \times N$ matrix as the outer product of two *different* vectors, $M \approx u^\dagger v$. The solution, according to the Eckart–Young theorem, is to start with the singular value decomposition: $M = \sum_{j=1}^{N} u_j^\dagger \sigma_j v_j$, where σ_j are scalars (the "singular values") and u_j, v_j are normalized complex vectors. Take the largest (in absolute value) singular value σ_1, and then the closest possible outer-product approximation to M is $u_1^\dagger(\sigma_1 v_1)$.

Algorithms for calculating the SVD are implemented in all linear-algebra program libraries, and faster algorithms that find only the largest-absolute-value term are also common. As an alternative method, the "power method" can be used [21]: If u is an initial guess for u_1, then $MM^\dagger u$ is a better guess; likewise, if v is the initial guess for v_1, then $M^\dagger M v$ is a better guess.

Getting back to the real problem, we want an outer product of a vector with *itself*, $M \approx v^T v$. If M were symmetric, the Eckart–Young theorem would automatically give a decomposition of this form. However, M may not be symmetric. A few approaches are possible: M can be symmetrized (i.e., use $(M + M^T)/2$) before applying the power method; or just the left or just the right singular vector can be used; or, as proposed in Ref. [21], both the left and the right singular values can be separately computed, then averaged directly or indirectly in the next iteration of the algorithm. Tests showed that all these algorithms succeeded with roughly equal probability, so for convenience, the left singular vectors were used exclusively. More specifically, in Step 4, if u is the previous guess for the $E(t)$ vector and M is the E_{FROG} matrix in "outer-product form" (see Ref. [21]), the $MM^\dagger u$ was the next guess for $E(t)$.

Unfortunately, due to the complexity of the pulses, this algorithm converged rather unreliably. The algorithm was therefore modified and extended in a few ways as described next.

First, while Steps 2 and 4 work with the time-domain pulse $E(t)$ in Ref. [21], it is equally possible to use the frequency domain. More specifically:

$$\text{Time domain:} \quad E_{\text{FROG}}(\nu,\tau) = \int_{-\infty}^{\infty} E(t)E(t-\tau)e^{-2\pi i\nu t}dt \tag{3.7}$$

$$\text{Frequency domain:} \quad E_{\text{FROG}}(\nu,\tau) = \int_{-\infty}^{\infty} \tilde{E}(\nu')\tilde{E}(\nu-\nu')e^{-2\pi i\nu' t}d\nu' \tag{3.8}$$

where

$$\tilde{E}(\nu) = \int_{-\infty}^{\infty} E(t)e^{-2\pi i\nu t}dt, \qquad E(t) = \int_{-\infty}^{\infty} \tilde{E}(\nu)e^{2\pi i\nu t}d\nu$$

When finding the best-guesss E starting from E_{FROG}, one can do an outer-product approximation in either the time or the frequency domain. In fact, by switching back and forth, one winds up with a more robust procedure, less liable to become trapped in a local minimum.

Second, for complex pulses, the algorithm has a severe aliasing problem, i.e. the pulse "wraps around" connecting the beginning and end. (This "wrapping around" is due to the FFTs at the core of the algorithm.) In theory, one ought to use enough data that the field is very close to zero for the first 25% and last 25% of datapoints in the time-domain or frequency domain. However, for the complex FROG traces in this work, this would have required a huge number of datapoints—as many as 2048×2048 pixels—which would make the algorithm far too slow to use.

Part of the problem is that the time-domain and frequency-domain step size need to satisfy the FFT relation, Eq. (3.4). In order to pad the outer 25% of each side of the FROG trace with zeros (in both time and frequency direction), the number of pixels must increase by a factor of sixteen.

Although a fully-padded FROG trace is certainly ideal [17], one nevertheless wants to have an algorithm which is reasonably robust to a FROG trace approaching the edges of the available pixels. For this reason, a novel anti-aliasing algorithm was used. Four variants were used in various combinations. First, the anti-aliasing could occur during Step 2 or Step 4. Second, the anti-aliasing could occur in the time-domain or in the frequency-domain. The time-domain anti-aliasing is functionally very similar to padding the signal with zeros in the time-domain, while the frequency-domain anti-aliasing is similar to padding with zeros in the frequency domain.

For the exact implementation, consider the time-domain pulse data $(E_1, E_2, \ldots)$, and the

"outer product form" matrix from Eq. (3.6):

$$\begin{pmatrix} E_1E_1 & E_1E_2 & \cdots & E_1E_N \\ E_2E_1 & E_2E_2 & & E_2E_N \\ \vdots & & \ddots & \vdots \\ E_NE_1 & E_NE_2 & \cdots & E_NE_N \end{pmatrix}$$

The entries in the top-right and bottom-left parts of this matrix are the product of the field very early in the pulse with the field very late in the pulse, which would correspond to a large delay between the two copies of the pulse. Such a large delay, in fact, falls outside the range of delays measured in the input FROG trace ($-N/2$ to $+N/2$). They are only present due to aliasing. Therefore it makes more sense to set these values to zero.

Similarly, consider the frequency-domain outer product form matrix:

$$\begin{pmatrix} \tilde{E}_{-N/2}\tilde{E}_{-N/2} & \tilde{E}_{-N/2}\tilde{E}_{-N/2+1} & \cdots & \tilde{E}_{-N/2}\tilde{E}_{+N/2} \\ \tilde{E}_{-N/2+1}\tilde{E}_{-N/2} & \tilde{E}_{-N/2+1}\tilde{E}_{-N/2+1} & & \tilde{E}_{-N/2+1}\tilde{E}_{+N/2} \\ \vdots & & \ddots & \vdots \\ \tilde{E}_{+N/2}\tilde{E}_{-N/2} & \tilde{E}_{+N/2}\tilde{E}_{-N/2+1} & \cdots & \tilde{E}_{+N/2}\tilde{E}_{+N/2} \end{pmatrix}.$$

The entries in the top-left and bottom-right parts of this matrix are the product of very low frequency components with each other, or very high frequency components with each other. These correspond to very low and very high sum-frequencies. Such extreme frequencies, in fact, fall outside the range of frequencies measured in the input FROG trace. They are only present due to aliasing. Therefore it makes more sense to set these values to zero.

It is worth emphasizing that these are not meant to be particularly rigorous explanations; they oversimplify the coupling between time and frequency domain. Nevertheless, these modifications of the algorithm were found to be generally effective at improving the convergence and reliability.

As a final note, when multiple measurements were taken, the reconstruction of one pulse was sometimes used as the "seed" (initial guess) of the reconstruction of another pulse. However, this was used cautiously, cross-checking against randomly-seeded reconstructions, in order to be sure that any inferred similarities between pulses were true aspects of the pulses, not artifacts imprinted from using using the same seed.

3.5.5 Verification of FROG reliability

The accuracy and reliability of the FROG pulse characterization was verified by altering the Treacy grating-pair configuration. Based on the light frequency and diffraction angles, it was inferred that altering the separation between gratings by a distance d should change the linear chirp by $(2.89\,\mathrm{THz}^2\,\mathrm{cm}^{-1}) \times d$. Likewise, it should not affect the spectrum at all. Both of these expectations are convincingly demonstrated in Fig. 3.14.

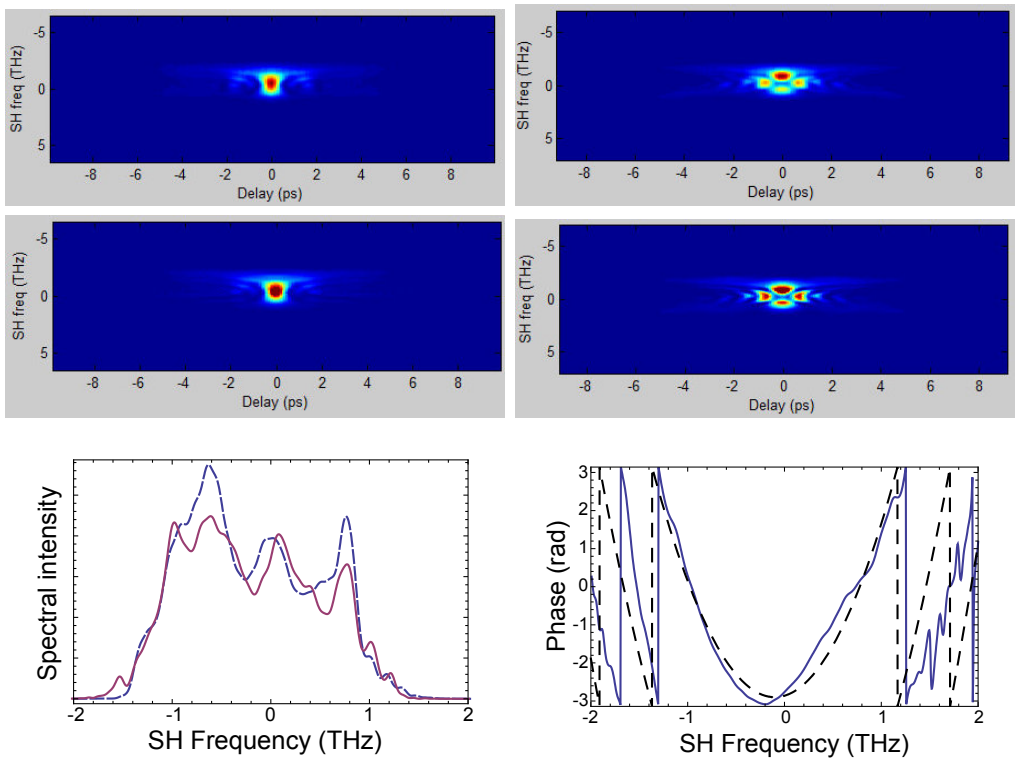

Figure 3.14 – FROG traces at two different Treacy settings. Top: Experimental data. Center: Best fit. Bottom left: As expected, the intensity spectrum is almost the same for the two Treacy settings. Bottom right: The difference in phase between the two settings is a parabola as expected. Note: The dashed line is the best-fit parabola, *not* the *ab initio* expected parabola. However, the two parabolas are consistent within the experimental uncertainty, which was due to difficulties in precisely measuring the grating-grating distance change.

