Thickness-independent transport in thin (001)-oriented cadmium arsenide films

David A. Kealhofer, Manik Goyal, Tyler N. Pardue, and Susanne Stemmer

Phys. Rev. B **104**, 035435 — Published 28 July 2021

DOI: 10.1103/PhysRevB.104.035435

Thickness-independent transport in thin (001)-oriented cadmium arsenide films

David A. Kealhofer[1], Manik Goyal[2], Tyler N. Pardue[2], and Susanne Stemmer[2*]

[1]Department of Physics, University of California, Santa Barbara, CA 93106-9530, USA.

[2]Materials Department, University of California, Santa Barbara, CA 93106-5050, USA.

*Corresponding author. Email: stemmer@mrl.ucsb.edu

Abstract

The three-dimensional Dirac semimetal is a parent phase for a variety of topological phases that can be generated by tuning parameters in material growth or device operation. Notably, it has recently been found that cadmium arsenide, which is ordinarily a three-dimensional Dirac semimetal, can nevertheless realize a three-dimensional topological insulator in (001)-oriented films about 50 nm thick. In this work, we study the quantum Hall effect in thin (001)-oriented cadmium arsenide films, their thickness ranging from 12 nm to 24 nm. When the carrier density is kept approximately constant across the different films, quantum transport reveals an identical underlying picture. The result is shown to be consistent with the transport's origin in the surface states of a three-dimensional topological insulator, but problematic for a perspective in which the quantum Hall effect originates from the confined subbands of the bulk band structure. These thin-film results complement previous studies of the quantum Hall effect in 50-nm-thick films.

1. Introduction

Three-dimensional (3D) Dirac semimetals like cadmium arsenide (Cd_3As_2) are considered 3D analogues of graphene because their low energy dispersion realizes the Dirac equation in three dimensions in the same way that graphene's low energy dispersion realizes it in two [1-4]. A practically important difference between a 3D Dirac semimetal and a 2D one such as graphene is that the third dimension provides a practical way of manipulating the bulk band structure. According to the bulk-boundary correspondence principle [5], such manipulation amounts to control over the topological surface states when the bulk band structure is characterized by a topological invariant different from that of the material surrounding it [6].

In Cd_3As_2, the subject of this work, the 3D Dirac nodes are due in part to a crystal symmetry (a fourfold rotation of the k_z axis) [2]. The nodes, which lie along [001] (i.e. parallel to k_z), result in surface states that vary according to the nodes' projection onto the relevant surface Brillouin zone [2]. It is thought, for example, that the surface states of (112)-oriented films consist of pairs of arcs that join the projected Dirac nodes in the surface Brillouin zone [2,7], similar to the Fermi arcs in the closely related Weyl semimetals [8].

In this work, however, we focus on (001)-oriented thin films, in which the nodes project onto the same point in the corresponding surface Brillouin zone. The surface state in this case is not Fermi arc-like. Instead, thin (001)-oriented films realize a 3D topological insulator (TI). The origin of the 3D TI state is the band inversion at the center of the Brillouin zone [2]. At each interface of the film, which lies between a compound semiconductor layer and a gate dielectric, the resolution of the band inversion results in a 2D Dirac surface state [9-11]. A key feature of these thin film heterostructures is the energy offset between the two surface states that arises from the different band offsets on either surface or interface. In a previous study, we showed that this

picture of the 3D TI surface state explains the sequences of filling factors in the quantum Hall effect measured in high magnetic fields [12]. Another test of the 3D TI picture is that, within a relevant range of film thickness, the Landau level spectrum and transport in magnetic fields should be insensitive to the thickness. This is the focus of the present work.

The remainder of this Article is organized as follows: In Section 2, we discuss a phenomenological model for the TI surface state and its relevant predictions. Then we turn to our experiments. The methods are described in Section 3, and the results—quantum Hall data from four samples of thicknesses between 12 and 28 nm—in Section 4. Section 5 contains a discussion of the experimental data with reference to the surface state and subband models. We conclude (Section 6) with a summary and some suggestions for future experiments.

