Resonance-enhanced optical nonlinearity in the Weyl semimetal TaAs

Shreyas Patankar, Liang Wu, Baozhu Lu, Manita Rai, Jason D. Tran, T. Morimoto, Daniel E. Parker, Adolfo G. Grushin, N. L. Nair, J. G. Analytis, J. E. Moore, J. Orenstein, and D. H. Torchinsky

Phys. Rev. B **98**, 165113 — Published 10 October 2018

DOI: 10.1103/PhysRevB.98.165113

Resonance-enhanced optical nonlinearity in the Weyl semimetal TaAs

Shreyas Patankar,[1, 2] Liang Wu,[1, 2, 3] B. Z. Lu,[4, 5] M. Rai,[4, 5] J. D. Tran,[4, 5] T. Morimoto,[1] D. Parker,[1] Adolfo G. Grushin,[1, 6] N. L. Nair,[1] J. G. Analytis,[1, 2] J. E. Moore,[1, 2] J. Orenstein,[1, 2, *] and D. H. Torchinsky[4, 5, †]

[1] *Department of Physics, University of California, Berkeley, California 94720, USA*
[2] *Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3] *Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104*
[4] *Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, USA*
[5] *Temple Materials Institute, Temple University, Philadelphia, Pennsylvania 19122, USA*
[6] *Institut Néel, CNRS and Université Grenoble Alpes, F-38042 Grenoble, France*

(Dated: September 4, 2018)

The second-order conductivity of a material, $\sigma^{(2)}$, relating current to the square of electric field, is nonzero only when inversion symmetry is broken, unlike the conventional linear conductivity. Second-order nonlinear optical responses are thus powerful tools in basic research as probes of symmetry breaking; they are also central to optical technology as the basis for generating photocurrents and frequency doubling. The recent surge of interest in Weyl semimetals with acentric crystal structures has led to the discovery of a host of $\sigma^{(2)}$-related phenomena in this class of materials, such as polarization-selective conversion of light to dc current (photogalvanic effects) and the observation of giant second-harmonic generation (SHG) efficiency in TaAs at photon energy 1.5 eV. Here, we present measurements of the SHG spectrum of TaAs revealing that the response at 1.5 eV corresponds to the high-energy tail of a resonance at 0.7 eV, at which point the second harmonic conductivity is approximately 200 times larger than seen in the standard candle nonlinear crystal, GaAs. This remarkably large SHG response provokes the question of ultimate limits on $\sigma^{(2)}$, which we address by a new theorem relating frequency-integrated nonlinear response functions to the third cumulant (or "skewness") of the polarization distribution function in the ground state.

I. INTRODUCTION

Beginning with the work of Pierre Curie [1], a comprehensive theory of the role of symmetry in classifying the transport properties and optical response of materials has been developed. More recently there has been a growing appreciation that symmetry constraints combine with the geometry and topology of the relevant Hilbert space to shape the electrodynamic response functions of crystalline solids and artificial meta-crystals. [2] The newly discovered Weyl semimetals are currently a focus of intense research as examples of systems in which the combination of symmetry-breaking and bandstructure geometry may lead to novel and/or enhanced electronic response functions.

In the case of the Weyl semimetal TaAs, which is a focus of this work, breaking of inversion symmetry allows non-degenerate linearly dispersing electron bands to cross at isolated points, or Weyl nodes, in momentum space. Topologically, Weyl nodes are monopoles of the Berry curvature of the electron wavefunctions, [3] whose presence generates disconnected lines of Fermi contour (Fermi arcs) at the surface. [4] The existence of Weyl nodes in the bandstructure of TaAs was confirmed through the observation of the predicted Fermi arcs by angle-resolved photoemission. [5–7]

Following this demonstration, research has focused on discovering the defining electromagnetic response functions of Weyl semimetals. As the current linearly proportional to a single potential does not show very distinctive behavior, [8, 9] interest has focused on higher-order response functions, such as the electrical conductivity in the presence of a magnetic field [10–12] and the second-order optical conductivity, $\sigma^{(2)}$. The latter response function, which is allowed only in media that break inversion, describes the current generated to second-order in a time-varying electric field and quantifies a wide-variety of optical phenomena, including sum and difference frequency generation.

For excitation within a narrow band of frequency centered on ω_0, the second-order current is centered at $\omega \approx 0$ and $\omega \approx 2\omega_0$. The radiation arising from the frequency-doubled current is the phenomenon of second harmonic generation (SHG). The $\omega \approx 0$ (DC) current goes by a variety of names, including optical rectification, circular and linear photogalvanic effects [13–15], and shift current[16, 17] that reflect the variety of underlying mechanisms and dependences on the polarization state of the light at ω_0. Shift current, for example, denotes the contribution to the linear photogalvanic that is a property of the intrinsic bandstructure, as distinguished from extrinsic components that arise from asymmetric scattering off defects and phonons.

To date, much of the theoretical work on nonlinear optical effects in TaAs and other Weyl semimetals has focused on the role of Berry monopoles in generating quantized, and/or strongly enhanced second-order responses. The possibility of a quantized circular photogalvanic response arising from Berry monopoles was addressed by de Juan et al., [18] who found that quantization requires

* jworenstein@lbl.gov
† dtorchin@temple.edu

lowering of symmetry beyond inversion breaking such that all mirror planes are removed. Other theoretical predictions suggested that non-quantized circular and linear photogalvanic effects can be large, although their existence depends crucially on deviations from the simplest form of Weyl node, requiring either tilting [19] or curvature [20] of the bands near the crossing point. Experimentally, linear and circular photogalvanic currents have been detected in TaAs, [21, 22] by excitation with 100 meV photons. Ma et al. [21] determined the chirality of the Weyl nodes in TaAs from the sign of the photocurrent and reported a photocurrent amplitude smaller than the theoretical prediction of Ref. [19]. On the other hand Osterhoudt et al. [22] observed a large amplitude photogalvanic response in the same material, although using a different, smaller scale electrode configuration.

In this work, we focus on the companion phenomenon to intrinsic photogalvanic effects, namely the sum-frequency response that leads to SHG. This study was motivated by a previous report that TaAs, as well as closely related Weyl materials TaP and NbAs, exhibited the largest SHG response of any known crystal when photoexcited at 800 nm, which corresponds to a photon energy of 1.5 eV. [23] The existence of Berry monopoles is not expected to play a critical role in generating this phenomenon, as the energy scale for the Weyl-like electron dispersion does not exceed approximately 100 meV. In this paper we attempt to identify the factors that *do* contribute to the uniquely large SHG response of the TaAs family of Weyl semimetals.

In Section II we present measurements that spectrally resolve the SHG response function, extending previous results beyond the single 1.5 eV photon energy to a broad range from 0.5-1.5 eV. These measurements reveal that 1.5 eV actually lies in the high-frequency tail of a far stronger resonant response near 0.7 eV. In Section III we address the question of the roles of symmetry breaking and band geometry in determining the amplitude of the nonlinear optical response function. We demonstrate theoretically a general relationship, applicable to both SHG and shift current, between the strength of $\sigma^{(2)}$ and a measure of the polarity of the material. As TaAs is conducting, this measure cannot be simply the ground state polarization. We show instead that the correct measure of polarity for nonlinear properties is the skew of the polarization distribution function, which is embodied in a gauge-invariant cumulant of the bandstructure. In Section III we introduce a phenomenological description of the nonlinear response function that makes use of the analytical solution for a three-dimensional array of ferroelectric (Rice-Mele) chains.

II. EXPERIMENTAL RESULTS

Figure 1 shows a schematic of the apparatus used to measure the spectrum of the SHG response tensor, which is formally written as $\sigma_{ijk}(2\omega;\omega,\omega)$ and hereafter shortened to $\sigma^{\rm shg}_{ijk}(\omega)$, or simply σ_{ijk}. The measurements are enabled by a laser/optical parametric amplifier system that generates pulses of duration $\sim$50 fs at a 5 kHz repetition rate in the range of photon energies 0.5 - 1.5 eV. As the crystals are opaque (optical penetration depth of approximately 200 nm [24]) the SHG intensity is measured in reflection. As shown in the Figure 1a., a polarizer P, and a halfwave plate in front of the sample select the orientation of the electric field (referred to by $\hat{\mathbf{e}}$) of the fundamental light at frequency ω, and a second analyzing polarizer A selects the polarization of the light at frequency 2ω that reaches the detector.

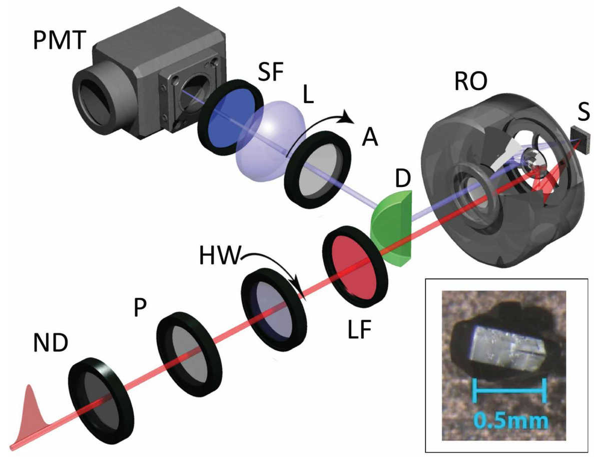

FIG. 1. **Schematic of the experimental apparatus used to study the spectrum of the nonlinear optical response in TaAs.** Laser pulses of pulse width 50 fs, at a repetition rate of 5 kHz with photon energy between 0.5 eV - 1.6 eV are generated by an optical parametric amplifier coupled to an regeneratively amplified laser. The incident radiation passes through a neutral density filter (ND), polarizer (P), and a low pass filter (LF), to lower the incident intensity, and to remove spurious polarizations, and wavelengths, respectively. A halfwave plate (HW) is used to rotate the incident linear polarization over 360^o. The incident laser radiation is then focused on to the TaAs sample (S) using a reflective objective (RO), which is insensitive to the photon energy, in contrast with conventional refractive microscope objectives. The SHG radiation from the sample is collimated by the RO, and picked off by a D-mirror (D), and passed through an analyzing polarizer (A), whose polarization state is determined by the scan being measured (see text). The radiation then passes through a stack of short pass filters (SF) to remove any incident light, and then focused by a lens (L) onto a photomultiplier tube (PMT) for detection. Inset: A single crystal of TaAs with the (112) facet visible.

The structure of TaAs (point group $4mm$) contains two perpendicular glide planes, which are equivalent to mirror planes for optical response functions [25, 26]. The normal directions of these two planes determine the two equivalent directions of the tetragonal unit cell, which we label as x and y. The direction perpendicular to the xy plane is the polar, or z axis. The $4mm$ point group allows three independent nonvanishing components of the conductivity tensor, σ_{zzz}, $\sigma_{zxx} = \sigma_{zyy}$, and $\sigma_{xzx} = \sigma_{yzy}$. To avoid using a surface obtained by cutting and subsequent polishing, which are known to affect the nonlinear optical response, we perform nonlinear reflectance measurements using the (112) surface, as it forms a naturally flat and smooth growth facet. Figure 1b shows a photograph of the (112) surface used for the measurement. The two high symmetry directions in the plane of this surface are [1,-1,0], which is perpendicular to the polar axis, and [1,1,-1].

We performed two pairs of polarization scans at each wavelength. In the first pair, $\hat{\mathbf{e}}$ and A are co-rotated with their relative angle fixed at 0 and 90°. This is equivalent to rotating the sample, but with the advantage that the position of the laser focus on the sample surface remains constant. These scans are primarily used to determine the high symmetry directions. However, as will be discussed below, they provide additional evidence for the existence of a sharp resonance in the nonlinear response function in the infrared. In the second pair of scans A is fixed parallel to either of the [1,-1,0] or [1,1,-1] high symmetry directions and $\hat{\mathbf{e}}$ is rotated over 360°. From this pair of scans we obtain the three combinations of tensor components that are available from reflection measurements performed on the (112) surface: $|\sigma_{zxx}|$, $|\sigma_{xzx}|$, and $|\sigma_{\text{eff}}| \equiv |\sigma_{zzz}+2\sigma_{zxx}+4\sigma_{xzx}|$ [27]. Analysis of the polarization scans at a given wavelength determines the relative magnitudes of these tensor components, i.e., $|\sigma_{zxx}|/|\sigma_{\text{eff}}|$ and $|\sigma_{xzx}|/|\sigma_{\text{eff}}|$.

Measurements of absolute, as opposed to relative, magnitudes of σ_{ijk} components over a broad range of frequency were accomplished by using GaAs to calibrate the response at 1.5 eV. Determining the effective $\sigma_{ijk}(\omega)$ at lower photon energies required characterization of the wavelength dependence of all components of the optical setup, such as wavelength-sorting filters, attenuators, and photodetectors, as well as spectral variation of the laser focal spot size and pulse duration. An extensive discussion of the calibration procedure is provided in Appendix A.