3.6 Demonstration of pulse shortening with phase plate

The pulse from the laser had a weak low-frequency wing; when the Treacy compressor was set to optimize the main pulse, the wing arrived many picoseconds later. A razor-edge in the compressor was used to cut off the wing, in order to better focus on the main pulse.

3.6.1 Configuring plate in laser

An important requirement, discovered early in testing, is the lens shown in Fig. 3.2(b). The cylindrical lens is supposed to focus light on the plate. In a double-pass configuration, it is impossible to achieve perfect focus in both passes; nevertheless, by putting the focus at the mirror, using a long focal length, and keeping the plate near the mirror, the requirement is adequately met.

The purpose of the lens is to reduce spatial distortion of the pulse—the phenomenon where the left half of the light spot, for example, might have a very different length and chirp profile than the right half. This phenomenon was strikingly visible in FROG, as spatially shifting the beam would change which part of the spot was measured by the FROG, and could dramatically alter the FROG trace. Another clear indication was a FROG trace lacking the appropriate symmetry between positive and negative delay. This could occur because the positive and negative delay parts of the signal in the FROG trace are generated from different (spatial) parts of the beam.

Spatial distortion results from the plate when a single frequency overlaps one or more step-edges. Depending on the exact propagation direction of the light (within the beam), the energy passing through the two sides of the step-edge may interfere constructively or destructively. In particular, the laser entered the Treacy grating pair as a collimated beam about 1mm in diameter. Without a lens, any given frequency component will maintain about a 1mm spread, which meant that, in a typical mask position, a large proportion of light frequencies passed through two or even more different step heights, thus getting spatially redirected.

Due to space constraints, the one-lens setup shown in Fig. 3.2(b) was not used; instead, one lens was used to focus, and a different lens was used to re-collimate.

3.6.2 Pulse characterization and theoretical improvement

The phase and intensity of the original pulse in both time and frequency domain are shown in Fig. 3.15, when the Treacy was set to a near-optimal position (in terms of minimal autocorrelation FWHM). Using this data, it is possible to calculate theoretically how much improvement might be possible if the linear and quadratic chirp were *ideally* corrected. Mathematically, the phase was mathematically modified by adding every possible combination of linear and quadratic chirps, and the resulting pulses were ranked by some measure of time-domain pulse sharpness. These tests indicated that it should be possible to only modestly improve the pulse by adjusting linear and quadratic chirp: The FWHM can be reduced by about 25% (but the "improved" pulse actually has much larger wings); the RMS pulse-length can be reduced by about 15%, and the inferred SHG from the pulse can be increased by about 40%, according to these calculations. (The possible improvement including ideal cubic chirp correction was only slightly better.) These modest improvements can be contrasted with the dramatic improvement in Fig. 3.4 if the pulse truly starts with quadratic chirp, rather than the much-higher-order chirp of this pulse. The fundamental reason is that the frequency-domain phase (see bottom of Fig. 3.15) is already quite flat where the intensity is high. At the same time, the part where the pulse could be improved, i.e. on the wings where the phase is not flat, the phase varies so suddenly and rapidly that a quadratic or even cubic chirp cannot well describe it. (This pulse could be sharpened in a more traditional way by filtering its high and low frequency sides, but this comes at the cost of reduced intensity.)

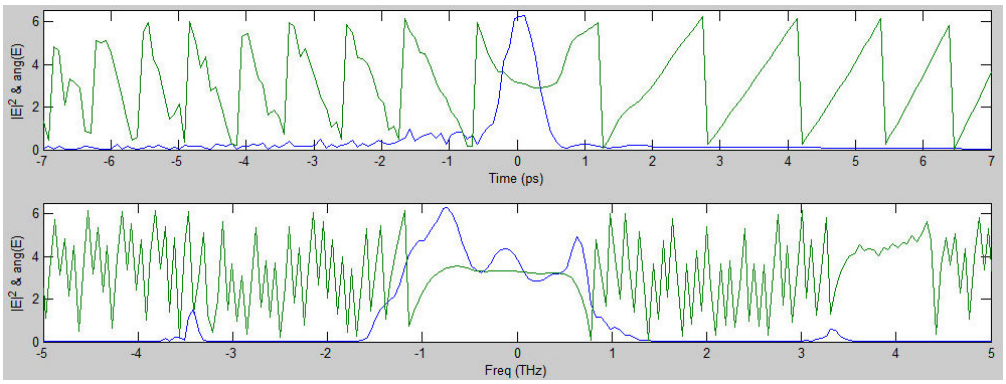

Figure 3.15 – Pulse characterized by FROG without the phase plate. Top: Time domain. Bottom: Frequency domain. Blue curves: Intensity. Green curves: Phase.

3.6.3 Measured pulse improvement due to plate

As mentioned, the chirp profile of the laser under test was not a good match to the plate design, consisting mainly of higher-than-cubic-order chirp. Nevertheless, we confirmed that the plate was functioning as expected in the laser, and we did in fact confirm a modest improvement in the laser pulse from the properly-positioned plate.

Fig. 3.16 confirms that the plate has the expected detailed effect on the pulse phase. Moreover, the plate can, in fact improve the pulse in line with the modest expectations described in Sec. 3.6.2: As shown in Fig. 3.17, the plate increases the peak intensity by about 12%, and decreases FWHM by 20%.

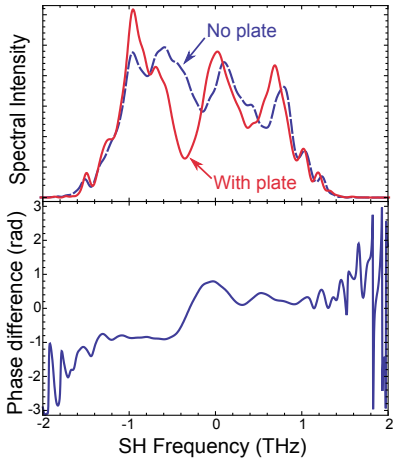

Figure 3.16 – Phase plate alters the pulse exactly as expected. In this case, the plate was put so that there was one step-edge in the middle of the beam. As expected, the intensity went somewhat down at the step-edge (due to spatial pulse distortion, see Sec. 3.6.1), while the phase jumps by $2\pi/3$.

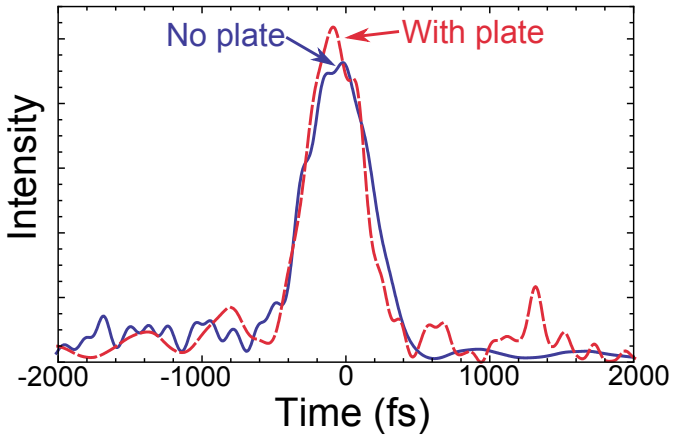

Figure 3.17 – Compared to the best pulse achievable without the plate, the plate here has increased the peak intensity and reduced the width. (This plot compares the best pulse obtained without the plate to the best pulse obtained with the plate; these were actually at different Treacy settings. Therefore the "no plate" measurement here is a different pulse than the one in Fig. 3.16.)

3.7 Conclusions and future work

For reasons described in Sec. 3.6.2, the detailed quadratic-chirp-canceling design of the plate was not specifically useful for improving the pulse of this particular laser. In fact, the data shown in Figs. 3.16–3.17 had just one step-edge in the beam. Nevertheless, the phase modification shown in Fig. 3.16 is a gratifying validation of the underlying idea, the plate construction, and the FROG methodology for validation.

Testing additionally drew attention to the importance of spatial pulse distortion associated with the step-edges. Although this is mitigated by the lens (see Fig. 3.2(b)), it could not be eliminated particularly well, because geometric constraints made it impossible to put the plate very close to the mirror. (Unless the plate is right at the mirror, it is impossible to have the light frequencies focused on the plate in *both* of the two passes of light.) This frustrated efforts to investigate the effect of the plate where the step-edges were too close together: A pulse with complex spatial distortion cannot be analyzed by FROG. In future work, a number of improvements are possible: the Treacy grating compressor could be redesigned to allow the plate to be placed right next to the mirror; the steps could be deposited directly onto the mirror rather than using a separate plate; a single-pass configuration could be used; or at least, a very long focal length lens could increase the confocal parameter and allow both passes to be better in focus.

More importantly, in future work, we plan to use a different laser, where the chirp profile is a better match to the design of the plate. This will allow a clearer test of the extent to which the plate is capable of improving pulse quality. Work in that direction is ongoing.

3.8 References

[1] Weiner, A. M. *Ultrafast Optics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, (2009).

[2] Walmsley, I., Waxer, L., and Dorrer, C. *Rev. Sci. Instrum.* **72**, 1 (2001).

[3] Tsuda, H., Takenouchi, H., Hirano, A., Kurokawa, T., and Okamoto, K. *J. Lightwave Technol.* **18**, 1139–1147 (2000).

[4] Wang, X., Kikuchi, K., and Takushima, Y. *IEICE Trans. Electron.* **E82-C**, 1407 (1999).

[5] Weiner, A. M. *Rev. Sci. Instrum.* **71**, 1929 (2000).