2. Model

We apply a simple continuum model for the surface states of a 3D TI to parametrize the relevant underlying physics [13]. Near the center of the Brillouin zone, we expect that the surface states can be described by the following Hamiltonian:

$$H_0 = \left[\hbar v_F\left(k_x\sigma_y - k_y\sigma_x\right) + \frac{\Delta_i}{2}\mathbb{1}\right] \otimes \tau_z + \frac{\Delta_h}{2}\mathbb{1} \otimes \tau_x, \quad (1)$$

where the σ_i and τ_i are Pauli matrices, referring to a spin degree of freedom and a surface pseudospin degree of freedom, respectively. The k_i are the 2D crystal momenta, and v_F, the Fermi velocity, parametrizes the steepness of the (identical) Dirac cones when $\Delta_h = 0$. The term Δ_i is formally an inversion-breaking term in the sense that τ_z is an inversion symmetry operator. In a more concrete sense, it describes the effect of the inter-surface energy difference that we ascribe to the asymmetry of the Cd_3As_2 heterostructure, that is, the difference in band offsets on either side of the film. Finally, Δ_h allows a gap to open via the hybridization of the two surfaces. This

is expected to be relevant only in films thin enough for there to be significant wavefunction overlap between the surface states on the opposite surfaces.

In the absence of a magnetic field, the spectrum of this Hamiltonian is

$$E_{\alpha,\beta}(k) = \alpha\sqrt{(\hbar v_F k)^2 + (\Delta/2)^2 + \beta\,\hbar v_F |k| \Delta_i}, \tag{2}$$

where $\alpha, \beta = \pm 1$ independently and we substitute $\Delta = \sqrt{\Delta_i^2 + \Delta_h^2}$. Obviously, in the circumstance where $\Delta_i = \Delta_h = 0$, the result is a doubly degenerate Dirac cone.

In a magnetic field $\vec{B} = B\,\hat{z}$, the Zeeman effect acts on the real spin degree of freedom and is captured by the addition of the following term to the Hamiltonian, $H_Z = g^* \mu_B \mathrm{B}\, \sigma_z \otimes \mathbb{1}$, where g^* is an effective g factor, and μ_B is the Bohr magneton. More important, though, is the quantization of the spectrum into Landau levels. The calculation of this is accomplished by the Peierls substitution, $\hbar\vec{k} \to \vec{\Pi} = \hbar\vec{k} + e\vec{A}$, where we use the gauge $\vec{A} = xB\,\hat{y}$, and e is the magnitude of the electron charge. Because $\vec{k}$ and $\vec{A}$ do not commute, we introduce the ladder operators $a = (2eB\hbar)^{-1/2}(\Pi_y + i\,\Pi_x)$ and $a^\dagger = (2eB\hbar)^{-1/2}(\Pi_y - i\,\Pi_x)$. The resulting Hamiltonian *H(B)* reads:

$$H(B) = \begin{pmatrix} \frac{\Delta_i}{2} + g^*\mu_B B & -\sqrt{b(B)}\,a & \frac{\Delta_h}{2} & 0 \\ -\sqrt{b(B)}\,a^\dagger & \frac{\Delta_i}{2} - g^*\mu_B B & 0 & \frac{\Delta_h}{2} \\ \frac{\Delta_h}{2} & 0 & -\frac{\Delta_i}{2} + g^*\mu_B B & +\sqrt{b(B)}\,a \\ 0 & \frac{\Delta_h}{2} & +\sqrt{b(B)}\,a^\dagger & -\frac{\Delta_i}{2} - g^*\mu_B B \end{pmatrix}, \tag{3}$$

where we have used the shorthand $b(B) \coloneqq 2eB\hbar v_F^2$. The ladder operators are associated to states $|n\rangle$, where *n* is an integer ≥ 0, for which $a^\dagger\,|n\rangle = \sqrt{(n+1)}\,|\,n+1\rangle$, $a\;|n>0\rangle = \sqrt{n}\;|n-1\rangle$, and $a\,|0\rangle = 0$.

As long as *n > 0*, the eigenvectors $\Psi_n(B)$ of *H(B)* have the form:

$$\Psi_{n>0} = \begin{pmatrix} \Phi_{1,n}|n-1\rangle \\ \Phi_{2,n}|n\rangle \\ \Phi_{3,n}|n-1\rangle \\ \Phi_{4,n}\,|n\,\rangle \end{pmatrix},$$

where the $\Phi_{i,n}$ are numbers. When $n = 0$, however,

$$\Psi_{n=0} = \begin{pmatrix} 0 \\ \Phi_{2,0}|0\rangle \\ 0 \\ \Phi_{4,0}\,|0\,\rangle \end{pmatrix}.$$

Combining all these, we find the spectrum in a perpendicular field to be

$$E_{\alpha,\beta,n>0}(B) = \alpha\sqrt{(g^*\mu_B B)^2 + (\Delta/2)^2 + n\,b(B) + \beta\sqrt{n\,b(B)\Delta_i^2 + (g^*\mu_B B)^2\Delta^2}},$$

where, as above, $\Delta = \sqrt{\Delta_i^2 + \Delta_h^2}$, and $\alpha, \beta = \pm 1$ independently. For $n = 0$, $E_{\beta,\,n=0} = -g^*\mu_B B + \beta\Delta$.