Polarization scans were performed at incident wavelengths in the interval from 800 to 2500 nm; a complete library of these data is available upon request. Figure 2(a-c) shows representative scans at wavelengths 800, 1560, and 2200 nm (1.55, 0.80, and 0.56 eV, respectively) with co-rotating parallel polarizations. The co-rotation plot at 800 nm illustrates the extreme anisotropy of the nonlinear optical response of TaAs. The solid curve through the data points is proportional to $|\sigma_{zzz}|^2 \cos^6\theta$, which is the dependence predicted for a crystal that responds only to E parallel to its polar axis and radiates second harmonic light that is likewise fully z-polarized. The polar plots at longer wavelength indicate that the amplitudes of the off-diagonal components, $|\sigma_{zxx}|$ and $|\sigma_{xzx}|$, begin to increase relative to $|\sigma_{\text{eff}}|$ as the fundamental photon energy approaches 0.7 eV from above, although they remain approximately an order of magnitude smaller, as is shown below.

Figure 3 illustrates the spectral dependence of the SHG response function of a TaAs crystal at room temperature. Figure 3a shows $|\sigma_{\text{eff}}|$, $|\sigma_{zxx}|$, and $|\sigma_{xzx}|$ vs. fundamental photon energy as solid circles (the solid lines are guides to the eye). The y-axis scale expresses σ^{shg} in units of the conductance quantum per Volt, to facilitate comparison with theory. In SI units, the corresponding peak value of σ^{shg} at 0.7 eV is approximately 5×10^{-3} $(\Omega\ \text{V})^{-1}$. The dashed line is the spectrum of σ_{xyz} in GaAs as reported in Ref. [28], multiplied by a factor 10 so that it can be compared with the TaAs spectra. As reported previously, the SH response measured at 1.5 eV in TaAs exceeds the peak response of GaAs by a factor ~ 10. The measurements reported here show that $|\sigma_{\text{eff}}|$ becomes even larger with decreasing frequency, reaching a peak near 0.7 eV where it is $\sim 2 \times 10^2$ larger than the maximum response of GaAs.

Although the diagonal response function $|\sigma_{zzz}|$ is not determined directly from the SHG intensity, its range of values is highly constrained through analysis of the polarization scans [27]. The shaded region in Figure 3b illustrates the upper and lower bounds on $|\sigma_{zzz}|$. The dashed line that lies within the shaded region is a best fit to a phenomenological model described below, which is seen to capture the basic features of the resonance enhancement of $|\sigma_{zzz}|$.

III. NONLINEAR CONDUCTIVITY SUM RULE

We now turn to a theoretical discussion of the fundamental causes for this giant second harmonic response. As this response is large it is reasonable to start with a discussion of bounds or ultimate limits on SHG, which is relevant not only to SHG, but to sensitivity of photodetectors and efficiency of solar cells based on intrinsic photogalvanic effects as well. As with linear response, the true measure of the strength of the nonlinearity is the frequency-integrated response function, rather than its amplitude at any single frequency. In the case of linear response, the frequency integral corresponds to the well-known f-sum rule. We have discovered an analogous sum rule that connects second harmonic generation and the modern theory of polarization. The rule is elucidated in this section, then in the subsequent section we see it in action with a minimal model for TaAs.

The SHG response in Fig. 3 appears to consist of a single peak and thus we may assume that it arises from a single resonance between two bands. For the two-band case, there is close relationship between the current generated at twice the excitation frequency and at zero frequency. The real part of the zzz component of the SHG tensor can be expressed in terms of the shift current tensor as follows [29],

$$\text{Re}\left\{\sigma_{zzz}^{\text{shg}}(\omega)\right\} = -\frac{1}{2}\sigma_{zzz}^{\text{shift}}(\omega) + \sigma_{zzz}^{\text{shift}}(2\omega). \qquad (1)$$

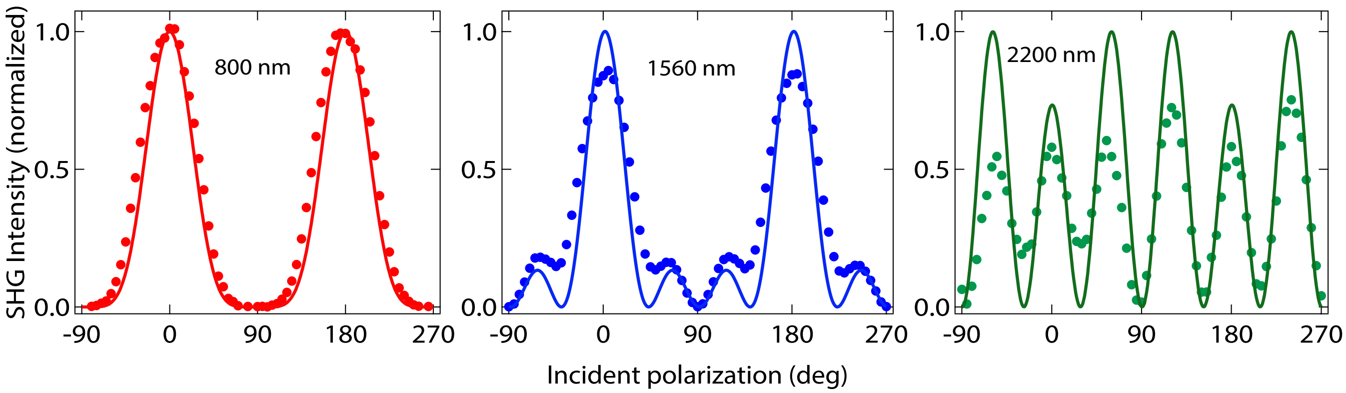

FIG. 2. **Second harmonic generation polarimetry.** SHG intensity measured in the "parallel" scan as a function of the incident polarization angle, plotted on a normalized scale, for three different incident photon wavelengths. The polar pattern changes from a two-fold pattern at shorter wavelengths (800 nm=1.55 eV) to a six-fold pattern at longer wavelengths (2200 nm=0.56 eV)

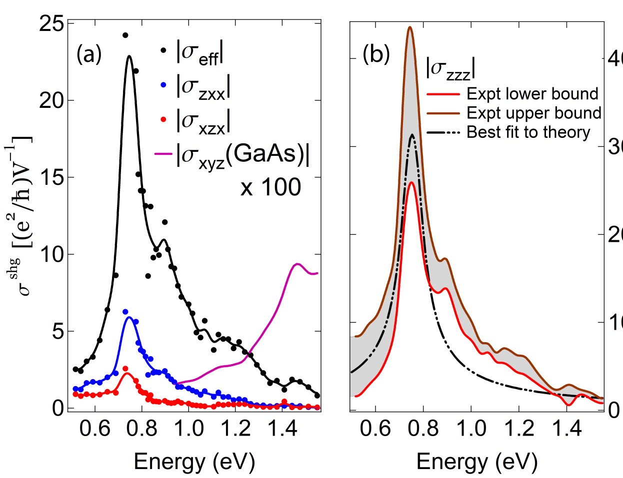

FIG. 3. **Measured nonlinear optical conductivity as a function of incident photon energy.** (a) Spectra of the conductivity components $|\sigma_{zxx}|$, $|\sigma_{xzx}|$, and $|\sigma_{\mathrm{eff}}| = |\sigma_{zzz} + 4\sigma_{xzx} + 2\sigma_{zxx}|$. The solid lines are guides to the eye. The nonlinear optical conductivity of GaAs, $|\sigma_{xyz}|$ is multiplied by 100 and plotted for comparison. (b) The (expt) lower and upper bounds of $|\sigma_{zzz}|$ are estimated using measured values of $|\sigma_{\mathrm{eff}}|$, $|\sigma_{zxx}|$, and $|\sigma_{xzx}|$ from experiment. The dashed black line depicts the best fit to $|\sigma_{zzz}|$ using the phenomenological model described in the text.

We use a shorthand notation $\sigma^{\mathrm{shg}}_{zzz}(\omega)$ for the second-harmonic response $\sigma_{zzz}(2\omega;\omega,\omega)$. Similarly, $\sigma^{\mathrm{shift}}_{zzz}(\omega)$ represents the shift current response, $\sigma_{zzz}(0;\omega,\omega)$. The above relation expresses a fundamental link between the SHG and shift current response functions for the two-band case.

Equation (1), which is derived in Appendix B.2, shows that the SHG response is the sum of two terms that are opposite in sign but have the same functional form when the frequency argument is scaled by a factor of two. The first ("one-photon") term corresponds to the resonant condition $\hbar\omega = E_{\mathrm{gap}}(k)$ and the second ("two-photon") term corresponds to $\hbar\omega = E_{\mathrm{gap}}(\mathbf{k})/2$, where $E_{\mathrm{gap}}(\mathbf{k})$ is the energy gap at wavevector $\mathbf{k}$. Clearly the total $\sigma^{\mathrm{shg}}_{zzz}$ response integrates to zero, but we shall show below that for some models each shift current term has a geometrical interpretation in terms of the skewness or intrinsic asymmetry of the ground-state polarization distribution.

In Appendix C we derive the following sum rule for the shift current conductivity in the two-band approximation, which holds regardless of dimensionality d and the range of electron hopping amplitude,

$$\Sigma^{\mathrm{shift}}_{z} \equiv \int \sigma^{\mathrm{shift}}_{zzz}(\omega)\, d\omega = \frac{2\pi e^3}{V\hbar^2} C_3, \tag{2}$$

where,

$$C_3 \;=\; -\frac{V}{(2\pi)^d}\int d^dk\ \mathrm{Im}\left[c_3 - 3c_2c_1 + 2c_1^3\right], \tag{3}$$

$c_n \equiv \langle u_0(k)|(i\partial_k)^n|u_0(k)\rangle$, and the periodic gauge is assumed for the valence-band Bloch wavefunction $\psi_{0k}(r) = u_{0k}(r)e^{ikr}$. The quantity C_3 is a member of a set of gauge invariant quantities, C_n, that are cumulants of the electronic polarization [30, 31]. Briefly, the quantity C_1 is exactly the average macroscopic polarization, which coincides with the first moment of the polarization distribution [32]. Accordingly, C_3 is the third cumulant, or "skewness" of the distribution, which vanishes in the presence of inversion symmetry. C_1 is controlled by the center-of-mass location of the polarization, while C_3 describes the intrinsic asymmetry in the shape of the polarization distribution, independent of its center-of-mass location.

In addition to providing a very satisfying connection between the nonlinear response and ground state fluctuations of the polarization, equation (2) speeds the search for effective Hamiltonians with large nonlinear response. If a material is known to have a large skew in its polarization distribution, then the sum rule guarantees it will have a large shift current and, in turn, large second harmonic generation. In the next section we will see this in action with a minimal model for TaAs.

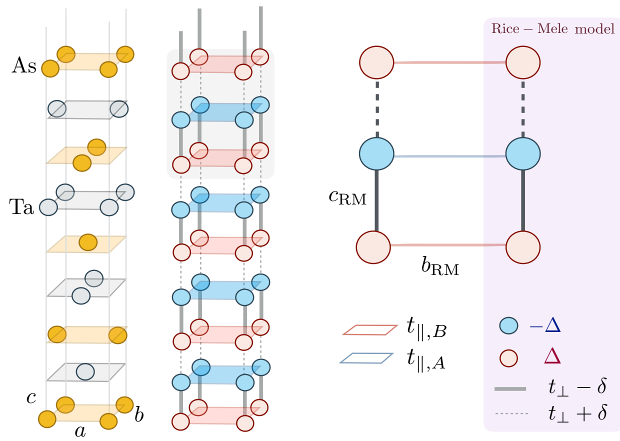

FIG. 4. **The Rice-Mele model.** Depiction of the model of coupled Rice-Mele chains (center) used to reproduce the experimentally observed second harmonic generation in TaAs (left). A zoom on the shaded unit-cell is depicted on the right specifying the different hopping terms present in the model.

IV. PHENOMENOLOGICAL MODEL OF THE SHG SPECTRUM

Under the assumption that TaAs is a weakly interacting system, $\sigma^{\rm shg}_{ijk}(\omega)$ can in principle be calculated by combining *ab initio* bandstructure and the general theory of the second order nonlinear response function [17]. In reality this is a challenging calculation, as it requires knowledge of the gradients of the wavefunction throughout the Brillouin zone with high resolution in momentum. Nevertheless, a first principles calculation of $\sigma^{\rm shift}_{ijk}(\omega)$ has been published recently[33], although restricted to the energy range from 20-200 meV and thus not directly relevant to our measurements. Given the lack of a first principles calculation for energies of order 1 eV, we present below a phenomenological theory based on an analytical solution for a model Hamiltonian. This Hamiltonian is not intended as a tight-binding parameterization of the TaAs bandstructure, rather our purpose is to highlight some of the features that a successful first principles theory must reproduce.