[6] Jiang, Z., Yang, S., Leaird, D. E., and Weiner, A. M. *Opt. Lett.* **30**, 1449–1451 (2005).

[7] Gaudiosi, D. M. and Akbulut, M. U.S. patent 7903326: Static phase mask for high-order spectral phase control in a hybrid chirped pulse amplifier system (2011).

[8] Kane, S. and Squier, J. *IEEE J. Quantum Elect.* **31**, 2052–2057 (1995).

[9] Kane, S. and Squier, J. *J. Opt. Soc. Am. B* **14**(3), 661–665 (1997).

[10] Jain, A. R., U.S. patent 6791736: Optical device for dispersion compensation (2004).

[11] Liu, J. and Chao, S. *Appl. Optics* **43**, 3442–3448 (2004).

[12] Durkin, M., Ibsen, M., Cole, M., and Laming, R. *Electron. Lett.* **33**, 1891–1893 (1997).

[13] Gualda, E. J., Gómez-Pavón, L. C., and Torres, J. P. *J. Mod. Optic.* **52**, 1197 (2005).

[14] Jablonski, M., Takushima, Y., and Kikuchi, K. *J. Lightwave Technol.* **19**, 1194–1205 (2001).

[15] Fells, J., Kanellopoulos, S., Bennett, P., Baker, V., Priddle, H., Lee, W., Collar, A., Rogers, C., Goodchild, D., Feced, R., Pugh, B., Clements, S., and Hadjifotiou, A. *IEEE Photonic Tech. Lett.* **13**, 984–986 (2001).

[16] Trebino, R., DeLong, K. W., Fittinghoff, D. N., Sweetser, J. N., Krumbügel, M. A., Richman, B. A., and Kane, D. J. *Rev. Sci. Instrum.* **68**, 3277–3295 (1997).

[17] Trebino, R. *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*. Kluwer Academic Publishers, Boston, (2002).

[18] O'Shea, P., Kimmel, M., Gu, X., and Trebino, R. *Opt. Lett.* **26**, 932–934 (2001).

[19] O'Shea, P., Akturk, S., Kimmel, M., and Trebino, R. *Appl. Phys. B* **79**, 683–691 (2004).

[20] Cohen, J., Lee, D., Chauhan, V., Vaughan, P., and Trebino, R. *Opt. Express* **18**, 17484–17497 (2010).

[21] Kane, D. J. *IEEE J. Quantum Elect.* **35**, 421–431 (1999).

[22] Byrnes, S. J. Frequency Resolved Optical Gating (FROG) (MATLAB file exchange ID 34986), `http://www.mathworks.com/matlabcentral/fileexchange/34986-frequency-resolved-optical-gating-frog`, (2012).

4 Field-effect photovoltaics

4.1 Background and overview

In this chapter, we discuss *Screening-engineered field effect photovoltaics*, a new architecture for solar cells. (Non-solar-related applications of the architecture are discussed briefly in Sec. 4.4.) In the remainder of this section, we give a motivation and overview, and in Sections 4.2 and 4.3 we describe two flavors of the architecture, including design principles, simulations, and experiments. We give some conclusions in Sec. 4.4.

4.1.1 Undopable materials

The most common basis of solar cells is the p-n junction. Unfortunately, lots of promising materials for these cannot be doped to both p- and n-type without degradation—or in some cases, cannot be doped to both p- and n-type at all. Some examples follow:

- Oxides, sulfides, and other II-VI semiconductors are promising solar cell materials due to their abundance and low materials processing costs [1]. Unfortunately, they tend to self-compensate, such that their doping cannot be reversed. For example, Cu_2O might be a promising solar cell candidate, but is natively p-type. Despite much effort, there is no convincing evidence that n-type Cu_2O has ever been made; indeed, there is evidence that n-type doping is impossible [2]. Likewise, ZnO would be a promising candidate for UV-LEDs, but cannot be made p-type (excepting rare reports of low-quality materials) [3]. Another important example is $In_xGa_{1-x}N$, which is natively n-type. As x increases, p-type $In_xGa_{1-x}N$ becomes a progressively worse semiconducting material. This has been the primary roadblock to making efficient green diode lasers.
- Nanoparticle (quantum dot) films like PbS and PbSe are promising candidate materials for solar cells [4]; however, doping is quite difficult as impurities tend to segregate to the surfaces of the particles.
- Amorphous silicon (a-Si) is natively intrinsic (neither p nor n-type). It can be doped p- or n-type, but only at significant reduction in minority carrier lifetime. For this reason, a-Si solar cells are traditionally made as p-i-n junctions, mitigating but not eliminating the effects of the low-quality doped layers.
- Polymers and other organics are usually impossible to dope both p and n in the same material system. However, these will not be discussed below, as the large exciton binding energy in organic materials means that an abrupt heterojunction interface is required to separate charge; an electric field is not sufficient. Therefore these are not promising candidates for the field-effect architecture.

Since p-n junctions are impractical or impossible with these materials, other architectures are sometimes used. In commercial applications, the p-n heterojunction is common, most

prominently with p-CdTe / n-CdS and with p-$CuIn_xGa_{1-x}Se_2$ / n-CdS. The most common difficulty with heterojunctions relate to incompatibilities between the two materials, particularly high interface recombination, poor work-function match leading to weak built-in fields, materials processing incompatibility, or harmful effects of interdiffusion. For example, the CdTe-CdS junction yields a remarkably efficient device, as long as the junction is annealed in the presence of chlorine [5]. Nevertheless, tradeoffs are inevitable. The CdS is not perfectly transparent, and the light that it absorbs ends up going to waste [5]. One might like to replace CdS with a different material which produces an equally good charge-separating interface, but with higher transparency. Unfortunately, no such material is known to exist.

Another alternative to the p-n junction is the Schottky junction. Unfortunately, these tend to perform quite poorly in practice. There are two related reasons. First, they usually have rather low built-in voltages, around half the bandgap or less. The open-circuit voltage of a solar cell cannot possibly be higher than the built-in voltage, as there would be no driving force to collect the current. Second, a metal-semiconductor interface has a continuum of states spanning the semiconductor's bandgap, and therefore has a very high interface recombination velocity. Even worse, this high-recombination-velocity interface is right at the depletion region, where it causes maximal damage to the solar cell performance [6].

In the context of crystalline silicon, there is a special variant of the Schottky junction called MOS-IL (metal-oxide-semiconductor inversion-layer), where a thin tunnel junction separates the metal and semiconductor. This both lowers the recombination velocity and increases the barrier height, with the latter due in part to introduction of electronegative caesium atoms at the interface. While these may have quite high efficiency [7], they have can have poor stability [8], and more importantly, it is unclear to what extent this device concept can be applied in other material systems, since it seems to rely on special properties of the Si-SiO_2 interface.

4.1.2 Field effect architecture

In this work, we propose a new solar cell architecture based on the field effect. As shown in Fig. 4.1, the field from the gate creates a depletion and inversion layer in the semiconductor. The overall device is planar, which is helpful for manufacturability.

There are a number of technical challenges that are immediately apparent from glancing at Fig. 4.1. First, the top electrode is expected to screen the semiconductor from feeling the gate. Overcoming this problem is a major focus of this work; we use the term "screening engineering" to characterize the methods for addressing it. Second, holding the gate at a large voltage would seem to undermine the ability of the photovoltaic device to generate its own power. But in fact, as long as the gate insulator has low leakage, the gate draws negligible power. For example, in the experimental results shown below, the gate uses up less than 1% of the power generated. Third, using three electrodes rather than two would seem to add undesirable complication to the device, making it harder to integrate and install, and also making it harder to wire multiple cells in series. Fortunately, there are ways around this.

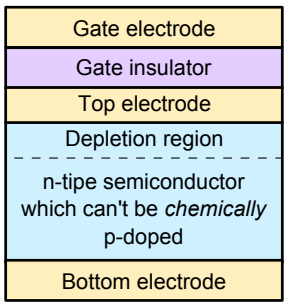

Figure 4.1 – Overview of device architecture. The negative gate voltage causes depletion in the semiconductor. (Alternatively, the semiconductor could be p-type, with positive gate.) *Screening engineering* is the task of designing a top electrode which will not screen the gate from affecting the semiconductor.

Instead of a conventional gate electrode, a similar effect could be created by, for example, a ferroelectric gate, or by engineering the interface to contain fixed charge of the appropriate sign. Alternatively, the bottom electrode and gate can be wired together (a "self-gating" configuration), as discussed below.

Since both the photovoltaic effect and field effect have been well understood for many decades, it is hardly surprising that they have been combined from time to time in the past. For example, in one line of research [9–11], gates have been used to improve semiconductor-interface surface passivation, repelling minority carriers from surfaces not covered by contacts. In another line of research [12–14], a gate was used to partially replace one of the doped layers in an amorphous silicon solar cell. Under the grid contact, however, the layer was doped as usual.

In all these cases, however, the gate was used *in addition* to a conventional p-n contact, and if the conventional p-n contact has poor quality, the gate can do very little to improve the cell. The reason is that a cell performance is related to the *logarithm* of the overall saturation current. (The saturation current is J_0 in the classic photovoltaic diode equation $J = J_0(\exp(qV/k_BT)-1)-J_{\text{photo}}$ [6]; a lower saturation current leads to higher performance.) For example, if the conventional contact is reduced to cover only 10% of the surface, with no surface recombination whatsoever in the non-contacted areas, then the saturation current can go down by as much as a factor of 10 (Fig. 4.2). Unfortunately, this provides a surprisingly small overall benefit to the cell's performance, raising the open-circuit voltage by just $-\frac{k_BT}{e}\ln(10\%) = 0.06$ V [15]. The goal of this work is to show that larger improvements are possible with screening engineering.

4.1.3 Screening engineering

We will discuss two methods of screening engineering.