The effect of tuning the model parameters Δ_h and Δ_i is illustrated in Fig. 1. If both $\Delta_h = 0$ and $\Delta_i = 0$, then the two Dirac cones are degenerate everywhere, as shown in Fig. 1(a). A finite Δ_i has the effect of shifting each cone relative to the other in energy, as seen in Fig. 1(b), so that the energy difference between the Dirac points is equal to Δ_i. By contrast, the effect of Δ_h, shown in Fig. 1(c), is to open a gap at the Dirac point; no degeneracy is split. Far from $k = 0$, the dispersion looks like that of Fig. 1(a). If both Δ_i and Δ_h are nonzero, the case of Fig. 1(d), then the dispersion far from Γ looks like that of Fig. 1(b), while a gap opens at $k = 0$.

Since Δ_i essentially tunes a splitting while Δ_h opens a gap, small changes in Δ_i substantially affect the Landau level spectrum, in contrast to even fairly large changes in Δ_h. This difference is illustrated in Figs. 1(e-f). In Fig. 1(e), Δ_i is fixed at 75 meV, and the spectra of Eq. (3) are plotted for different values of Δ_h, ranging from 0 to 60 meV. The effect of changing Δ_h is subtle and most

noticeable for the lowest Landau levels at low field, or, in the quantum Hall regime, the smallest filling factors. A spectroscopic experiment, sensitive to quantitative shifts in the Landau level energies, could in principle be sensitive to the shift of the lowest couple of Landau levels. Such energy shifts are, however, invisible to a transport experiment. Equivalently small changes in Δ_i, by contrast, drastically affect the Landau level spectrum. This point is illustrated in Fig. 1(f). Here Δ_h, now, is set at a fixed value and spectra are plotted for various Δ_i, ranging from 0 to 60 meV. A small change in Δ_i results in both splitting and shifting of the Landau levels, causing the locations of crossed Landau levels to change. In an experiment—a Hall measurement—this could be seen as a change in the sequence of quantum Hall filling factors observed as the external field is ramped, or as a change in the interplay of multiple frequencies in quantum oscillations.

3. Experimental Methods

Capped (001)-oriented Cd_3As_2 films were grown by molecular beam epitaxy and fabricated into gated Hall bar devices. Details regarding the growth and structural and electronic characterization of the resulting structures have been reported elsewhere [12,14-16]. The samples consist of a (100) GaSb substrate, cut 3º toward (111)B, onto which was grown a buffer layer of $In_xAl_{1-x}Sb$, a Cd_3As_2 layer, and finally a thin GaSb cap. Where noted, an Al_2O_3 gate dielectric was deposited *ex situ* using atomic layer deposition after the as-grown devices were first measured. The gate metal, on top of the dielectric, lies above the region containing Hall bar's voltage leads, and a dc bias is applied between the gate metal and the Cd_3As_2 film; the carrier density is determined from the low-field Hall effect. Quasi-dc Hall measurements were performed in a Quantum Design PPMS Dynacool using standard lock-in techniques and a 1 μA current. Raw resistance data were binned and interpolated before being symmetrized (R_{xx}) or antisymmetrized

(R_{xy}) with respect to B. The thickness of each sample was determined from cross-sections using transmission electron microscopy.

4. Results

We refer to transport measurements on four samples, A, B, C, and D. The samples differ in the thickness of the Cd_3As_2 layer, which is 12 nm for sample A, 14 nm for sample B, 18 nm for sample C, and 24 nm for sample D. Table I lists the film thicknesses, the carrier density and Hall mobility extracted from the traces in Figs. 2 and 3. Before we compare these four samples, we examine sample D in detail.

Figure 2 shows magnetotransport data from sample D. At magnetic fields below about 5 T, the plateaus are weak and proceed according to an apparent degeneracy of two, i.e. the filling factor ν steps from 10 to 8 to 6, as can be seen from panel (b), with no hint of other dips in the longitudinal magnetoresistance, R_{xx}, that might reveal the missing odd filling factors [panel (a)]. Around 6 T, the peak in R_{xx} is, however, clearly split, corresponding to a suppressed (that is, not observed) plateau at $\nu = 5$, and the plateau at $\nu = 3$ is more clearly recorded. The resolution of these odd-numbered plateaus at higher field is a main feature of these data, and, except for a different background, it is repeated in samples A, B, and C, as discussed below.