The key features of the nonlinear conductivity in TaAs – strong anisotropy and a single, sharp resonant peak – suggest that a quasi-one dimensional, two-band model is sufficient to capture much of the physics involved. The minimal model required to reproduce this physics is a modified version of the Rice-Mele (RM) Hamiltonian [34], which is a one-dimensional (1D) tight-binding model of a semiconductor with broken inversion symmetry. The RM model played a crucial role in the development of the "modern theory of polarization"[31, 32, 35–38] and, more recently, was shown to describe the shift current spectrum in monochalcogenide semiconductors [39]. We believe that such a semiconductor-based model is appropriate because the energies probed in our experiment are much larger than either the Fermi energy [5, 7] or plasma frequency [40] of TaAs. To describe the nonlinear response of a three dimensional system, we consider an array of RM chains aligned in the z direction with an interchain coupling. To be clear, we are not suggesting that the resulting quasi-one dimensional model reproduces the bandstructure of TaAs; rather it serves as a minimal model capable of reproducing the essential features of the resonant nonlinear response found experimentally.

The RM Hamiltonian is,

$$H_{RM} = t\cos(k_z a/2)\sigma_x + \delta \sin(k_z a/2)\sigma_y + \Delta\sigma_z, \quad (4)$$

where the σ_i are the Pauli matrices. As shown in Figure 4, H_{RM} describes a 1D chain of atoms along the z direction, in which inversion symmetry is broken by alternating on-site energies ($\pm\Delta$) and hopping amplitudes ($t \pm \delta$). Despite its relative simplicity, the RM model has played a key role in the development of the modern theory of polarization [32] and recently in strategies to enhance nonlinear optical response functions [41].

The optical response of the RM Hamiltonian captures qualitative features of the nonlinear optical response in TaAs, as $\sigma^{\rm shg}_{ijk}(\omega)$ is polarized parallel to the polar axis and exhibits a sharp peak at the threshold for absorption. The RM model can be modified to include an interchain coupling to smooth the 1D van Hove singularity that is obtained for an isolated chain. To extend the model to 3D, we consider an array of RM chains parallel to z and allow electrons to hop laterally in the A and B atom layers (see Figure 4). The interchain hopping adds an additional term $t_{AB}(\cos k_x a_x + \cos k_y a_y)$ to the coefficient of σ_z in H_{RM}, where $t_{AB} \equiv t_{\parallel,A} - t_{\parallel,B}$. The shift current spectrum in this model can be calculated following [42] (see [27]),

$$\sigma^{\rm shift}_{zzz}(\omega) = \frac{2e^3}{\hbar\Delta}\left(\frac{1}{4\pi}\right)^3 \frac{c^2}{ab} F(\tilde{\omega}; \tilde{\delta}, \tilde{t}, \tilde{t}_{AB}), \quad (5)$$

where F is a dimensionless function of frequency ω and the parameters of H_{RM} (the tildes indicate normalization by Δ); a, b, and c are lattice constants in the x, y, and z directions, respectively [27]. The real part of the SHG conductivity can easily be expressed in terms of the shift conductivity by using Eq. (1).

The dash-dotted curve in Figure 3b is a fit to $|\sigma^{\rm shg}_{zzz}|$ based on the two-photon term in Eq. (1). Despite the simplicity of the model relative to the complexity of the TaAs bandstructure, the quasi-one dimensional version of the RM model describes the data remarkably well. The peak position, low, and high energy tails of the spectrum are best described using parameters $\tilde{t} = 1.5$, $\tilde{\delta} = 1.4$, $\tilde{t}_{AB} = 0.02$, and $\Delta = 0.428$. The corresponding bandstructure consists of relatively flat valence and conduction bands separated by about 1.4 eV, as is required to produce a the narrow resonance near 0.7 eV. Finally, we use the TaAs lattice parameters, $c = 1.165$ nm, and $b = a = 0.344$ nm as input to the dimensionless geometric factor c^2/ab in Eq. 5.

We note that our minimal model of the SHG resonance implies a corresponding resonant peak in the linear conductivity $\sigma_{zz}(\omega)$ near 1.4 eV. The ratio of the spectral weight of the nonlinear to linear resonance is $(e/\hbar\omega_0)\,(C_3/C_2)$, and in the RM model $C_3/C_2 = 2c/3$. Based on this ratio, and the fact that the width of the nonlinear resonance peak is half that of the linear one, the model predicts $\sigma_{zz}(\omega) \approx 2 \times 10^4\ \Omega^{-1}$-cm^{-1} at the peak energy of 1.4 eV. This prediction is as yet to be tested, as to date the linear optical conductivity in TaAs has been measured by reflection from the (001) surface, probing only $\sigma_{xx}(\omega)$ or $\sigma_{yy}(\omega)$. Measuring $\sigma_{zz}(\omega)$ on TaAs is not straightforward as this requires a large (100) surface which does not arise naturally during growth. Moreover it has been found that cutting and polishing the surface of TaAs for reflectivity measurement results in a loss of the bulk property.

V. SUMMARY AND CONCLUSIONS

In Section II of this paper we reported SHG spectra of TaAs, demonstrating a ten-fold resonant enhancement of the response at 0.7 eV as compared with the previously reported, and already quite large, response at 1.5 eV. The large amplitude led us to consider bounds on the strength of optical nonlinearity, as embodied in the frequency-integrated second-order conductivity (or spectral weight). In Section III we presented a new theorem that links the spectral weight of the nonlinear conductivity to the geometry of the bandstructure. The link is closely related to the connection between geometry and ferroelectricity in the “modern theory of polarization.” We showed that for general two-band models the spectral weight is proportional to the skew of the polarization distribution, which itself is proportional to the third gauge-invariant cumulant, C_3. This result is analogous to the relation between the polarization and C_1, which is the integral of the Berry connection over the Brillouin zone.

In Section IV we combined the insight gained from the spectral weight theorem with a model of a quasi-1D polar semiconductor to identify the factors contributing to large SHG in TaAs. Specifically, we introduced the Rice-Mele chain, which is parameterized by alternating on-site energies $(\pm\Delta)$ and nearest-neighbor hopping amplitudes $(t \pm \delta)$. We note that although this model is very useful, the relationship between the RM phenomenology and the actual electronic wavefunctions and bandstructure of TaAs remains to be understood.

We found that the third-cumulant for the RM model (and indeed all nearest-neighbor 1D Hamiltonians) is bounded, such that $C_3 \leq 0.3c^2$ (for each spin), where c is the dimension of the 1D unit cell [27]. For a 3D array of weakly interacting RM chains, the spectral weight of $\sigma^{shg}_{zzz}(\omega)$, or Σ_z is bounded by,

$$\Sigma_z \leq \frac{e^2}{\hbar} C_3 P_z, \tag{6}$$

where $P_z \equiv ec/V$ is the “polarization quantum” and V is the 3D unit cell volume.

Maximizing Σ_z requires, first, a strongly polar chain, that is one that approximately saturates the bound $C_3/c^2 \leq 0.3$. For the RM model, strong polarity occurs when the parameters t, δ, and Δ are roughly equal, as shown in Ref. [43]. A second large contributing factor is the dimensionless ratio of the square of the lattice parameter along the chain axis, c^2, to area per chain, ab. We note that as there are four Ta-As chains per unit cell, the spatial packing factor c^2/ab is large, ≈ 45. In addition, the peak SHG response will become large when the total spectral is concentrated in a narrow resonance, as will occur in the RM parameterization for $t \approx \delta$, *i.e.*, the limit of weakly interacting dimers. Finally, we note that the bound on C_3 obtained in nearest-neighbor models is not an ultimate limitation on nonlinear response. This bound is exceeded when next-neighbor hopping is introduced, and can diverge when the fundamental gap is driven to zero (see Appendix C.5). This result is consistent with the recent analysis of Tan and Rappe [41], who suggested that Σ_z can be greatly enhanced when the range of inter-site hopping becomes larger than the lattice parameter.

VI. ACKNOWLEDGEMENTS

Measurements and modeling were performed at the Lawrence Berkeley National Laboratory in the Quantum Materials program supported by the Director, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Spectroscopy measurements were performed at the Department of Physics, Temple University. The authors would like to thank B. Xu for sharing ellipsometry data of TaAs. We would also like to thank Balasz Hetényi for helpful communication. J.O. and L.W. received support for performing and analyzing optical measurements from the Gordon and Betty Moore Foundation’s EPiQS Initiative through Grant GBMF4537 to J.O. at UC Berkeley. Sample synthesis by N.N. and J.A. was supported by the Gordon and Betty Moore Foundations EPiQS Initiative through Grant GBMF4374. Work by N.N. and J.A. was supported by the Office of Naval Research under the Electrical Sensors and Network Research Division, Award No. N00014-15-1-2674. T.M. and J.E.M. were supported by the Quantum Materials program under the U.S. Department of Energy. T.M was supported by the Gordon and Betty Moore Foundation’s EPiQS Initiative Theory Center Grant to UC Berkeley. A.G.G. was supported by the

Marie Curie Program under EC Grant agreement No. 653846. J.E.M. received support for travel from the Simons Foundation. D.H.T. acknowledges startup funds from Temple University. D.P. received support from the NSF GRFP, DGE 1752814. The authors would like to thank Nobumichi Tamura for his help in performing crystal diffraction and orientation on beamline 12.3.2 at the Advanced Light Source. N. Tamura and the ALS were supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Appendix A: Data acquisition and processing

1. Extracting susceptibility components from SHG polar patterns

To calculate the components of the third rank susceptibility tensor $\overleftrightarrow{\chi}^{(2)}$ from the measured data, we start with the relation defining the second harmonic susceptibility in terms of the electric fields of the incident $\mathbf{E}^{1\omega}$ and radiated $\mathbf{E}^{2\omega}$ fields:

$$E_i^{2\omega} = \chi_{ijk}^{\text{shg}} E_j^{1\omega} E_k^{1\omega}, \tag{A1}$$

where a sum over repeated indices is implied. We use the notation $\hat{x}, \hat{y}, \hat{z}$, to correspond to the crystalline axes $[1,0,0]$, $[0,1,0]$, $[0,0,1]$, and for brevity, we shall drop the superscript "shg" in the following. The crystal structure of TaAs corresponds to the point group $4mm$, which allows only three distinct components of the χ_{ijk} tensor: χ_{zzz}; $\chi_{zxx} = \chi_{zyy}$; $\chi_{xxz} = \chi_{xzx} = \chi_{yyz} = \chi_{yzy}$. As noted in the main text, the experiments were performed by measuring the second harmonic intensity with light incident on the naturally occuring (112) facet of a TaAs single crystal. The SHG intensity was measured in three polarization "channels":

1. *parallel*: the polarization state of the analyzer is set to be parallel to the polarization state of the incident radiation, while both rotate from $0°$ to $360°$
2. *vertical*: the polarization state of the analyzer is set to be parallel to the in-plane polar direction $(1,1,-1)$ of the TaAs crystal, while the polarization state of the incident radiation rotates from $0°$ to $360°$
3. *horizontal*: the polarization state of the analyzer is set to be perpendicular to the in-plane polar direction $(1,1,-1)$ of the TaAs crystal, while the polarization state of the incident radiation rotates from $0°$ to $360°$

It is useful to label the in-plane component of the polar axis $[1,1,-1]$ as the axis "γ", and correspondingly label the perpendicular in-plane direction $[-1,1,0]$ as the axis "α". If the incident light has intensity I_0, with linearly polarized electric field expressed by $\mathbf{E}^{1\omega} = \sqrt{I_0}\,(\hat{\alpha}\sin\theta + \hat{\gamma}\cos\theta)$, in the parallel channel, the intensity can be expressed in terms of the angle θ of the incident polarization state with respect to the "γ" axis:

$$\begin{aligned} I_{\text{para}}(\theta) &= \left|E_\alpha^{2\omega}\sin\theta + E_\gamma^{2\omega}\cos\theta\right|^2 \\ &= \left|\frac{1}{\sqrt{2}}\left(-E_x^{2\omega} + E_y^{2\omega}\right)\sin\theta + \frac{1}{\sqrt{3}}\left(E_x^{2\omega} + E_y^{2\omega} - E_z^{2\omega}\right)\cos\theta\right|^2. \end{aligned}$$