In Sec. 4.2, we discuss using graphene as the top electrode. Graphene, a single-atom-thick

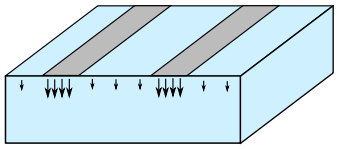

Figure 4.2 – The sources of saturation current are schematically indicated in the case of high contact recombination. By covering only a small area with contacts and using the field effect to create a well-passivated surface elsewhere, the saturation current is reduced [15]. However, this only has limited effectiveness in improving the cell quality, which goes as the logarithm of saturation current.

conductor, does not fully screen the gate both because it is thinner than its Debye length, and because has a low density-of-states. We will show theoretical and experimental results for monolayer and bilayer graphene, and we will discuss more generally the engineering requirements on this type of system.

In Sec. 4.3, we discuss using a "nanoporous" electrode, consisting (for example) of an interconnected network of thin (nanoscale) wires. Again, we will show theoretical and experimental results, and discuss the general engineering requirements.

4.2 Graphene for field-effect control

Graphene is one-atom-thick sheet of sp^2-bonded carbon atoms. It has been investigated as a transparent electrode in solar cells [16,17] thanks to its promising combination of conductivity, flexibility, transparency, and raw-material abundance (unlike, for example, the common transparent conductor indium tin oxide, which contains the rare element indium).

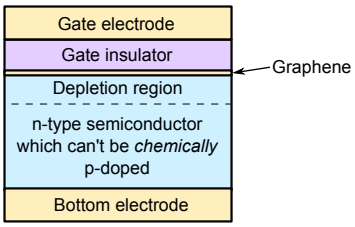

Figure 4.3 – Schematic of device architecture with graphene. For the more detailed experimental geometry, see Sec. 4.2.4.

We are particularly interested in graphene as a non-screening electrode, as shown in Fig. 4.3. Fundamentally, we are taking advantage of the low density-of-states of graphene, which in turn is due to both its thinness and its unusual electronic band structure. The mechanism can be viewed two ways. First, one can say that graphene is thinner than its own Debye length, and therefore allows fields to penetrate through. Second, one can say

that the gate fills electronic states, raising or lowering the effective graphene workfunction, which in turn changes the Schottky barrier height.

These mechanisms are closely related to the concept of *quantum capacitance* [18, 19], which was first discussed in the context of the "2-dimensional electron gases" (2DEGs) in III-V semiconductor quantum wells. (In other contexts it has been called *chemical capacitance* [20]). A voltmeter measures differences in the electrochemical potential of electrons, also called Fermi level. Electrochemical potential, as the name implies, is the sum of electric potential and (internal) chemical potential of free electrons. For example, a p-n junction in thermal equilibrium has an electric potential gradient across it (the "built-in field"), balanced by an equal and opposite chemical potential gradient; a voltmeter across the junction would read 0V despite the presence of an electric potential drop. Applying this general principle to a capacitor, moving charge from one plate to the other can create an electrochemical potential drop by creating an electric potential drop, or creating a chemical potential drop, or both. The former corresponds to the conventional (or "geometrical") capacitance, while the latter corresponds to quantum capacitance. The low quantum capacitance of graphene [19] allows electric fields to penetrate it [18].

4.2.1 System modeling methods

Modeling gated graphene devices requires solving three equations self-consistently.

First,

$$Q_{\text{graphene}} = D_{\text{gate}} - D_{\text{semi}}, \tag{4.1}$$

i.e. the charge on graphene (per unit area) equals the discontinuity between the displacement field in the gate insulator (which equals the charge on the gate electrode), and the displacement field at the surface of the semiconductor.

Second,

$$\chi_{\text{graphene}} = \chi^{\text{CNP}}_{\text{graphene}} + \Delta\chi(Q_{\text{graphene}}) \tag{4.2}$$

where χ is workfunction, $\chi^{\text{CNP}}_{\text{graphene}}$ is the workfunction of graphene at its charge-neutral point, and $\Delta\chi(Q)$ is a function, related to the density of states, that inputs the net charge on the graphene sheet and outputs the shift in Fermi level. Since extra electrons (negative charge) occupy higher-energy states (lower ionization energy), $\Delta\chi(Q)$ is an increasing function of Q. Its formula for monolayer graphene is [21]:

$$\Delta\chi(Q) = \text{sign}(Q)\frac{\hbar v_F}{e}\left|\pi Q/e\right|^{1/2} \tag{4.3}$$

where $e > 0$ is the elementary charge and the parameter $v_F = 1 \times 10^6$ m/s is called the *Fermi velocity* of graphene, related to its band structure. $\chi^{\text{CNP}}_{\text{graphene}}$ was estimated as 4.6 eV [21].

Third, the drift-diffusion-poisson equations must be satisfied, which take as inputs χ_{graphene} and the voltage between the semiconductor electrodes, and return as outputs D_{semi} and the semiconductor's current. The method for this is described in the next section; for now,

this step can be thought of as a black box. In practice, this aspect of the simulation was completed prior to any graphene-specific analysis, using the program COMSOL. Thousands of combinations of χ_{graphene} and V were simulated. The results were exported to a different software program, Mathematica, where they were interpolated into a smooth function and used to calculate self-consistent I-V curves depending on graphene and gate properties.

Three different systems were modeled: Single-layer graphene, bilayer graphene, and graphite ("infinite-layer graphene").

For single-layer graphene, we performed the calculation by starting with the initial guess $\chi_{\text{graphene}} = \chi^{\text{CNP}}_{\text{graphene}}$. We looked up the corresponding simulation in the COMSOL output to get the displacement field at the surface of the semiconductor, D_{semi}, which gives updated guesses for Q_{graphene} and then χ_{graphene}, at which point the process is repeated. To help numerical stability, χ_{graphene} was updated each time to only halfway between the old value and the value inferred from D_{semi}. Convergence was tested in all cases.

In practice, there are often charged impurities at the graphene interface, typically negative (hole-doping) and below 10^{12} e/cm^2. These were not explicitly included in the simulations above because they correspond merely to a linear shift in D_{gate}.

For bilayer graphene, the procedure is analogous. The experimental system was not lattice-matched bilayer graphene; instead, one chemical-vapor-deposition-grown monolayer graphene sheet was transferred, then a second sheet with a random angle offset. Therefore, instead of using the bilayer graphene material band structure, the two sheets were treated as two separate electron systems, each screening each other and the semiconductor. The distance between graphene sheets was set at 3.3 Å, while the static dielectric constant between them was set to $2.4\epsilon_0$ [22]. Therefore the D-fields in all three regions (gate insulator, semiconductor, and between the two graphene sheets) could be calculated and related to the free charge on the graphene sheets. The workfunction of the graphene sheet adjacent to the semiconductor was treated as a free parameter, which we deduced by recursive bisection. As before, this workfunction is used to calculate D_{semi} (using the COMSOL output), which in turn allows calculation of the D field between the two sheets. This information can be combined with the fact that the two sheets are electrically contacted (and therefore have equal electrochemical potentials), in order to infer the charge on the next graphene sheet. This finally gives D in the gate insulator, which can be compared with the expected value (D_{gate}) to move into the next round of recursive bisection.

The procedure for infinite-layer graphene is a straightforward extension of the procedure for bilayer graphene. Of course, as expected, the value of D_{gate} does not affect the infinite-layer system's behavior.

4.2.2 Semiconductor modeling methods

The flow of charge in the semiconductor was modeled via the drift-diffusion-poisson equations, using the finite element method, as implemented in COMSOL software (version 4.1).

The equations to be solved are:

$$\begin{aligned}
\mathbf{J}_n/e &= D_n\nabla n + n\mu_n\mathbf{E} && (4.4)\\
\mathbf{J}_p/e &= -D_p\nabla p + p\mu_p\mathbf{E} && (4.5)\\
\nabla\cdot\mathbf{J}_n &= -G && (4.6)\\
\nabla\cdot\mathbf{J}_p &= -G && (4.7)\\
\mathbf{E} &= -\nabla\phi && (4.8)\\
\nabla\cdot(\epsilon\mathbf{E}) &= e(p-n) && (4.9)
\end{aligned}$$

where n, p are the electron and hole densities, $\mathbf{J}_n, \mathbf{J}_p$ are the corresponding current densities, D_n, D_p are diffusion coefficients, μ_n, μ_p are mobilities, G is the *net* generation rate (the rate at which electron-hole pairs are photogenerated, minus the rate at which they recombine), and $e > 0$ is the elementary charge. The diffusion coefficients are inferred from mobility by the Einstein relations: $D_{n,p} = (k_B T/e)\mu_{n,p}$. In the simulations shown here, semiconductor parameters for silicon were used [23], as shown in Table 1. Not all of these parameters are independent: $n_i = \sqrt{N_C N_V}\exp(-E_{\text{gap}}/2k_B T)$.

Symbol	Value used	Definition
N_C	2.8×10^{19} cm^{-3}	Conduction band effective density of states [23]
N_V	2.65×10^{19} cm^{-3}	Valence band effective density of states [23]
E_{gap}	1.12 eV	Bandgap [23]
μ_n	1000 cm^2/V·s	Electron mobility [23]
μ_p	500 cm^2/V·s	Hole mobility [23]
X_{semi}	4.0 eV	Conduction band minimum relative to vacuum [23]
N_D	10^{15} cm^{-3}	n-type dopant density
ϵ_{semi}	11.8	Semiconductor dielectric constant [23]
n_i	1.1×10^{16} cm^{-3}	Intrinsic carrier concentration [23]
τ_n	100 μs	Electron minority carrier lifetime
τ_p	100 μs	Hole minority carrier lifetime
A^*/A	2.1	Effective Richardson constant correction [24]

Table 1 – Simulation parameters. Data is drawn from sources where indicated, or otherwise chosen to correspond to experimental parameters.