We wish to express this in terms of the tunable quantities $E_z^{1\omega} = -\frac{E_\gamma^{1\omega}}{\sqrt{3}}$ and $E_\gamma^{1\omega} = \sqrt{I_0}\cos\theta$. To do this, we first expand $\boldsymbol{E}^{2\omega}$ in terms of $\boldsymbol{E}^{1\omega}$ via $E_x^{2\omega} = 2\chi_{xxz}E_x^{1\omega}E_z^{1\omega}$, $E_y^{2\omega} = 2\chi_{yyz}E_y^{1\omega}E_z^{1\omega}$, $E_z^{2\omega} = \chi_{zxx}\left(E_x^{1\omega}\right)^2 + \chi_{zyy}\left(E_y^{1\omega}\right)^2 + \chi_{zzz}\left(E_z^{1\omega}\right)^2$. and transform back from the xyz to the $\alpha\gamma$ basis via $E_x^{1\omega} = -\frac{E_\alpha^{1\omega}}{\sqrt{2}} + \frac{E_\gamma^{1\omega}}{\sqrt{3}}$, $E_y^{1\omega} = \frac{E_\alpha^{1\omega}}{\sqrt{2}} + \frac{E_\gamma^{1\omega}}{\sqrt{3}}$, $E_z^{1\omega} = -\frac{E_\gamma^{1\omega}}{\sqrt{3}}$. This yields

$$\begin{aligned} I_{\text{para}}(\theta) &= \frac{I_0^2}{3}\left|\chi_{xxz}\cos\theta\sin^2\theta + \left(\frac{4}{3}\chi_{xxz}\cos^2\theta + \chi_{zxx}\left(\sin^2\theta + \frac{2}{3}\cos^2\theta\right) + \frac{1}{3}\chi_{zzz}\cos^2\theta\right)\cos\theta\right|^2 \\ &= \frac{I_0^2}{27}\left|3\chi_{xxz}\cos\theta\sin^2\theta + \left(4\chi_{xxz}\cos^2\theta + \chi_{zxx}\left(3\sin^2\theta + 2\cos^2\theta\right) + \chi_{zzz}\cos^2\theta\right)\cos\theta\right|^2 \\ &= \frac{I_0^2}{27}\left|(6\chi_{xxz} + 3\chi_{zxx})\cos\theta\sin^2\theta + (4\chi_{xxz} + 2\chi_{zxx} + \chi_{zzz})\cos^3\theta\right|^2. \end{aligned}$$

We define $\chi_{\text{eff}} = \chi_{zzz} + 2\chi_{zxx} + 4\chi_{xxz}$, which gives us the expression,

$$I_{\text{para}} = \frac{I_0^2}{27}\left|3(2\chi_{xxz} + \chi_{zxx})\cos\theta\sin^2\theta + \chi_{\text{eff}}\cos^3\theta\right|^2. \tag{A2}$$

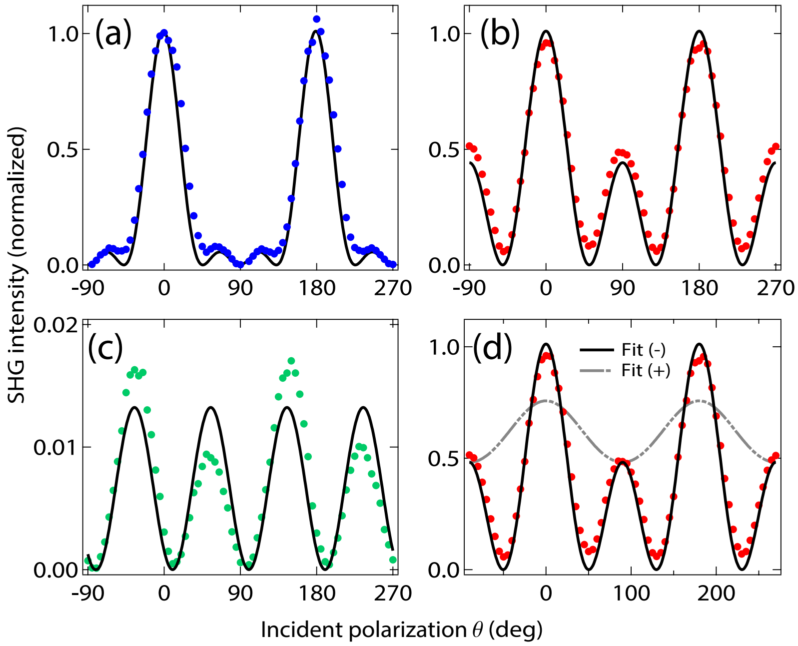

FIG. 5. Minimized mean-squared deviation fits (solid lines) to the SHG intensity data (points) as a function of the incident polarization angle θ, in three polarization channels (a) parallel, (b) vertical, (c) horizontal, for incident photon wavelength $\lambda = 1400$nm, plotted on a normalized scale. (d) Fits to the SHG intensity in the vertical channel assuming the relative sign between χ_{zxx} and χ_{zzz} to be $(+)$ (dashed line), and $(-)$ (solid line).

We can similarly obtain expressions for the other two channels:

$$I_{\text{vertical}}(\theta) = \left|E_3^{2\omega}\right|^2 = \left|\frac{1}{\sqrt{3}}\left(E_x^{2\omega} + E_y^{2\omega} - E_z^{2\omega}\right)\right|^2 = \frac{1}{27}\left(3\chi_{zxx}\sin^2\theta + \chi_{\text{eff}}\cos^2\theta\right), \tag{A3}$$

and

$$I_{\text{horiz}}(\theta) = \left|E_\alpha^{2\omega}\right|^2 = \left|\frac{1}{\sqrt{2}}\left(-E_x^{2\omega} + E_y^{2\omega}\right)\right|^2 = \frac{1}{3}I_0^2|\chi_{xxz}|^2\ \sin 2\theta. \tag{A4}$$

Figure 5d shows the SHG intensity data as a function of the polarization of incident photons in the parallel, vertical, and horizontal channel, with fits to the expressions in equation (A2), (A3), (A4) respectively. The amplitudes of the fitting parameters $|\chi_{xxz}|, |\chi_{zxx}|, |\chi_{\text{eff}}|$ thus obtained are then multiplied by various wavelength-dependent correction factors described below.

Although the fitting scheme described above is not sufficient to obtain the complete amplitude and phase information for the three components $|\chi_{xxz}|, |\chi_{zxx}|, |\chi_{\text{eff}}|$, we are nevertheless able to definitively say that for all incident photon energies, χ_{zxx} and χ_{zzz} have a relative phase of π. Figure 5 shows the best fits to the data at incident photon wavelength $\lambda = 1400$ nm in the "vertical" channel, using two contrasting assumptions for the relative signs of the fitting parameters, $(+1)$ or (-1). The fits with relative $(-)$ signs are are observed to be better approximations to the data.

With the additional information about the relative signs, the calculated values of $|\chi_{\text{eff}}| = |\chi_{zzz} + 2\chi_{zxx} + 4\chi_{xxz}|$ can be used to place bounds on the values of $|\chi_{zzz}|$ with an upper bound (UB) and lower bound (LB) expressed as (see figure 3 in the main text):

$$|\chi_{zzz}^{UB}| = |\chi_{\text{eff}}| + 2|\chi_{zxx}| + 4|\chi_{xxz}| \tag{A5}$$

$$|\chi_{zzz}^{LB}| = |\chi_{\text{eff}}| + 2|\chi_{zxx}| - 4|\chi_{xxz}|. \tag{A6}$$

2. Wavelength dependent correction factors

We provide a brief summary of the factors that must necessarily be taken into consideration to accurately determine the second harmonic susceptibility tensor elements χ_{ijk}^{shg} and the procedure we followed to experimentally measure them.

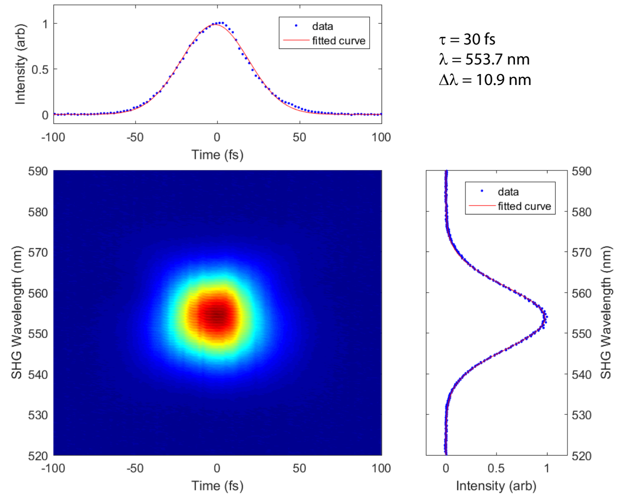

FIG. 6. Measured FROG spectrum for input fields $\lambda = 1100$ nm (nominally) showing the projection of the data along the temporal (above) and spectral (right) axes. Fits of the projected data to Guassian models, as described in the text, produced the observed pulse duration τ, the center SHG wavelength λ and the spectral width of the pulse $\Delta\lambda$.

The electric field of the incident laser pulse can be well approximated by a simple Gaussian field as

$$\begin{aligned}\mathbf{E}_\omega(\mathbf{r},t) \;=\; & E_0 e^{-i(k_0 z-\omega t)} \exp\left(-\frac{x^2+y^2}{w_0^2}\right) \\ & \times \exp\left[-\left(\frac{z-ct}{c\tau}\right)^2\right]\hat{x}\end{aligned} \tag{A7}$$

where E_0 is the electric field amplitude, k_0 is the field wavevector magnitude, ω is the fundamental light frequency, c is the speed of light, and w_0 is the electric field beam waist [44], here taken as the point from which the SHG field arises. We have assumed that the beam is transversely polarized purely in the $\hat{x}$ direction for convenience. As a consequence of Poynting's Theorem, the relationship between the integrated intensity of the incident, fundamental frequency field I_ω and the associated electric field can be given by

$$I_\omega = \frac{1}{2}\epsilon_0 \int \left\langle |\mathbf{E}_\omega(\mathbf{r},t)|^2 \right\rangle d\mathbf{r} \tag{A8}$$

where the brackets denote time averaging. For the field as given by equation (A7), the integration is easily performed to yield

$$I_\omega \propto E_0^2 w_0^2 \tau \tag{A9}$$

The penetration depth of the wavelengths used in this study is relatively short in TaAs. Thus we may consider the limit of non-depleted incident fields. The radiated second harmonic electric field is then proportional to the second-order induced polarization as $\mathbf{E}_{2\omega}(\mathbf{r},t) \propto \mathbf{P}_{2\omega}(\mathbf{r},t) = \epsilon_0 \overleftrightarrow{\boldsymbol{\chi}}^{\mathrm{shg}} \mathbf{E}_\omega(\mathbf{r},t)\mathbf{E}_\omega(\mathbf{r},t)$. Hence, the intensity of the second harmonic field $I_{2\omega}$ is given by

$$\begin{aligned} I_{2\omega} &= \frac{1}{2}\epsilon_0 \int \left\langle |\mathbf{E}_{2\omega}(\mathbf{r},t)|^2 \right\rangle d\mathbf{r} \\ &\propto \int \left\langle |\overleftrightarrow{\boldsymbol{\chi}}^{\mathrm{shg}}|^2 |\mathbf{E}_\omega(\mathbf{r},t)|^4 \right\rangle d\mathbf{r} \\ &\propto (E_0^2 \chi_{ijk}^{\mathrm{shg}})^2 w_0^2 \tau. \end{aligned} \tag{A10}$$

When this latter result is combined with equation (A9) to eliminate the electric field E_0, we determine that $I_{2\omega} \propto (\chi_{ijk}^{\mathrm{shg}})^2 I_\omega^2 / w_0^2 \tau$. However, for Gaussian fields such as in equation (A7), the beam waist at the focus w_0 can be described in terms of the incident beam diameter d, lens focal length f and beam wavelength λ as $w_0 \propto f\lambda/d$, yielding

$$I_{2\omega} \propto (\chi_{ijk}^{\mathrm{shg}})^2 \frac{I_\omega^2 d^2}{f^2 \lambda^2 \tau}. \tag{A11}$$

Therefore, accurate measurement of the elements of χ^{shg} must be controlled for incident integrated pulse intensity (equivalently, energy), diameter, wavelength and duration as well as the extrinsic factor of the lens focal length. Since the value for $\chi_{ijk}^{\mathrm{shg}}$ at 800 nm is known from previous study [23], all measurements were normalized to our measured response at 800 nm. In addition, the relative enhancement of second harmonic response between TaAs and GaAs was checked independently using a two color Er:doped fiber laser with source photon wavelengths of 1550 nm and 780 nm respectively.

3. Experimental measurement of correction factors

a. Intrinsic factors

Even though a commercially supplied wavelength separator was employed, a number of parasitic wavelengths were present due to a combination of leakage of other responses from the Optical Parametric Amplifier, as well as their sum-frequency, difference-frequency, and second harmonic responses among the signal, idler and second stage pump fields. At each individual wavelength, we computed these parasitic wavelengths and assembled a combination of longpass and shortpass filters of minimum ND8 to remove the parasitics before measuring the incident power. This power was recorded after the last longpass filter (LF in figure 1 of the main text), and a set of reflective neutral density filters of value ND chosen to maintain a calculated incident fluence of roughly ~ 20 mJ/cm^2. As the measured average power P is simply proportional to the energy per pulse through a factor of the repetition rate, the value $P/10^{ND}$ was divided from the measured power to provide a measure of the incident integrated intensity I_ω.

The spectral and temporal characteristics of the OPA beam were measured using a homebuilt Frequency Resolved Optical Gating (FROG) apparatus over the range

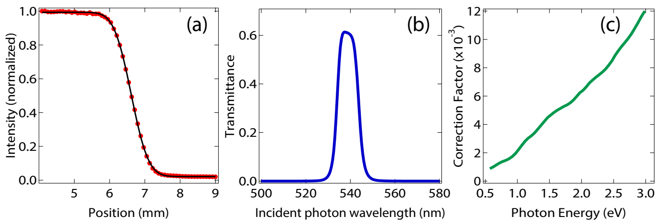

FIG. 7. (a) Representative data from knife-edge measurement of the beam diameter and corresponding fit for $\lambda_0 = 1100$ nm incident light. (b) Experimentally measured transmittance of the 540 nm bandpass filter. (c) The linear optical correction factor $BP(\omega)$ as a function of incident photon energy.

from incident wavelength $\lambda = 800$ nm to $\lambda = 2000$ nm. In all cases, the FROG measurement was performed on the beam before its entrance into the reflective objective (i.e., in between LF and RO in figure 1 of the main text) in order to account for the filtering and dispersive characteristics of the preceding optics. While we did not use the same ND filters for the FROG measurement as we did in the experiments, we did use the same number of filters of identical material (UV grade fused silica) and thickness as those used in data collection. A representative FROG trace is shown in figure 6. The FROG apparatus used a 100 μm thick β-Barium Oxide (BBO) crystal, the radiated second harmonic spectrum was determined via simulation with lab2.de software[45] to be virtually identical to that which would derive from a thin layer of a bulk response reflection experiment as occurred here on TaAs.

The pulse duration of the incident second harmonic field was determined from numerically summing the FROG trace along the spectral dimension and fitting the resulting data to a Gaussian of the form $I_{2\omega}(t) = I_0 \exp\left[-\left((t-t_0)/\tau\right)^2\right]$, as shown in the top panel of figure 6. We note that from evaluation of the convolution of a Gaussian with itself, the pulse duration at 2ω is related to that at ω through a numerical factor as $\tau_{2\omega} = \tau_\omega/\sqrt{2}$. In order to determine the radiated central 2ω wavelength and bandwidth, we numerically summed the FROG trace along the temporal dimension and fit the result to a Gaussian of the form $I_{2\omega}(\lambda) = I_0 \exp\left[-\left((\lambda-\lambda_0)/\Delta\lambda\right)^2\right]$, as shown on the righthand panel. These data allowed us to determine the fundamental wavelength as $2\lambda_0$ and the spectral bandwidth of the SHG pulse $\Delta\lambda$.

The knife-edge method was used to measure the beam diameter w_0 immediately before the reflective objective as a function of incident wavelength λ for all wavelengths in the study. Representative data at $\lambda = 1100$ nm and the corresponding fit to the complementary error function $I(x) = I_0 \,\mathrm{erfc}\left(\sqrt{2}(x-x_0)/w_0\right)$ are shown in figure 7. A basic Gaussian beam propagation could be used to determine the beam at the focus. However, this was unnecessary since the identical, all-reflective optical setup was used for every data point, necessitating only measurement of the beam diameter w_0.

b. Extrinsic factors

Experimentally, the measured second harmonic intensity is also proportional to the spectral response of the detector $\mathcal{D}(\lambda)$ and the transmittance $\mathcal{T}(\lambda)$ of the filters in front of the detector.

The detector spectral response $\mathcal{D}(\lambda)$ was determined from the manufacturer specifications for each detector and divided out from the measured intensity. In the range from $\lambda = 800$ nm to $\lambda = 1600$ nm, we used a Hamamatsu R12829 in a Hamamatsu C12597-01 socket for power supply. The photomultiplier output was directly connected to the current input of a Zurich Instruments MFLI digitizing lockin which yielded a current value proportional to the incident SHG power. The range from $\lambda = 1560$ nm to $\lambda = 1700$ nm was measured with an unbiased Thorlabs FDS010 photodiode, while measurements from $\lambda = 1600$ nm to $\lambda = 2600$ nm performed with a Thorlabs FGA01 InGaAs photodiode. For their respective wavelength ranges, the photodiodes were connected to the input of a Cremat CR-Z-110 charge preamplifier followed by a CR-S-8μs shaping amplifier whose output were connected to the voltage input of the lockin. The overlap between the wavelength ranges of the three detectors allowed us to account for the current to voltage conversion of the charge integrating electronics in relation to the PMT, permitting the Si photodiode to be a bridge between the PMT and the InGaAs detector.

In all cases, we used a set of two shortpass filters and a single bandpass filter centered at the nominal second harmonic wavelength, which determined the spectral transmittance $\mathcal{T}(\lambda)$. These filters were chosen in order to provide attenuation of >ND10 at the fundamental wavelength while not attenuating the SHG by more than a factor of 2. We verified the transmittance of the filters to match the supplier's specifications through individual UV-Vis measurements, as shown in figure 7. It was necessary to account for the fact that the bandwidth of the filter was more narrow than that of the second harmonic pulse, in particular at the longest wavelengths measured. Thus, after dividing by $\mathcal{D}(\lambda)$, the true incident SHG power was determined by dividing out the numerically integrated product of the measured filter data and a normalized Gaussian representing the fitted parameters of the measured second harmonic spectrum the FROG traces as

$$\frac{1}{\sqrt{\pi}\Delta\lambda} \int_{-\infty}^{\infty} \mathcal{T}(\lambda) \exp\left[-((\lambda-\lambda_0)/\Delta\lambda)^2\right] d\lambda. \quad \text{(A12)}$$

4. Translating reflection geometry parameters to bulk nonlinear conductivity

To calculate the bulk nonlinear optical parameters $\chi^{\rm shg}$, and $\sigma^{\rm shg}$, we also need to correct for the linear optical parameters of TaAs. We use the formula derived by Bloembergen and Pershan[23, 46] to account for the linear optical properties such as the refractive index and transmission coefficients:

$$\chi_R^{\rm shg} = \frac{\chi^{\rm shg}}{\left(\sqrt{\epsilon(2\omega)}+\sqrt{\epsilon(\omega)}\right)\left(\sqrt{\epsilon(2\omega)}+1\right)}T(\omega)^2 \quad \text{(A13)}$$

where $\chi_R^{\rm shg}$ denotes the nonlinear susceptibility in reflection geometry, as measured in the experiment, $\epsilon(\omega)$, and $T(\omega)$, are respectively the dielectric constant and the transmission coefficients of TaAs.

We define the Bloembergen-Pershan correction coefficient

$$BP(\omega) \equiv \frac{T(\omega)^2}{\left(\sqrt{\epsilon(2\omega)}+\sqrt{\epsilon(\omega)}\right)\left(\sqrt{\epsilon(2\omega)}+1\right)}, \quad \text{(A14)}$$

which can be calculated as a function of incident photon energy based on values of the dielectric constant as measured by ellipsometry.[47] (See figure 7.) Optical conductivity is more conveniently modeled than susceptibility, and can be obtained using the relation: $\sigma_{ijk}^{\rm shg}(\omega) = -2i\epsilon_0\,\omega\,\chi_{ijk}^{\rm shg}(\omega)$

Appendix B: Quasi-One Dimensional Model for Second Harmonic Generation

1. Model details

This Appendix describes the details of the phenomenological model of Rice-Mele chains used in the main text, and details the calculation of its second harmonic response. The model is inspired by the polar character of the TaAs lattice as well as the observation that σ_{zzz} is the dominant component of the SHG response. This is reminiscent of a 1D system, in which σ_{zzz} is the only non-zero component. We consider a quasi-one dimensional model composed of 1D Rice Mele chains in the z-direction [34] with only weak couplings in the xy plane. In momentum space it takes the form of a two band model $H_{\boldsymbol{k}} = \boldsymbol{d_k}\cdot\boldsymbol{\sigma}$ with

$$\begin{aligned} d_x &= t\sin(ck_z/2) \\ d_y &= \delta\cos(ck_z/2) \\ d_z &= \Delta + t_{AB}(\cos(bk_x)+\cos(bk_y)). \end{aligned} \quad \text{(B1)}$$

Here b and c are the dimension of the unit cell in the x and z directions. Along lines in the z-direction there is a staggered onsite potential $\pm\Delta$ and staggered hopping $t\pm\delta$. The parameter $t_{AB} = t_{\|,A} - t_{\|,B} \ll \Delta$ represents the difference between the inter-chain hopping strengths. When $t_{AB}=0$, this reduces to an ensemble of independent Rice-Mele models. The overall energy scale is set by fixing Δ; there are three independent parameters of the model: $\tilde{t} = t/\Delta$, $\tilde{\delta} = \delta/\Delta$, and $\tilde{t}_{AB} = t_{AB}/\Delta$.

2. Calculation of SHG Response

The second-order conductivity tensor is defined via

$$J_a(\omega_0) = \sum_{b,c}\sigma_{abc}(\omega_0;\omega_1,\omega_1)E_b(\omega_1)E_c(\omega_2) \quad \text{(B2)}$$

where $\omega_0 = \omega_1 + \omega_2$ is the frequency of the emitted photon The second harmonic response is $\sigma_{abc}^{\rm shg}(\omega) = \sigma_{abc}(2\omega;\omega,\omega)$. We will also frequently invoke the shift current, $\sigma_{abc}^{\rm shift}(\omega) = \sigma_{abc}(0;-\omega,\omega)$. The shift current may be thought of as the "solar panel response": the amount of DC current generated under illumination.

In two band models, the diagonal components of σ_{abc} are particularly simple. For $\omega>0$,[20, 29]

$$\mathrm{Re}\,\sigma_{aaa}^{\rm shift}(\omega) = \frac{2\pi e^3}{\hbar^2\omega^2}\int[d\boldsymbol{k}]f_{10}v_{01}^a w_{10}^{aa}\delta(\omega_{10}-\omega) \quad \text{(B3)}$$

$$\mathrm{Re}\,\sigma_{aaa}^{\rm shg}(\omega) = \frac{\pi e^3}{2\hbar^2\omega^2}\int[d\boldsymbol{k}]\Big[f_{10}v_{01}^a w_{10}^{aa}\delta(\omega_{10}-2\omega) - 2f_{10}v_{01}^a w_{10}^{aa}\delta(\omega_{10}-\omega)\Big] \quad \text{(B4)}$$

where $\int[d\boldsymbol{k}] = \int\frac{d^d\boldsymbol{k}}{(2\pi)^d}$ is the normalized integral over the Brillouin zone, 0 and 1 refer to the valence and conduction bands respectively, f_{01} is the difference in Fermi factors, ω_{01} is the difference in band frequencies, $v^a = \partial_{k_a}H$ is the velocity operator, and $w^{aa} = \partial_{k_a}\partial_{k_a}H$. Taken together, these lead to the convenient identity (see equation (3) of the main text)

$$\mathrm{Re}\,\sigma_{aaa}^{\rm shg}(\omega) = -\frac{1}{2}\mathrm{Re}\,\sigma_{aaa}^{\rm shift}(\omega) + \mathrm{Re}\,\sigma_{aaa}^{\rm shift}(2\omega). \quad \text{(B5)}$$

Again, these equations only hold for two band models.