The workfunction of the back contact was set to make it an ideal (uncharged) ohmic contact; i.e. the metal's workfunction was set to $X_{\text{semi}} + \frac{E_{\text{gap}}}{2} - \frac{k_B T}{e}\ln\frac{N_D}{n_i} = 4.26$ eV.

The incident sunlight was inferred from the "AM1.5G" reference solar spectrum [25]. By numerically integrating, the number of photons above silicon's bandgap is estimated as $G_{inc} = 2.74 \times 10^{21}$ photons/(m^{2}s). The depth-dependent absorption profile was then

estimated as $G(z) = G_{inc}\alpha e^{\alpha z}$, where $z < 0$ is the depth below the top surface and $\alpha = 3 \times 10^3/\text{cm}$ is a typical absorption coefficient for the above-bandgap light in silicon. (A more detailed approach would take into account the variation of α with frequency, but for the purpose of these simulations, a very accurate light absorption profile was not necessary. Indeed, tests confirmed that the detailed absorption profile had no effect on the comparisons or qualitative results discussed here.) The thickness of the silicon was set at 10 μm, which is unrealistically small but easier to simulate; again, the simulated thickness had no effect on the conclusions below.

Shockley-Reed-Hall (defect) recombination was calculated by the usual formula:

$$R_{SRH} = \frac{np - n_i^2}{(p + n_i)\tau_n + (n + n_i)\tau_p}$$

Auger recombination was negligible in all simulations, and therefore was normally not calculated.

Back-surface recombination was set to zero, as this effect is negligible (compared to other recombination sources) in well-constructed cells using back-surface-fields [6]. For the front surface, minority carrier recombination was assumed to be infinite, while majority carriers satisfy the Crowell-Sze model combining drift-diffusion and thermionic emission [26]. In this model, one uses a majority-carrier recombination velocity

$$v_R = \frac{A^* T^2}{eN_C} = 5 \times 10^4 \text{ m/s}$$

The barrier heights were calculated in the framework of the Schottky-Mott model—i.e. the vacuum level is assumed to be continuous, and therefore the barrier height can be related to the metal workfunction and the semiconductor bandgap and electron affinity [23, 27] (The semiconductor conduction-band-minimum was taken to be 4.0 eV below vacuum, a typical value for silicon [23].) In reality, the Schottky-Mott model is a rather poor predictor of barrier heights [23,27], which vary significantly less than workfunctions. However, the model is likely to be more accurate than usual in this situation, because the barrier height is being changed *in situ*, without any complication from changing surface reconstruction, dangling bonds, ionic migration, and so forth.

Image-force lowering of the Schottky barrier is taken into account, using the formula $\Delta\phi = (eD_{\text{semi}}/(4\pi))^{1/2}$ [23]. Since the correction was small, we took it into account in a low-order manner rather than self-consistently: A simulation was done without image-force lowering; then the D_{semi} from that simulation was used to calculate $\Delta\phi$; then a new simulation with lower barrier was used for final results.

The modeled system depends on only one coordinate (z), and is a uniform extrusion in the other two dimensions. Therefore it is analyzed as a 1D system.

Various tricks were used to increase the likelihood that the simulation would converge. The system was in fact simulated in a two-step process. First, the thermodynamic-equilibrium

potentials were calculated. This simulation converges easily because there is only one variable to solve for, the potential. (The electron and hole concentrations, and hence also the charge density, is related to the potential via the Boltzmann distribution.) This first step was used as the initial conditions for the second step, the coupled drift-diffusion-poisson-equations simulation, which converged much less reliably. (When the first step did not produce good enough initial conditions for the second step to converge, the minority carrier concentration was multiplied everywhere by $10^6 - 10^{10}$.) When this simulation converged, the voltages were gradually swept up and down, outputting the current in each case to get an I-V curve, and using each converged simulation as the initial conditions of the next. When the finite-element simulation did not converge, various steps were tried including refining and reshaping the mesh, and changing calculation parameters including voltage step size and convergence threshold. When the simulation did converge, similar steps were done regardless, in order to check that the solution was reliable and robust.

4.2.3 Modeling results

Results from these simulations are shown in Figs. 4.4–4.5. A few points are particularly worth noting.

First, realistic gate fields D_{gate} for graphene can reach about $1.5 \times 10^{13} e/\text{cm}^2$ for Si-SiO_2 gates [28], and as high as $6 \times 10^{13} e/\text{cm}^2$ for electrolytic gates [29]. Therefore, these models suggest that it should be quite possible to substantially enhance, an even saturate, the Schottky barrier height, achieving barriers well beyond what would arise from intrinsic material properties alone. This should also be possible in bilayer graphene, although larger fields are required. Second, quite high efficiencies should be possible, notwithstanding the high contact recombination assumed by the model. Third, the workfunction of graphene depends quite weakly on device voltage; i.e. as an I-V curve is traversed (holding the gate constant), the workfunction remains largely unchanged. Therefore, the I-V curve of a given device is expected to look like a normal Schottky diode. In reality, graphene-semiconductor and related junctions often have I-V curves with unusual shape (for example, Refs. [16, 30], and see also below), but the reasons remain poorly understood.

4.2.4 Experimental results

Gated-graphene-on-silicon devices were made experimentally in collaboration with the Alex Zettl group. Starting with an n-type ($N_D \approx 10^{16}$ cm^{-3}) Si wafer, 100nm thermal oxide was grown by a dry anneal. After coating with PMMA resist, electron-beam lithography followed by 5:1 buffered HF etch was used to etch a 2mm×2mm square window down to the silicon. Graphene was grown by chemical vapor deposition, following the method of Ref. [31]. A graphene sheet on PMMA was placed down, covering all of the Si window and some of the SiO_2, and the PMMA was dissolved in acetone. (This process was repeated twice for "bilayer" samples.) A Cr/Au contact was evaporated onto an area where the graphene was

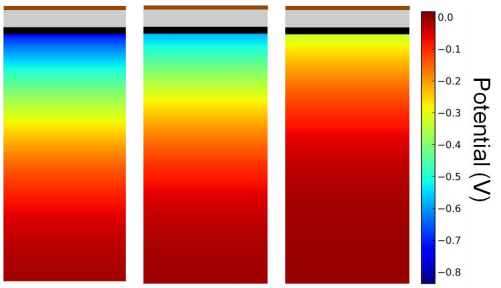

Figure 4.4 – Simulated electric potential (in V) in gated-graphene-semiconductor devices, with $1.5 \times 10^{13} e/\mathrm{cm}^2$ gate charge, at short circuit under illumination. Brown is gate, gray is gate insulator, black is graphene, colors represent semiconductor. Left: Single-layer graphene; Center: Bilayer graphene; Right: Graphite (infinite layers).

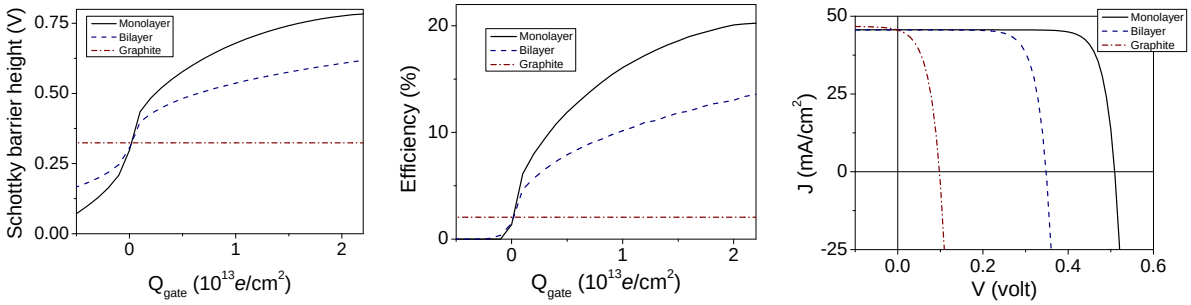

Figure 4.5 – Simulated Schottky barrier height (left) and solar cell efficiency (center) as a function of gate charge. Right: I-V curves at $1.5 \times 10^{13} e/\mathrm{cm}^2$ gate charge.

sitting on SiO_2 (not Si), patterned by shadow mask. For a back contact to the Si, the SiO_2 was removed with 5:1 buffered HF, then 70nm of Al was evaporated. An ionic liquid gate was used: A contact was put on SiO_2 away from the graphene, then the ionic liquid EMI-BTI (1-ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide, Sigma Aldrich) was dropped to cover both the contact and the graphene.

The I-V curves are shown in Fig. 4.6, under a 1-sun solar simulator. (The gate leakage current was negligible compared to the current through the device.) The gate clearly has a strong effect on the device, quadrupling the power generation efficiency from $\approx 0.5\%$ to $\approx 1.8\%$. (While the relative numbers are reliable, the absolute efficiency measurements should be taken with a grain of salt for such a small device [32].) Unfortunately, the unusual I-V curve shape (see previous section), combined with the high series resistance, precluded a measurement of the Schottky barrier height, which would have made a more useful comparison with theory.

In future work, the efficiency could be improved by optimized device structures, particularly to lower the series resistance, and other semiconductors can be explored. We note that the device can function as a switch in addition to a solar cell, an aspect recently explored in independent work [33].

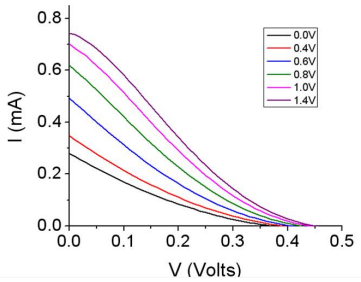

Figure 4.6 – I-V curves for gated monolayer graphene on silicon devices. Gate voltage is indicated.

4.3 Nanoporous electrodes for field-effect control

An alternative type of screening-engineered electrode is a "nanoporous" electrode. An example would be an interconnected percolating network of metal nanowires, laying flat on the surface of a semiconductor, e.g. Fig. 4.7. These sorts of electrodes have already been explored as solar cell electrodes for other reasons—they are transparent, conducting, flexible, inexpensive, and (unlike indium-tin-oxide) do not necessarily rely on rare chemical elements [34–37]. From an optics perspective, a properly engineered nanowire film can be even better than transparent; it can enhance light absorption in the semiconductor by plasmonically scattering incident light into oblique angles where it is more likely to be absorbed [38, 39].

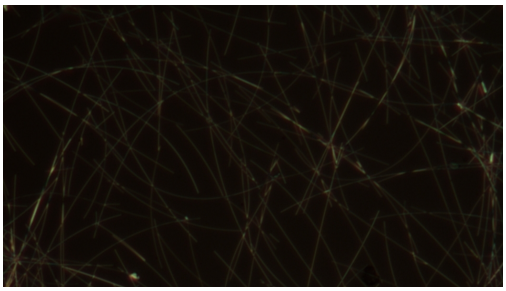

Figure 4.7 – Example of a nanoporous electrode: A percolating silver nanowire film. Image at 60X; average nanowire length is 25μm. Nanowires dropcast from a solution (SeaShell Technology) onto a glass substrate.

Intuitively, one expects that such an electrode would allow a gate field to pass through to the semiconductor beneath, without fully screening it. (This is analogous to how, in classical electrostatics, an electric field can penetrate through a porous capacitor plate [40].) Below, we will discuss in more detail the extent to which this occurs.

4.3.1 Design overview

Above (see Sec. 4.1.2 and Fig. 4.2), we discussed how reducing the surface coverage fraction of a metal in a Schottky junction solar cell can increase its performance, but only to a limited

extent (increasing voltage by $\frac{k_B T}{e} \ln(1/f)$, where f is the surface coverage fraction). Fig. 4.8 shows how nanotechnology can enable a significantly larger performance boost. When the lateral dimension of the electrode finger w is much smaller than the depth d at which the potential far from the electrode matches the Schottky interface potential (see Figs. 4.8(a-b)), one expects the gate field to spread under the electrode as shown in Fig. 4.8(c). Therefore the *effective* electron barrier height is larger than the intrinsic Schottky barrier height; in fact, the electrons in bulk have to pass over a larger saddle-point barrier before reaching the electrode. On the other hand, if $d \ll w$, as in Fig. 4.8(d), then the gate field has essentially no effect on the semiconductor beneath the contacts. This, therefore, corresponds to the case discussed above, with only the limited $\frac{k_B T}{e} \ln(1/f)$ voltage improvement.

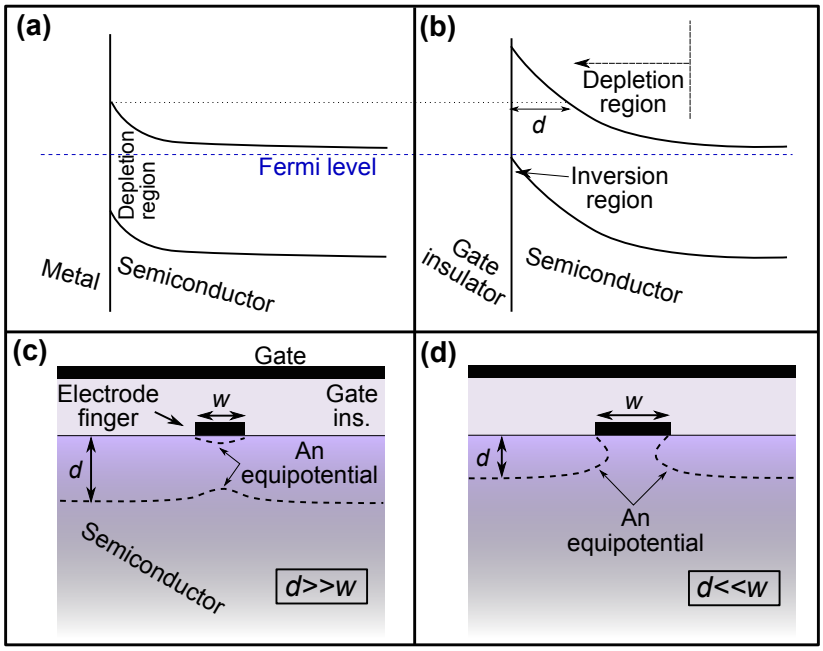

Figure 4.8 – Illustration for rule-of-thumb (Eq. 4.11) dictating nanofinger width. (a) A band-diagram for a semiconductor-metal Schottky junction in thermodynamic equilibrium. (b) At a semiconductor-gate-insulator interface, there is stronger band bending, leading to inversion at the surface. The quantity d is the depth with the same potential as at the semiconductor-metal interface. (c) A cross-sectional view of a device where the depth $d \gg w$, the electrode nanofinger width. The two dashed lines are at the same potential. The effect of the gate "spreads under" the electrode, creating a saddle point barrier that blocks electron flow into the electrode. (d) In a device where $d \ll w$, the field cannot get under the electrode. Therefore the gate can only yield a limited improvement in the device.

In order to achieve good screening engineering, according to this principle, the maximum allowable electrode finger width depends on the length-scale over which the potential varies in the semiconductor. For more heavily-doped semiconductors, the depletion width is smaller and the potential varies more rapidly. Therefore, the electrode requirement be-

comes more stringent as doping increases. To be more specific, we assume that the Schottky barrier puts the metal Fermi level in exactly the middle of the semiconductor bandgap (the largest barrier typically possible), and we use $d = w$ criterion for maximum electrode finger width (see Fig. 4.8). Assuming an n-type semiconductor, the Fermi level is $\frac{k_B T}{e} \ln \frac{N_C}{N_D}$ below the conduction-band minimum in bulk; the charge density in the depletion region is approximately $-N_D e$; and the Fermi level at the inverted semiconductor-insulator interface is approximately at the valence-band maximum. Using the Poisson equation, we get the formula

$$d \approx \sqrt{2\frac{E_{\text{gap}} - \frac{k_B T}{e} \ln \frac{N_C}{N_D}}{N_D e/\epsilon_s}} - \sqrt{2\frac{\frac{E_{\text{gap}}}{2} - \frac{k_B T}{e} \ln \frac{N_C}{N_D}}{N_D e/\epsilon_s}} \tag{4.10}$$

and a corresponding rough criterion for screening-engineered electrode width:

$$w \lesssim \sqrt{2\frac{E_{\text{gap}} - \frac{k_B T}{e} \ln \frac{N_C}{N_D}}{N_D e/\epsilon_s}} - \sqrt{2\frac{\frac{E_{\text{gap}}}{2} - \frac{k_B T}{e} \ln \frac{N_C}{N_D}}{N_D e/\epsilon_s}} \tag{4.11}$$

Table 2 shows the numbers, using the parameters of silicon as a typical example. For a semiconductor with unintentional doping around 10^{17} cm^{-3}, such as $\text{In}_x\text{Ga}_{1-x}\text{N}$, a screening-engineered porous electrode would need electrode finger widths of the order of 40 nm or less. Such electrodes exist [41, 42], but nevertheless may be a challenge, whereas lower-doped semiconductors would be easier (for example, lead salt quantum dots have doping around 10^{16} cm^{-3} [43]). On the other hand, the third column of Table 2 shows that the requirements on the gate insulator are not too stringent, even up to quite high doping.

Bulk doping	Maximum electrode finger width	Gate-induced charge needed to get to inversion	Prospects
10^{14} cm^{-3}	1.5 μm	3×10^{10} cm^{-2}	OK
10^{15} cm^{-3}	400 nm	1×10^{11} cm^{-2}	OK
10^{16} cm^{-3}	130 nm	3×10^{11} cm^{-2}	Maybe
10^{17} cm^{-3}	40 nm	1×10^{12} cm^{-2}	Challenging

Table 2 – Prospects for screening engineering varies as a function of semiconductor doping, assuming the material parameters of silicon. The second column is based on the electrode-width criterion Eq. 4.11. The "prospects" is based on the discussion in the text.

Table 2 can be supplemented with some additional considerations, still based on the idea of Fig. 4.8. First, a higher intrinsic Schottky barrier leads to more stringent length requirements for screening engineering. Second, under forward bias, the semiconductor fields are modified so as to ease the length requirements (i.e., the short-circuit figure in Table 2 may somewhat overstate the difficulty).

4.3.2 Modeling methods

To test more carefully the ideas above, we performed semiconductor modeling, along the lines of Sec. 4.2.2. In this situation, a 2D finite-elements model was necessary, while the third dimension was a uniform extrusion. An example structure is shown in Fig. 4.9, although the actual simulation area (as indicated in the figure) is only half of one repeating unit, with mirror boundary conditions filling in the rest of the infinite structure.

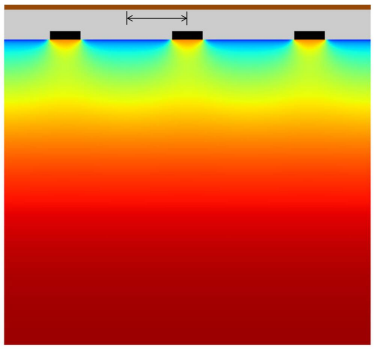

Figure 4.9 – An example of a periodically-repeating structure being simulated. Brown is gate, gray is gate insulator, black is electrode, and the colors correspond to voltages in the semiconductor. The arrow near the top indicates the boundaries of a computer simulation, using mirror boundary conditions. This is a two-dimensional cross-section; the third dimension need not be simulated explicitly, as it is a uniform extrusion.

All electrodes were taken to be ideal, perfectly-conducting metals. Therefore their interiors were not actually part of the simulation; instead, their surfaces were used as equipotential boundary conditions. In order to focus on the electronic, not optical, effects, the light absorption was treated in the same simplified manner of Sec. 4.2.2. (It was checked that, for the results shown here, shadowing and other effects had negligible consequences on the simulation results.) All other parameters and approximations are as in Sec. 4.2.2.