We now evaluate these equations for our specific model. We focus on the two-photon term of the SHG. Differentiating equation (B1),

$$\hat{v}^z = \frac{c}{\hbar 2}\left[-t\sin\left(\frac{k_z c}{2}\right)\sigma^x + \delta\cos\left(\frac{k_z c}{2}\right)\sigma^y\right] \tag{B6}$$

and

$$\hat{w}^{zz} = \frac{c^2}{4\hbar}\left[-t\cos\left(\frac{k_z c}{2}\right)\sigma^x - \delta\sin\left(\frac{k_z c}{2}\right)\sigma^y\right]. \tag{B7}$$

Hence, using $\sigma^i\sigma^j = i\varepsilon^{ij}_k\sigma^k$,

$$\hat{v}^z\hat{w}^{zz} = \frac{1}{\hbar^2}\left(\frac{c}{2}\right)^3\left[\frac{t^2-\delta^2}{2}\sin\left(\frac{2k_z c}{2}\right) + it\delta\right]\sigma^z. \tag{B8}$$

We must now evaluate the matrix element $\langle 0|\sigma^z|0\rangle$. Adopting the Bloch sphere representation,

$$\boldsymbol{d} = d\left(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta\right) \tag{B9}$$

which leads to the eigenvector equation $H|0\rangle = \varepsilon_0|0\rangle$ with

$$|0\rangle = \begin{pmatrix} \cos\dfrac{\theta}{2} \\ \sin\dfrac{\theta}{2}e^{i\varphi} \end{pmatrix}, \tag{B10}$$

and eigenvalue $\varepsilon_v = -d$. Therefore

$$\langle 0|\sigma^z|0\rangle = \left(\cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2}\right) = \cos\theta = \frac{d_z}{d}. \tag{B11}$$

Combining (B8) and (B11) shows that the two-photon contribution is

$$\mathrm{Re}\,\sigma^{\mathrm{shg2p}}_{zzz}(\omega) = \frac{\pi e^e}{2\hbar^2\omega^2}\int[d\boldsymbol{k}]f_{10}v^z_{01}w^{zz}_{10}\delta(\omega_{01}-2\omega) = i\frac{\pi e^3}{2\hbar^2\omega^2}\int[d\boldsymbol{k}](-1)\frac{1}{\hbar^2}\left(\frac{c}{2}\right)^3 it\delta\frac{d_z}{d}\delta(\omega_{10}-2\omega). \tag{B12}$$

Note that the $\sin(k_z c)$ term integrates to zero and can be ignored. What remains is to de-dimensionalize the integral and simplify the expression somewhat. Define $\varepsilon(k) = d(k) = \hbar\omega_{10}/2$. Then

$$\mathrm{Re}\,\sigma^{\mathrm{shg2p}}_{zzz}(\omega) = \frac{\pi e^3}{2\hbar^2}\left(\frac{c}{2}\right)^3 t\delta\int\frac{d^3\boldsymbol{k}}{(2\pi)^3}\frac{d_z}{\varepsilon(k)\omega^2}\frac{\hbar}{2}\delta(\varepsilon(k)-\hbar\omega). \tag{B13}$$

To account for physical broadening of the peak, we use a Lorentzian $\delta(\varepsilon-\omega) \to \dfrac{1}{\pi}\dfrac{\gamma}{(\varepsilon-\omega)^2+\gamma^2}$. Furtheremore, we convert the remaining factors of ω into ε to remove an unphysical diverges at $\omega \to 0$ in numerical evaluation. So

$$\mathrm{Re}\,\sigma^{\mathrm{shg2p}}_{zzz}(\omega) = \frac{\pi e^3}{\hbar}\left(\frac{c}{2}\right)^3 t\delta\frac{1}{4\pi}\int[d\boldsymbol{k}]\frac{d_z}{\varepsilon(k)^3}\frac{\gamma}{(\varepsilon(k)-\omega)^2+\gamma^2}. \tag{B14}$$

Now we de-dimensionalize by normalizing energies to Δ, and henceforth represent Δ-normalizes quantities with a tilde. For example, $\tilde{t} \equiv t/\Delta$. Furthermore, let us define $X \equiv k_x b, Y \equiv k_y b, Z \equiv k_z c$. This gives

$$\widetilde{\varepsilon} = \sqrt{\tilde{\delta}^2\cos(Z/2)^2 + \tilde{t}^2\sin(Z/2)^2 + \widetilde{d_z}^2} \text{ and } \widetilde{d_z} = 1 + \tilde{t}_{AB}\left[\cos(X)+\cos(Y)\right]. \tag{B15}$$

Thus

$$\sigma^{\mathrm{shg2p}}_{zzz}(\omega) = \frac{e^3}{\hbar}\left(\frac{c}{2}\right)^3\frac{\tilde{t}\tilde{\delta}}{\Delta}\frac{1}{4}\frac{1}{(2\pi)^3 abc}\int_{-\pi}^{\pi}dX\,dY\,dZ\,\frac{\widetilde{d_z}}{\tilde{\varepsilon}(k)^3}\frac{\tilde{\gamma}}{(\tilde{\omega}-\tilde{\varepsilon}(k))^2+\tilde{\gamma}^2}. \tag{B16}$$

To match with experimental data, one should multiply by degeneracy factors that account for the spin degeneracy $g_s = 2$ and orbital degeneracy $g_O = 4$ for the number of Rice-Mele chains per unit cell. Rearranging, we arrive at the two-photon contribution to the SHG

$$\sigma^{\mathrm{shg2p}}_{zzz}(\omega) = \left[\frac{e^2}{\hbar}\frac{e}{\Delta}\right]\frac{c^2}{b^2}\left(\frac{1}{4\pi}\right)^3 F(\tilde{\omega};\tilde{\delta},\tilde{t},\tilde{t}_{AB},\tilde{\gamma}) \tag{B17}$$

where the term in brackets has units of conductance per Volt, and the integral has been re-packaged as

$$F(\tilde{\omega};\tilde{\delta},\tilde{t},\tilde{t}_{AB},\tilde{\gamma}) = g_s g_O\frac{\tilde{\delta}\tilde{t}}{4}\int_{-\pi}^{\pi}dX\int_{-\pi}^{\pi}dY\int_{-\pi}^{\pi}dZ\,\frac{1+\tilde{t}_{AB}\left[\cos(X)+\cos(Y)\right]}{\left(\tilde{\delta}^2\cos(Z/2)^2+\tilde{t}^2\sin(Z/2)^2+\left(1+\tilde{t}_{AB}\left[\cos(X)+\cos(Y)\right]\right)^2\right)^{3/2}}$$
$$\times\frac{\tilde{\gamma}}{\left(\tilde{\omega}-\left[\tilde{\delta}^2\cos(Z/2)^2+\tilde{t}^2\sin(Z/2)^2+\left(1+\tilde{t}_{AB}\left[\cos(X)+\cos(Y)\right]\right)^2\right]^{1/2}\right)^2+\tilde{\gamma}^2}. \tag{B18}$$

By Equation (B5), $\operatorname{Re}\sigma^{\rm shift}_{zzz}(\omega) = -2\sigma^{\rm shg2p}_{zzz}(\omega)$ (see equation (2) of the main text).

It is worth examining the special case where $t_{AB} = 0$, i.e. an ensemble of uncoupled Rice-Mele chains. Here the integral becomes analytically tractable. Converting the Lorentzian back to a δ function,

$$\int_0^\infty d\omega\ F\left(\widetilde{\omega};\tilde{\delta},\tilde{t},\tilde{t}_{AB}\to 0,\tilde{\gamma}\to 0\right) \tag{B19}$$

$$= \frac{4\pi^3\Delta g_s g_O\tilde{\delta}\tilde{t}}{\hbar}\int_{-\pi}^{\pi}\frac{dZ}{\left(\tilde{\delta}^2\cos(Z/2)^2+\tilde{t}^2\sin(Z/2)^2+1\right)^{3/2}} \tag{B20}$$

$$= g_s g_0\frac{\Delta}{\hbar}\frac{2(2\pi)^3\tilde{\delta}\tilde{t}}{4}\frac{E_2\left(\frac{\tilde{\delta}^2-\tilde{t}^2}{1+\tilde{\delta}^2}\right)}{\left(1+\tilde{t}^2\right)\sqrt{1+\tilde{\delta}^2}}, \tag{B21}$$

where E_2 is the complete elliptic integral of the second kind. Therefore, using the fact that the spectral weights of the shift current and two-photon contribution to the SHG are equal,

$$\Sigma^{\rm shift}_z = 2\Sigma^{\rm shg\ 2p}_z = \left[\frac{e^3}{\hbar^2}\right]\frac{c^2}{b^2}G(\tilde{\delta},\tilde{t}) \tag{B22}$$

where

$$G(\tilde{\delta},\tilde{t}) \equiv g_s g_O\frac{\tilde{\delta}\tilde{t}}{8}\frac{E_2\left(\frac{\tilde{\delta}^2-\tilde{t}^2}{1+\tilde{\delta}^2}\right)}{\left(1+\tilde{t}^2\right)\sqrt{1+\tilde{\delta}^2}}. \tag{B23}$$

The function G is bounded and attains its global maximum at $G(\sqrt{2},\sqrt{2}) = \pi 3^{-3/2} \approx 0.604$ (see equation (4) of the main text). This shows that the total spectral weight in the two-photon contribution is bounded, i.e. there is a maximum amount of SHG for Rice-Mele models.

Appendix C: Gauge-Invariant Cumulants and a new sum rule

This appendix reviews the theoretical machinery of gauge-invariant cumulants (GICs) that underlies the relationship between ground-state polarization distributions and the frequency-integrated nonlinear response described in the last part of the main text. We restrict the analysis to two-band tight-binding models, relevant for this work, to show that this connection can be made exact. It generalizes the known relationships between the Berry connection and polarization, as well as the relation between the spread of Wannier functions with linear response. The last part of the Appendices shows how this connection gives a guide to construct a model whose spectral weight exceeds the maximum spectral weight possible in the Rice-Mele model [34].

1. The polarization distribution

Let us start with the macroscopic polarization of a solid [31, 32, 35–38] and Kohn's theory of the insulating state [48]. Consider a solid with N electrons and M nuclei in any dimension d. The macroscopic polarization operator is

$$\widehat{\boldsymbol{P}} = \frac{1}{V}\left(e\widehat{\boldsymbol{X}} + q_{\rm nuc}\widehat{\boldsymbol{X}}_{\rm nuc}\right), \tag{C1}$$

where $\widehat{\boldsymbol{X}} = \sum_{i=1}^{N}\hat{\boldsymbol{x}}_i$ is the center-of-mass position of the electrons (resp. $\widehat{X}_{\rm nuc}$ of the nuclei). Within the clamped nuclei approximation, the nuclei do not move and we can set $\widehat{X}_{\rm nuc} = 0$ by a choice of coordinates. The macroscopic polarization is the expectation $\langle\widehat{\boldsymbol{P}}\rangle$.

The GICs are a systematic way to extract further information from $\widehat{\boldsymbol{P}}$ that we now exploit. Computing $\langle\widehat{\boldsymbol{P}}\rangle = \langle\Psi|\widehat{\boldsymbol{P}}|\Psi\rangle$, averages over many different centers of charge. Following Souza *et al* [31] we define the *distribution* of those centers of charge, i.e. the spatial distribution of polarization, via

$$p(\boldsymbol{X}) = \left\langle\Psi\left|\delta(\widehat{\boldsymbol{X}}-\boldsymbol{X})\right|\Psi\right\rangle. \tag{C2}$$

This should be interpreted as the probability that the center of charge is exactly at the position $\boldsymbol{X}$. Different electronic properties of the solid have already been directly linked to $p(\boldsymbol{X})$, including the Berry connection, localization length of Wannier functions, and the f-sum rule[31].We can now add the second-harmonic to that list. As a cautionary note, the polarization is distinct from other real-space distributions in solids; $p(\boldsymbol{X})$ is a different object from the electron density, or the Wannier functions. The subtle relationship between these is discussed carefully by Souza *et al.* [31].

The great utility of this distribution is tempered by the indirect way it must be computed, a complication due to

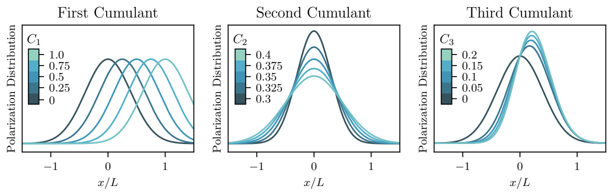

FIG. 8. A schematic of the first three cumulants of the polarization distribution. In each panel, one cumulant is varied while the others are held fixed. Note that the true polarization distribution is periodic in L.

the fact that the position operator $\widehat{\boldsymbol{X}}$ is ill-defined [49]. This implies that, even knowing the exact eigenvectors of the system, $p(X)$ cannot be straightforwardly computed and, moreover, it is not immediately clear if $p(X)$ is even a well-defined distribution. A key result of Souza *et al* is that $p(X)$ can be carefully defined, rendering its cumulants computable.

2. Definition of gauge-invariant cumulants

Recall that the first several cumulants C_n of a distribution $p(X)$ can be written in terms of the moments as

$$C_1^i = \langle X^i \rangle \tag{C3}$$

$$C_2^{ij} = \langle X^i X^j \rangle - \langle X^i \rangle \langle X^j \rangle \tag{C4}$$

$$\begin{aligned} C_3^{ijk} &= \langle X^i X^j X^k \rangle + 2\langle X^i \rangle \langle X^j \rangle \langle X^k \rangle \\ &\quad - \langle X^i X^j \rangle \langle X^k \rangle - \langle X^i X^k \rangle \langle X^j \rangle - \langle X^j X^k \rangle \langle X^i \rangle \end{aligned} \tag{C5}$$

where $i, j, k \in \{x, y, z\}$ are spatial indices. The first cumulant is the same as the first moment, the second is the variance, and the third cumulant is often called the skew of the distribution. A schematic of the first three cumulants is shown in figure (8).