One potentially important effect not simulated (due to technical limitations) was tunneling between the inverted semiconductor surface and the metal electrode. Particularly with electrochemical gating, a minority-carrier-rich region can be within a few nanometers of the metal interface, and can therefore tunnel. This effect is conceptually important, as it allows an open-circuit voltage larger than the built-in field at the metal-semiconductor interface. (Without the tunneling, this is impossible, as the drift-diffusion equations allow no driving force for charge extraction.) Even without tunneling, the simulations suggest that screening engineering has the expected befefits, as discussed in the next section.

In addition to a set of simulations with $V_{\text{gate}} = -10$ V, simulations were also done of a "self-gating" configuration. For these, the bottom electrode voltage was set equal to the gate voltage, corresponding to the situation where the two are wired together. This necessarily has a slightly lower gate-induced charge than a high-voltage gate, but this may be compensated by the practical benefits of having only two terminals at both the device and module level.

Self-gating simulations were performed by sweeping the top electrode voltage (rather than bottom), holding $V_{\text{gate}} = V_{\text{bottom}} = 0$. The workfunction of the gate was set to 5.1 eV, the value of silver—a high-workfunction electrode improves the gating for an n-type device [44].

4.3.3 Modeling results

Some sample finite-elements simulations are shown in Fig. 4.10. As predicted from Fig. 4.8, the sufficiently narrow fingers have a potential profile that forms a saddle point under the electrode, which in turn increases the effective barrier height. The corresponding I-V curves are shown on the right. While the results confirm the importance of a high intrinsic Schottky barrier, they also indicate how screening engineering can substantially improve the device.

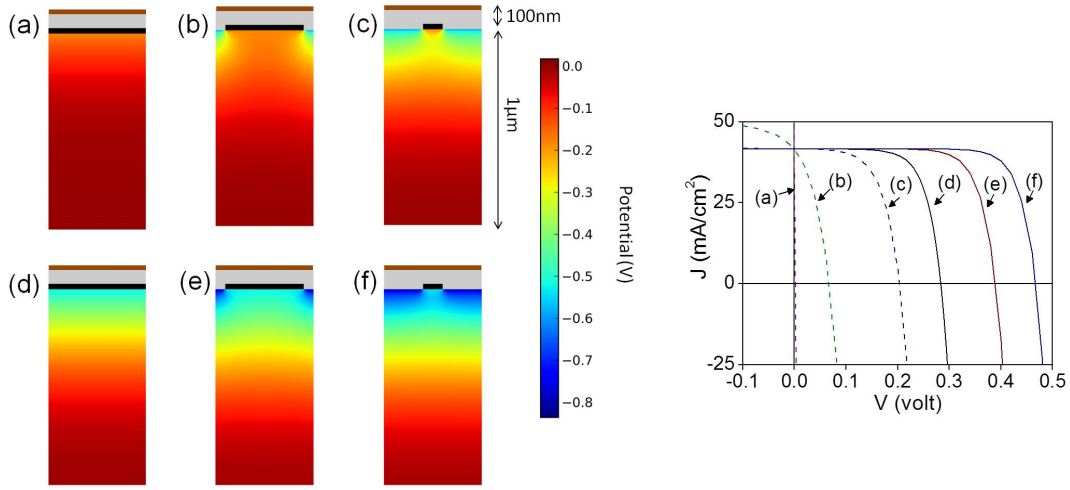

Figure 4.10 – Electric potential plots at short-circuit (left), and I-V curves (right) from simulations of nanofinger-electrode devices. In (a–c), the top electrode is "ohmic", with low (4.45 eV) workfunction. In (d–f), the top electrode is "Schottky", with higher (4.8 eV) workfunction. The top-electrode finger width is infinite (a,d), 400 nm (b,e), or 100 nm (c,f). The shape of the potential is consistent with the expectations from Fig. 4.8: For sufficiently narrow fingers (c,f), the potential forms a saddle point that increases the effective barrier height.

Fig. 4.11 shows how open-circuit voltage and power conversion efficiency depends on finger width. (These simulations have a fixed 10μm separation between fingers.) The "Ohmic" curves correspond to 4.45 eV workfunction electrode, while the "Schottky" curves correspond to 4.8 eV workfunction (the same values as above). (The term "ohmic" is justified by its I-V curve, Fig. 4.10, right side, (a).) The "Green model" [15] dotted curve in Fig. 4.11, as described in Sec. 4.1.2, is $V_{\text{OC}} = K + k_B T \ln(1/f)$, where K is a constant and f is the fraction of the surface covered by the metal fingers. This model accounts for purely the effect of partial surface passivation, but not screening engineering. Fig. 4.11 shows that narrow fingers dramatically outperform the Green model, and therefore it underlines the importance of screening engineering.

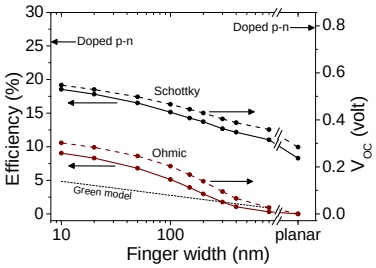

Figure 4.11 – Dependence of cell power conversion efficiency and open-circuit voltage on finger width. See text for details.

Fig. 4.11 also shows, for comparison, a "p-n junction" simulation. Here, an abrupt p-n junction was simulated, with planar ohmic contacts on both sides. To be consistent with the previous simulations, interface recombination at the ohmic contacts was set to zero. (See discussion above.) Therefore, the only efficiency limit was bulk minority carrier recombination. This gives the cells significantly higher performance than the screening-engineered design, where interface recombination continues to dominate even at small finger widths. Indeed, as expected from the discussion in Sec. 4.1.2, the screening-engineered field-effect architecture will not outperform an ideal p-n junction, but its reasonably high efficiency may make it a good alternative when p-n junctions cannot be manufactured, and particularly when interface recombination and/or low built-in voltage is the efficiency limitation.

Finally, Fig. 4.12 shows the potential in the two-electrode self-gated configuration. These simulations suggest that the self-gated device can perform with comparable, but modestly lower, efficiency than the device with a large externally-applied gate.

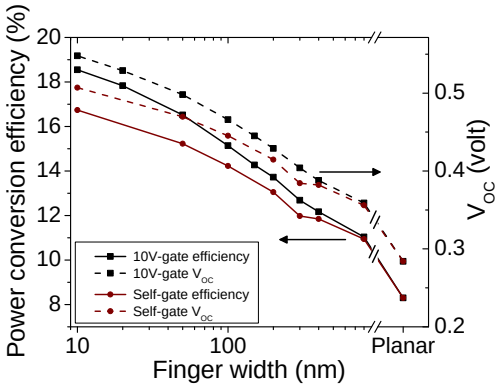

Figure 4.12 – Comparison between simulations of self-gated devices and externally-gated devices.

4.3.4 Experimental results

As a controlled proof of principle, experimental devices were made in collaboration with the Alex Zettl group, using electron-beam lithography to make metal nano-fingers on silicon. Two types of cells were made, one with ohmic top contacts, the other with Schottky.

The ohmic cells started with a p-type silicon wafer ($N_A \approx 10^{16}$ cm^{-3}), with 100nm of thermal oxide grown by dry anneal. After coating with PMMA resist, three rounds of electron-beam lithography were used to define, first, alignment marks, then second, an exposed silicon device area, then third, aluminum contacts (250nm width, 75nm thickness, 5μm spacing). For the back contact, the SiO_2 was etched away with 5:1 buffered HF, then aluminum contacts were evaporated. To ensure that both the back and front contacts are intrinsically ohmic, the cell was annealed in argon (150sccm, 475C, 30 minutes). The use of symmetric contacts means that any rectification is due to the field effect. Finally, the gate insulator and gate are added by evaporating 150nm SiO_2, 1.5nm chrome, and 12nm gold. (Calculations suggest that this stack should allow about 40% of the light to pass into the cell, while reflecting or absorbing the other 60%.)

The Schottky cells were similar, but where the metal fingers had a silicon-chrome Schottky interface (the fingers were 300nm width, 5nm chrome under 50nm gold thickness, 5μm spacing). Additionally, the silicon was $N_A \approx 3 \times 10^{15}$ cm^{-3}. The aluminum back contact was annealed before the Schottky contact was deposited.

The I-V curves are shown in Fig. 4.13, under a 1-sun solar simulator. (The gate leakage current was negligible compared to the current through the device.) The dramatic effect of the gate is evident, switching the device between ohmic and a moderate solar cell (the efficiency is estimated at 6% in the Schottky case, although as mentioned above, these absolute efficiency measurements may not be accurate for such a small device [32].)

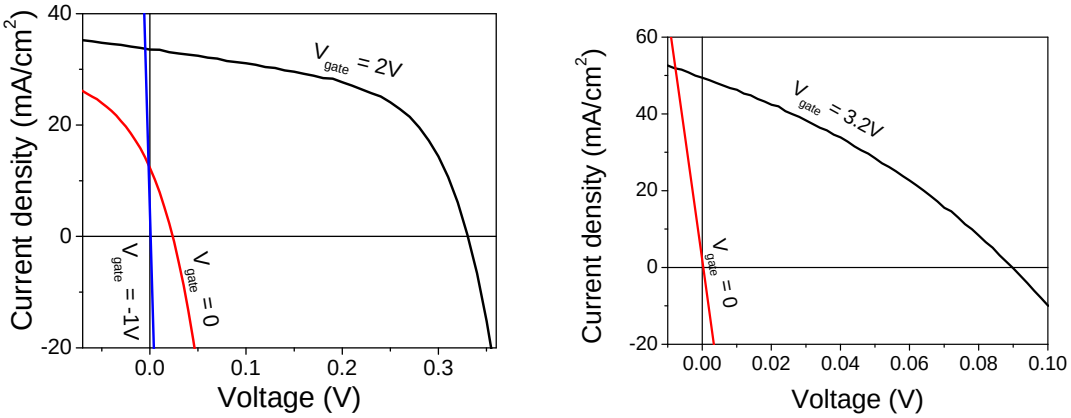

Figure 4.13 – Left: Experimental results with Schottky (Cr/Au) fingers. Right: Experimental results with ohmic (annealed Al) fingers. (Note different scales.)

4.4 Conclusion

To summarize, we have discussed design concepts, device modeling, and experimental results for two types of field-effect based solar cells: Graphene electrodes (or more generally, low-density-of-states electrodes), and nanofinger electrodes (or more generally, porous electrodes).

There are a few obvious directions for future research, some of which are already under way. First, we discussed the potential of these devices for improving solar cells from hard-to-dope semiconductors, and we hope to demonstrate that directly with such a semiconductor. (The results above used silicon because it is much easier to work with.) Second, the nanofinger device discussed above was made using electron-beam lithography, which is impractical for a large-area device such as a solar cell. An appealing alternative is to use metal-nanowire electrodes, which can be grown chemically in large quantities and then cast from solution over large areas [34–37]. Yet another alternative, which has been recently explored by a different group [45,46], is field-effect devices with "buckypaper" electrodes—i.e., electrodes from a network of carbon nanotubes. This combines the two device concepts, as it is simultaneously low-density-of-states and porous.

While we have focused in particular on the potential for improving solar cells, there are other applications worth exploring. First, light-emitting diodes (LEDs) and laser diodes could be made in a similar way in order to use hard-to-dope materials (like zinc oxide and group III-nitrides), and more generally control carrier spillover. Second, low-doped, hard-to-contact semiconductors (for example, p-type aluminum nitride) could benefit from a field-effect-induced inversion layer from which carriers could tunnel into the contact. Third, the device can function as an electrical switch; early explorations of this application by other groups have already found promising results [33,47,48].

4.5 References

[1] Wadia, C., Alivisatos, A. P., and Kammen, D. M. *Environ. Sci. Technol.* **43**, 2072–2077 (2009).

[2] Raebiger, H., Lany, S., and Zunger, A. *Phys. Rev. B* **76**, 045209 (2007).

[3] Look, D. C. and Claflin, B. *Phys. Status Solidi B* **241**, 624–630 (2004).

[4] Ma, W., Luther, J. M., Zheng, H., Wu, Y., and Alivisatos, A. P. *Nano Lett.* **9**, 1699–1703 (2009).

[5] Bonnet, D. *Practical Handbook of Photovoltaics*, chapter IIc-3, 334. Elsevier (2003).

[6] Green, M. A. *Solar cells : operating principles, technology, and system applications.* University of New South Wales, Kensington, NSW, (1986).

[7] Hezel, R. and Hoffmann, W. In *Proceedings of 3rd World Conference on Photovoltaic Energy Conversion*, volume 2, 1399–1402, (2003).

[8] Gomaa, N. G. *Renew. Energ.* **24**, 529–534 (2001).

[9] Aberle, A. G., Glunz, S., and Warta, W. *Sol. Energ. Mat. Sol. C.* **29**, 175–182 (1993).

[10] Bai, Y., Phillips, J. E., and Barnett, A. M. *25th PVSC; May 13-17, 1996; Washington, D.C.* , 425 (1996).

[11] Aberle, A. G. *Prog. Photovoltaics* **8**, 473–487 (2000).

[12] Miyazaki, K., Matsuki, N., Shinno, H., Fujioka, H., Oshima, M., and Koinuma, H. *B. Mater. Sci.* **22**, 729–733 (1999).

[13] De Cesare, G., Chicarella, F., Palma, F., Nobile, G., and Tucci, M. *Thin Solid Films* **427**, 166–170 (2003).

[14] Matsuki, N., Abiko, Y., Miyazaki, K., Kobayashi, M., Fujioka, H., and Koinuma, H. *Thin Solid Films* **486**, 210–213 (2005).

[15] Green, M. A. *Appl. Phys. Lett.* **27**, 287 (1975).

[16] Cox, M., Gorodetsky, A., Kim, B., Kim, K. S., Jia, Z., Kim, P., Nuckolls, C., and Kymissis, I. *Appl. Phys. Lett.* **98**, 123303 (2011).

[17] Gomez De Arco, L., Zhang, Y., Schlenker, C. W., Ryu, K., Thompson, M. E., and Zhou, C. *ACS Nano* **4**, 2865–2873 (2010).

[18] Luryi, S. *Appl. Phys. Lett.* **52**, 501–503 (1988).

[19] Miškovic, Z. L. and Upadhyaya, N. *Nanoscale Res. Lett.* **5**, 505–511 (2010).

[20] Bisquert, J. and Vikhrenko, V. S. *J. Phys. Chem. B* **108**, 2313–2322 (2004).

[21] Yu, Y., Zhao, Y., Ryu, S., Brus, L. E., Kim, K. S., and Kim, P. *Nano Lett.* **9**, 3430–3434 (2009).

[22] Lui, C. H., Li, Z., Mak, K. F., Cappelluti, E., and Heinz, T. F. *Nat. Phys.* **7**, 944–947 (2011).

[23] Sze, S. M. and Ng, K. K. *Physics of Semiconductor Devices.* Wiley-Interscience, Hoboken, 3 edition, (2007).

[24] Crowell, C. *Solid State Electron.* **8**, 395–399 (1965).

[25] American Society for Testing and Materials. http://rredc.nrel.gov/solar/spectra/am1.5/.

[26] Crowell, C. R. and Sze, S. M. *Solid State Electron.* **9**, 1035–1048 (1966).

[27] Myburg, G., Auret, F., Meyer, W., Louw, C., and van Staden, M. *Thin Solid Films* **325**, 181–186 (1998).

[28] Ju, L., Geng, B., Horng, J., Girit, C., Martin, M., Hao, Z., Bechtel, H. A., Liang, X., Zettl, A., Shen, Y. R., and Wang, F. *Nat. Nanotechnol.* **6**, 630–634 (2011).

[29] Chen, C., Park, C., Boudouris, B. W., Horng, J., Geng, B., Girit, C., Zettl, A., Crommie, M. F., Segalman, R. A., Louie, S. G., and Wang, F. *Nature* **471**, 617–620 (2011).

[30] Schriver, M., Regan, W., Loster, M., and Zettl, A. *Solid State Commun.* **150**, 561–563 (2010).

[31] Li, X., Magnuson, C. W., Venugopal, A., An, J., Suk, J. W., Han, B., Borysiak, M., Cai, W., Velamakanni, A., Zhu, Y., Fu, L., Vogel, E. M., Voelkl, E., Colombo, L., and Ruoff, R. S. *Nano Lett.* **10**, 4328–4334 (2010).

[32] Snaith, H. J. *Nat. Photonics* **6**, 337–340 (2012).

[33] Yang, H., Heo, J., Park, S., Song, H. J., Seo, D. H., Byun, K., Kim, P., Yoo, I., Chung, H., and Kim, K. *Science* **336**, 1140–1143 (2012).

[34] Hecht, D. S., Hu, L., and Irvin, G. *Adv. Mater.* **23**, 1482–1513 (2011).

[35] De, S., Higgins, T. M., Lyons, P. E., Doherty, E. M., Nirmalraj, P. N., Blau, W. J., Boland, J. J., and Coleman, J. N. *ACS Nano* **3**, 1767–1774 (2009).

[36] Hu, L., Kim, H. S., Lee, J., Peumans, P., and Cui, Y. *ACS Nano* **4**, 2955–2963 (2010).

[37] Wu, H., Hu, L., Rowell, M. W., Kong, D., Cha, J. J., McDonough, J. R., Zhu, J., Yang, Y., McGehee, M. D., and Cui, Y. *Nano Lett.* **10**, 4242–4248 (2010).

[38] Ferry, V. E., Munday, J. N., and Atwater, H. A. *Adv. Mater.* **22**, 4794–4808 (2010).

[39] Catrysse, P. B. and Fan, S. *Nano Lett.* **10**, 2944–2949 (2010).

[40] Grosser, J. and Schulz, H. *J. Phys. D Appl. Phys.* **22**, 723 (1989).

[41] Bai, J., Zhong, X., Jiang, S., Huang, Y., and Duan, X. *Nat. Nanotechnol.* **5**, 190–194 (2010).

[42] Barnes, T. M. and Blackburn, J. L. In *Transparent Electronics,* Facchetti, A. and Marks, T. J., editors, 185–211. John Wiley & Sons, Ltd (2010).

[43] Luther, J. M., Law, M., Beard, M. C., Song, Q., Reese, M. O., Ellingson, R. J., and Nozik, A. J. *Nano Lett.* **8**, 3488–3492 (2008).

[44] Lin, R., Lu, Q., Ranade, P., King, T., and Hu, C. *IEEE Electr. Device L.* **23**, 49 –51 (2002).

[45] Wadhwa, P., Liu, B., McCarthy, M. A., Wu, Z., and Rinzler, A. G. *Nano Lett.* **10**, 5001–5005 (2010).

[46] Wadhwa, P., Seol, G., Petterson, M. K., Guo, J., and Rinzler, A. G. *Nano Lett.* **11**, 2419–2423 (2011).

[47] Liu, B., McCarthy, M. A., Yoon, Y., Kim, D. Y., Wu, Z., So, F., Holloway, P. H., Reynolds, J. R., Guo, J., and Rinzler, A. G. *Adv. Mater.* **20**, 3605–3609 (2008).

[48] McCarthy, M. A., Liu, B., Donoghue, E. P., Kravchenko, I., Kim, D. Y., So, F., and Rinzler, A. G. *Science* **332**(6029), 570–573 (2011).