A convenient way to compute the moments is in terms of the characteristic function

$$\mathcal{C}(\boldsymbol{\alpha}) = \langle e^{-i\boldsymbol{\alpha}\cdot\boldsymbol{X}} \rangle, \tag{C6}$$

so that $\langle X^i X^j \cdots X^k \rangle = i\partial_{\alpha_i} i\partial_{\alpha_j} \cdots i\partial_{\alpha_k} C(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=0}$. The cumulants are likewise obtained by differentiating the log of the characteristic function

$$C_n^{ij\ldots k} = i\partial_{\alpha_i} i\partial_{\alpha_j} \cdots i\partial_{\alpha_k} \ln \mathcal{C}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=0}. \tag{C7}$$

Following [31], we define

$$\ln \mathcal{C}(\boldsymbol{\alpha}) = \frac{V}{(2\pi)^d} \int [d\boldsymbol{k}] \; \ln \langle \Psi_{\boldsymbol{k}} | e^{-i\boldsymbol{\alpha}\cdot\widehat{\boldsymbol{X}}} | \Psi_{\boldsymbol{k}+\boldsymbol{\alpha}} \rangle \tag{C8}$$

where V is the volume of the system, $\int [d\boldsymbol{k}]$ denotes the normalized integral over the Brillouin zone, and $|\Psi_k\rangle$ is the many-body wavefunction with boundary conditions twisted by e^{ikL}. In a single-particle description, $|\Psi_k\rangle = \prod_n u_{\boldsymbol{k}n} c^\dagger_{\boldsymbol{k}n} |0\rangle$ where the Bloch wavefunctions $\psi_{\boldsymbol{k}n}(\boldsymbol{r}) = u_{\boldsymbol{k}n}(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ are chosen to satisfy a smooth, periodic gauge: $\psi_{\boldsymbol{k}n} \equiv \psi_{\boldsymbol{k}+\boldsymbol{G},n}$ for reciprocal lattice vectors $\boldsymbol{G}$. The cumulants of the polarization distribution can be computed via differentiating (C8). The first three are

$$C_1^i = iV \int [d\boldsymbol{k}] \; \mathrm{Tr}\left[c_1^i\right] \tag{C9}$$

$$C_2^{ij} = i^2 V \int [d\boldsymbol{k}] \; \mathrm{Tr}\left[c_2^{ij} - c_1^i c_1^j\right] \tag{C10}$$

$$C_3^{ijk} = i^3 V \int [d\boldsymbol{k}] \; \mathrm{Tr}\left[c_3^{ijk} - c_2^{ij} c_1^k - c_2^{jk} c_1^i - c_2^{ki} c_1^j + 2 c_1^i c_1^j c_1^k\right] \tag{C11}$$

where the trace is over occupied bands, and the c's are matrices in the band space whose matrix elements are

$$c_1^i = i\langle u_{kn} | \partial_{k_i} u_{km} \rangle, \tag{C12}$$

$$c_2^{ij} = i^2 \langle u_{kn} | \partial_{k_i} \partial_{k_j} u_{km} \rangle, \tag{C13}$$

$$c_3^{ijk} = i^3 \langle u_{kn} | \partial_{k_i} \partial_{k_j} \partial_{k_j} u_{km} \rangle. \tag{C14}$$

One can show that the cumulants are gauge-invariant— hence the name — and purely real in the presence of time-reversal symmetry.

To compute the cumulants numerically, it is advantageous to use an alternative formulation (written here for a single occupied band in 1D)[30]

$$\Delta k \prod_{i=0}^{M-1} \ln \langle u_{k_i} | u_{k_{i+1}} \rangle = \sum_{n=1} \frac{(i\Delta k)^n}{n!} C_n \tag{C15}$$

where $k_{i+1} - k_i = \Delta k = 2\pi/M$, and again using a periodic gauge for the Bloch wavefunctions ψ_k. The cumulants can be measured by computing the left-hand side on successively finer meshes of k-points and fitting the result to a power series in Δk, the mesh size. The first several gauge-invariant cumulants in the Rice-Mele model have been computed recently, providing intuitive visualizations of the polarization distribution [30].

3. Interpretation and physical intuition

Each of the GICs has a dual interpretation: as a measure of the spatial properties of the electrons, or as an electromagnetic response. The first cumulant tells us the mean of the center of charge polarization, i.e. the total deviation of electrons from their companion nuclei [37]. Via equation (C1), the polarization of the solid has a straightforward formula in terms of the first cumulant

$$\langle \widehat{\boldsymbol{P}} \rangle = \frac{e}{V}\boldsymbol{C}_1 = e\sum_{n \text{ occ}} \int [d\boldsymbol{k}] \boldsymbol{A}_{\boldsymbol{k}n}, \tag{C16}$$

where $\boldsymbol{A}_{\boldsymbol{k}n}$ is the Berry connection. The last expression is familiar from the modern theory of polarization. Equation (C16) shows that, up to dimensionful prefactors, the first cumulant is the electrical response of the system in the absence of an external field, a purely static quantity.

The second cumulant measures the covariance of the polarization distribution. In an insulator, electrons are exponentially localized. Their localization is in the i ($= x, y, z$) direction is given by (no summation)[31]

$$\xi_i = \lim_{N\to\infty} \sqrt{\frac{1}{N} C_2^{ii}}. \tag{C17}$$

The second cumulant then gives a great deal of information about the character of the material. In a metal-insulator transition, for instance, the localization length diverges, and hence so does the second cumulant. Physically, the electrons in a metal are almost entirely delocalized, so the polarization distribution should be nearly flat, with variance on the order of the size of the system.

It is useful to contrast the information in C_2 with the Wannier functions. In 1d, the maximally localized Wannier functions have the property that the their centers are governed by the centers of polarization, and their variance is proportional to the squared localization length. We stress, however, that this does *not* imply the Wannier density is the same as the polarization distribution. In dimensions greater than one, there are no unique maximally-localized single-particle Wannier functions, and all Wannier functions have variance strictly larger than the squared localization length.

As an optical response, the second cumulant encodes the all-frequency linear response of the system. A standard application of the flucutation-dissipation theorem shows that the fluctuations in the polarization distribution (C_2) is related to the total current response of the system:[31]

$$\frac{\pi e^2}{V^2\hbar} C_2^{\alpha\beta} = \frac{d\omega}{\omega} \operatorname{Re} \sigma^{\alpha\beta}(\omega) \tag{C18}$$

where $\sigma^{\alpha\beta}$ is the linear conductivity.

4. A New Sum Rule

Now that we have seen the dual nature — spatial and optical — of the first two GICs, it is perhaps not too surprising that the third cumulant can give a non-linear sum rule. The third cumulant measures the skewness of the polarization distribution which, in one dimension, says if the left or right "shoulder" of the distribution is larger.

On the optical side, we find that the second cumulant is related to the second-harmonic response of the system. Given that the first cumulant determines the polarization and the second cumulant gives a linear sum rule, it is perhaps not too surprising that the third cumulant can give a non-linear sum rule. More precisely, we show that for a generic two-band model (see equation (5) of the main text)

$$\Sigma_a^{\text{shift}} = \frac{2\pi e^3}{V\hbar^2} C_3^a \tag{C19}$$

where

$$\Sigma_a^{\text{shift}} = \int_0^\infty d\omega \; \operatorname{Re} \sigma^{aaa}(0; -\omega, \omega) \tag{C20}$$

and, specializing to the case of two bands,

$$C_3^a = -V \int \frac{d^d\boldsymbol{k}}{(2\pi)^d} \operatorname{Im}\left[c_3 - 3c_1c_2 + 2c_1^3\right] \tag{C21}$$

where $c_n = \langle 0| \left(i\partial_{k_a}\right)^n |0\rangle$ for the valence band Bloch wavefunction $|0\rangle = |u_0(\boldsymbol{k})\rangle$.

Integrating (B4), the spectral weight of the shift current in a two-band model is

$$\Sigma_a^{\text{shift}} = \frac{2\pi e^3}{\hbar^2} \int [d\boldsymbol{k}] \frac{\operatorname{Im}[\langle 0|\partial_k H|1\rangle\langle 0|\partial_k^2 H|0\rangle]}{(\epsilon_1 - \epsilon_0)^2}, \tag{C22}$$

where $|0\rangle$ and $|1\rangle$ denote valence and conduction bands, respectively, and $\varepsilon_n = \varepsilon_n(k)$ are the band energies.

The Schrödinger equation and its k derivatives give

$$\begin{aligned}
&H|n\rangle = \epsilon_n|n\rangle, && \text{(C23)}\\
&\partial_k H|n\rangle + H|\partial_k n\rangle = \partial_k\epsilon_n|n\rangle + \epsilon_n|\partial_k n\rangle, && \text{(C24)}\\
&\partial_k^2 H|n\rangle + \partial_k H|\partial_k n\rangle + H|\partial_k^2 n\rangle && \text{(C25)}\\
&\quad = \partial_k^2\epsilon_n|n\rangle + \partial_k\epsilon_n|\partial_k n\rangle + \epsilon_n|\partial_k^2 n\rangle.
\end{aligned}$$

Taking inner products and applying $|0\rangle\langle 0| + |1\rangle\langle 1| = I$ implies

$$\langle 0|\partial_k H|1\rangle = (\epsilon_1 - \epsilon_0)\langle 0|\partial_k 1\rangle, \tag{C26}$$

$$\begin{aligned}\langle 1|\partial_k^2 H|0\rangle = {}& (\epsilon_0 - \epsilon_1)[\langle 1|\partial_k^2 0\rangle - 2\langle 1|\partial_k 0\rangle\langle 0|\partial_k 0\rangle] \\ &+ 2(\partial_k\epsilon_0 - \partial_k\epsilon_1)\langle 1|\partial_k 0\rangle.\end{aligned} \tag{C27}$$

Substituting these into the integral in the spectral weight yields

$$\begin{aligned}
&\frac{1}{(\epsilon_1-\epsilon_0)^2}\,\mathrm{Im}[\langle 0|\partial_k H|1\rangle\langle 0|\partial_k^2 H|0\rangle] \\
&= \mathrm{Im}\left\{\langle 0|\partial_k 1\rangle\left[-\langle 1|\partial_k^2 0\rangle + 2\langle 1|\partial_k 0\rangle\langle 0|\partial_k 0\rangle\right]\right\} \\
&\quad - \frac{\partial_k\epsilon_0-\partial_k\epsilon_1}{\epsilon_0-\epsilon_1}\,\mathrm{Im}[\langle 0|\partial_k 1\rangle\langle 1|\partial_k 0\rangle].
\end{aligned} \tag{C28}$$

We can drop the last term since $\langle 0|\partial_k 1\rangle\langle 1|\partial_k 0\rangle$ is real. Applying the resolution of the identity, the first term is

$$\begin{aligned}
\langle 0|\partial_k 1\rangle\langle 1|\partial_k^2 0\rangle &= -\langle\partial_k 0|\partial_k^2 0\rangle + \langle\partial_k 0|0\rangle\langle 0|\partial_k^2 0\rangle \\
&= -\partial_k(\langle 0|\partial_k^2 0\rangle) + \langle 0|\partial_k^3 0\rangle - \langle 0|\partial_k 0\rangle\langle 0|\partial_k^2 0\rangle \\
&= \partial_k c_2 + ic_3 - ic_1c_2,
\end{aligned} \tag{C29}$$

while the second term becomes

$$\begin{aligned}
&\langle 0|\partial_k 1\rangle\langle 1|\partial_k 0\rangle\langle 0|\partial_k 0\rangle \\
&= -\langle\partial_k 0|\partial_k 0\rangle\langle 0|\partial_k 0\rangle + \langle\partial_k 0|0\rangle\langle 0|\partial_k 0\rangle\langle 0|\partial_k 0\rangle \\
&= -\partial_k(\langle 0|\partial_k 0\rangle)\langle 0|\partial_k 0\rangle + \langle 0|\partial_k^2 0\rangle\langle 0|\partial_k 0\rangle - c_1^3 \\
&= +\frac{1}{2}\partial_k(c_1^2) + ic_1c_2 - ic_1^3.
\end{aligned}$$

The total derivatives vanish after integration, and we obtain equation (C19).

This sum rule leads to intriguing conclusions. For two-band models, non-linear optical responses can be understood as a facet of the spatial distribution of polarization. This provides a clear physical picture that may be more intuitive than the normal expressions for SHG, which involve k-space sums over virtual transitions. Moreover, the sum rule can *predict* the shift current — and hence the SHG response — of a material as a ground state property.

5. Upper bounds in the Rice-Mele model and beyond

In light of this relation between the SHG and C_3, it is worth revisiting the above fact, equation (B23), that there is a maximum SHG in Rice-Mele models. While this maximum is a good benchmark as to how much second-harmonic can be produced by a system, it is not an absolute bound. By designing a Hamiltonian with a large C_3, we will see this bound may be exceeded.

For concreteness, consider a generalization of the Rice-Mele model with a next-nearest neighbor hopping $H = H_{\mathrm{RM}} + H_{\mathrm{NNN}}$ where

$$\begin{aligned}
H_{\mathrm{RM}} &= \sum_n \Delta(-1)^n c_n^\dagger c_n \\
&\quad + \left(\frac{t}{2} + (-1)^n\frac{\delta}{2}\right)c_n^\dagger c_{n+1} + \mathrm{h.c.}
\end{aligned} \tag{C30}$$

$$\begin{aligned}
H_{\mathrm{NNN}} &= \sum_n \left(\frac{t'}{2} + \frac{\delta'}{2}\right)c_{An}^\dagger c_{A,n+1} \\
&\quad + \left(\frac{t'}{2} - \frac{\delta'}{2}\right)c_{Bn}^\dagger c_{B,n+1} + \mathrm{h.c.}
\end{aligned} \tag{C31}$$

In k-space this becomes

$$H = \sum_k \boldsymbol{c}_k^\dagger\left[h_{\mathrm{RM}} + h_{\mathrm{NNN}}\right]\boldsymbol{c}_k \tag{C32}$$

where

$$h_{\mathrm{RM}}(\boldsymbol{k}) = t\cos\left(\frac{kc}{2}\right)\sigma_x + \delta\sin\left(\frac{kc}{2}\right)\sigma_y + \Delta\sigma_z \tag{C33}$$

$$h_{\mathrm{NNN}}(\boldsymbol{k}) = t'\cos(kc)\,\mathrm{Id}_2 + \delta'\cos(ka)\sigma_z. \tag{C34}$$

By applying the maximum entropy method to approximately solve the moment problem [30], we can reconstruct the polarization distribution. Figure C 4 shows the reconstructed polarization distributions in the next-nearest neighbor model with parameters $\Delta = 1$, $\delta = \sqrt{2}$, $t = \sqrt{2} - 2\gamma$, $\delta' = \gamma$ and $t' = 0$. As γ is tuned past zero, C_3 increases, and exceeds the maximum possible in the Rice-Mele model. We can thus tune the model to achieve an arbitrary large C_3 and, at a metal-insulator transition, cause it to diverge. Finding materials realizing a large third cumulant would be excellent candidates for giant shift current or second harmonic generation.

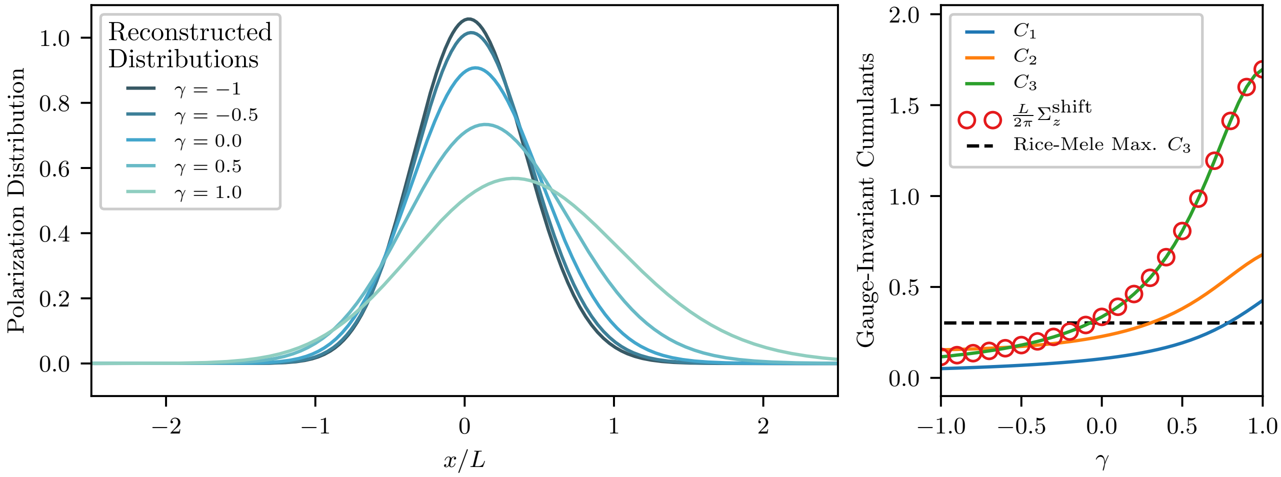

FIG. 9. (Left) Reconstructed polarization distributions in the extended Rice-Mele model via the maximum entropy method[30]. (Right) The gauge-invariant cumulants and spectral weight in the next-nearest neighbor extension of the Rice-Mele model. The maximum in the normal Rice-Mele model, equation (B23), is the black line, and the spectral weight as a function of γ is computed via equation C22. All parameters are given in the text.

[1] Curie, P. Sur la symétrie dans les phénomènes physiques, symétrie d'un champ électrique et d'un champ magnétique. *J. Phys. Theor. Appl.* **3**, 393–415 (1894).

[2] Xiao, D., Chang, M.-C. & Niu, Q. Berry phase effects on electronic properties. *Rev. Mod. Phys.* **82**, 1959–2007 (2010).

[3] Murakami, S. Phase transition between the quantum spin Hall and insulator phases in 3D: emergence of a topological gapless phase. *New J. Phys.* **9**, 356 (2007).

[4] Wan, X., Turner, A. M., Vishwanath, A. & Savrasov, S. Y. Topological semimetal and Fermi-arc surface states in the electronic structure of pyrochlore iridates. *Physical Review B* **83**, 205101 (2011).

[5] Lv, B. Q. *et al.* Experimental discovery of Weyl semimetal TaAs. *Physical Review X* **5**, 031013 (2015).

[6] Yang, L. *et al.* Weyl semimetal phase in the non-centrosymmetric compound TaAs. *Nature physics* **11**, 728 (2015).

[7] Xu, S.-Y. *et al.* Discovery of a Weyl fermion semimetal and topological Fermi arcs. *Science* **349**, 613–617 (2015).

[8] Zhong, S., Moore, J. E. & Souza, I. Gyrotropic magnetic effect and the magnetic moment on the Fermi surface. *Physical Review Letters* **116**, 077201 (2016).

[9] Ma, J. & Pesin, D. A. Chiral magnetic effect and natural optical activity in metals with or without Weyl points. *Physical Review B* **92**, 235205 (2015).

[10] Liang, T. *et al.* Ultrahigh mobility and giant magnetoresistance in the dirac semimetal Cd_3As_2. *Nature Materials* **14**, 280–284 (2014).

[11] Xiong, J. *et al.* Evidence for the chiral anomaly in the Dirac semimetal Na_3Bi. *Science* **350**, 413–416 (2015).

[12] Gooth, J. *et al.* Experimental signatures of the mixed axial–gravitational anomaly in the Weyl semimetal NbP. *Nature* **547**, 324 (2017).

[13] Belinicher, V. & Sturman, B. The photogalvanic effect in media lacking a center of symmetry. *Usp. Fiz. Nauk* **130**, 415 (1980).

[14] Belinicher, V., Ivchenko, E. & Sturman, B. Kinetic theory of the displacement photovoltaic effect in piezoelectrics. *Zh. Eksp. Teor. Fiz* **83**, 649–661 (1982).

[15] Golub, L., Ivchenko, E. L. & Spivak, B. Photocurrent in gyrotropic weyl semimetals. *JETP Letters* **105**, 782–785 (2017).

[16] von Baltz, R. & Kraut, W. Theory of the bulk photovoltaic effect in pure crystals. *Phys. Rev. B* **23**, 5590–5596 (1981).

[17] Sipe, J. & Shkrebtii, A. Second-order optical response in semiconductors. *Physical Review B* **61**, 5337 (2000).

[18] F. de Juan, T. M., AG Grushin & Moore, J. Quantized circular photogalvanic effect in Weyl semimetals. *Nat. Commun.* **8**, 15995 (2017).

[19] Chan, C.-K., Lindner, N. H., Refael, G. & Lee, P. A. Photocurrents in Weyl semimetals. *Physical Review B* **95**, 041104 (2017).

[20] Yang, X., Burch, K. & Ran, Y. Divergent bulk photovoltaic effect in Weyl semimetals. *arXiv preprint arXiv:1712.09363* (2017).

[21] Ma, Q. *et al.* Direct optical detection of Weyl fermion chirality in a topological semimetal. *Nature Physics* **13**, 842 (2017).

[22] Osterhoudt, G. B. *et al.* Colossal photovoltaic effect driven by the singular Berry curvature in a Weyl semimetal. *arXiv preprint arXiv:1712.04951* (2017).

[23] Wu, L. *et al.* Giant anisotropic nonlinear optical response in transition metal monopnictide Weyl semimetals. *Nature Physics* **13**, 350 (2017).

[24] Penetration depth calculated using refractive index, as measured by ellipsometry, obtained in private communication from Bing Xu, Institute of Phyiscs, Beijing.

[25] Weng, H., Fang, C., Fang, Z., Bernevig, B. A. & Dai, X. Weyl semimetal phase in noncentrosymmetric transition-metal monophosphides. *Physical Review X* **5**, 011029

(2015).

[26] Li, Z. *et al.* Weyl semimetal TaAs: Crystal growth, morphology, and thermodynamics. *Crystal Growth & Design* **16**, 1172–1175 (2016).

[27] See appendices for additional details.

[28] Bergfeld, S. & Daum, W. Second-harmonic generation in GaAs: Experiment versus theoretical predictions of $\chi^{(2)}_{xyz}$. *Physical Review Letters* **90**, 036801 (2003).

[29] Morimoto, T. & Nagaosa, N. Topological nature of nonlinear optical effects in solids. *Science Advances* **2**, e1501524 (2016).

[30] Yahyavi, M. & Hetényi, B. Reconstruction of the polarization distribution of the Rice-Mele model. *Physical Review A* **95**, 062104 (2017).

[31] Souza, I., Wilkens, T. & Martin, R. M. Polarization and localization in insulators: Generating function approach. *Physical Review B* **62**, 1666 (2000).

[32] Resta, R. & Vanderbilt, D. Theory of polarization: a modern approach. In *Physics of Ferroelectrics*, 31–68 (Springer, 2007).

[33] Zhang, Y. *et al.* Photogalvanic effect in weyl semimetals from first principles. *Phys. Rev. B* **97**, 241118 (2018).

[34] Rice, M. J. & Mele, E. J. Elementary excitations of a linearly conjugated diatomic polymer. *Physical Review Letters* **49**, 1455–1459 (1982).

[35] Zak, J. Berry's phase for energy bands in solids. *Physical Review Letters* **62**, 2747 (1989).

[36] King-Smith, R. & Vanderbilt, D. Theory of polarization of crystalline solids. *Physical Review B* **47**, 1651 (1993).

[37] Resta, R. Macroscopic polarization in crystalline dielectrics: the geometric phase approach. *Reviews of modern physics* **66**, 899 (1994).

[38] Ortiz, G. & Martin, R. M. Macroscopic polarization as a geometric quantum phase: Many-body formulation. *Physical Review B* **49**, 14202–14210 (1994).

[39] Rangel, T. *et al.* Large bulk photovoltaic effect and spontaneous polarization of single-layer monochalcogenides. *Physical Review Letters* **119**, 067402 (2017).

[40] Xu, B. *et al.* Optical spectroscopy of the Weyl semimetal TaAs. *Physical Review B* **93**, 121110 (2016).

[41] Tan, L. Z. & Rappe, A. M. Upper limit on nonlinear optical processes: shift current and second harmonic generation in extended systems. *arXiv preprint arXiv:1708.05433* (2017).

[42] Sipe, J. E. & Shkrebtii, A. I. Second-order optical response in semiconductors. *Physical Review B* **61**, 5337–5352 (2000).

[43] Fregoso, B. M., Morimoto, T. & Moore, J. E. Quantitative relationship between polarization differences and the zone-averaged shift photocurrent. *Physical Review B* **96**, 075421 (2017).

[44] Quesnel, B. & Mora, P. Theory and simulation of the interaction of ultraintense laser pulses with electrons in vacuum. *Physical Review E* **58**, 3719 (1998).

[45] Lab2 add on to LabVIEW® software (1999). URL `http://www.lab2.de`.

[46] Bloembergen, N. & Pershan, P. Light waves at the boundary of nonlinear media. *Physical Review* **128**, 606 (1962).

[47] The dielectric constant was measured using ellipsometry of TaAs by Bing Xu at Université de Fribourg, and obtained in private communication.

[48] Kohn, W. Theory of the insulating state. *Physical Review* **133**, A171 (1964).

[49] Resta, R. Quantum-mechanical position operator in extended systems. *Physical Review Letters* **80**, 1800 (1998).