Comparisons between these samples must be made at fixed carrier density. As shown elsewhere, the carrier density and mobility depend strongly on the surface Fermi level, which in turn depends on the chemical and other boundary conditions of the sample surface [17]. In Fig. 3, different carrier densities are achieved by adjusting the top gate bias. Additional traces shown are from as-grown films, i.e. without the deposition of a gate dielectric. As grown, the carrier density varies across the samples. Here, the as-grown carrier density in samples A and B is nearly the

same (about 6.5×10^{11} cm^{-2}), while it is highest in sample C (2.4×10^{12} cm^{-2}), and sample D's falls in between (1.4×10^{12} cm^{-2}).

The Hall data from sample D, shown in Fig. 3(a), demonstrate that the same spectrum is relevant across a wide range of carrier densities, which is equal to about 25% the total: there are no qualitative changes. In other words, the evolution of the longitudinal magnetoresistance with the magnetic field, $R_{xx}(B)$ is nearly the same across this range of carrier density. Starting at 14 T and tracing $R_{xx}(B)$ toward $B = 0$, two pairs of peaks in the magnetoresistance are evident, as discussed above in the context of Fig. 2, resulting in an apparent degeneracy factor of two at low field. As the carrier density differs, so does the shape of the double peak that obscures $\nu = 5$, being essentially a single peak for the lowest-density trace and most clearly two overlapping peaks in the highest-density one.

A comparison between samples A, B, and D is shown in Figure 3(b), for a density of about 6.5×10^{11} cm^{-2}. Sample D, the thickest, exhibits the longest classical and quantum scattering times (we deduce the difference in quantum scattering times from the onset of the quantum oscillations and the width of the oscillating features [18]). Ignoring the difference in the magnetoresistance background and the broadening of the oscillations, all three traces exhibit the same behavior. Following all three $R_{xx}(B)$ traces from high to low field, a double peak is visible around 11 T, more or less resolved according to the oscillation width, followed by another around 6 T, which in samples A and B is hardly resolved at all in R_{xx}, but slightly clearer in R_{xy}. The sequence of filling factors appears to be identical, and the oscillations match modulo the difference in carrier density.

The same picture is visible in Figs. 3(c) and (d). Panel (c) shows the same two traces for samples A and B against two slightly higher density traces for samples C and D. The apparent phase shift of the R_{xx} oscillations in sample C versus those in sample D is clearly due to the

difference in carrier density: the same sequence of filling factors is seen in R_{xy}. Sample C, whose thickness is intermediate between that of samples B and D, has a mobility comparable to that of sample D (see Table I).

Panel (d) shows traces at high density (approximately 2 × that of panels b and c) for samples B, C, and D. (In this panel, data for sample D were acquired without the deposition of a gate.) Once again, while the background between the three samples varies significantly, and can be largely attributed to differences in the scattering times that are characteristic of the thickness of each sample (see Table I), the quantum oscillations for each sample reveal the same underlying Landau level spectrum. Here, steps in two of the filling factor are only resolved in R_{xx} in sample C.

Across panels (b) through (d), minima in R_{xx} are close to zero, but some amount of parallel conductance exists, similar to other studies of topological insulators [13]. The observed parallel conductance is not through three-dimensional bulk states, due to the lack of trend with film thickness. The smallest R_{xx} values recorded for the thinnest sample, sample A, approach 22 Ω [panel (b)], whereas for samples C and D, the two thickest, the smallest values are 14 Ω [panel (c)] and 34 Ω [panel (b)], respectively. Since thicker films should support more channels for parallel conductance through the bulk, we would expect to see more parallel conductance as the film thickness increases, which is not what is observed.

We separate the oscillating part of the magnetoresistance from the slowly varying (classical) magnetoresistance. As can be seen from the raw data (Fig. 3) the non-oscillating background differs greatly between the four samples. In all cases, the procedure is to interpolate the raw data on a grid in $1/B$. Then a weighted polynomial is fit to a subset of the interpolated data

and then subtracted. The results are shown in Figs. 4(a-d). The Fourier transform of these traces reveals the frequency components of the oscillations. These are shown in Figs. 4(e-h).

At low carrier density, the Fourier transform for all samples appears to have a single large peak between about 15 and 20 T. A high-frequency peak (40 to 60 T, depending on the sample), is also visible—it is most prominent in sample D [panels (d) and (h) in Fig. 4]. It is not a higher harmonic of the fundamental frequency. Instead, it is due to the resolution of the two fans that appears at high field, which appears as a doubled peak at high field. If the Fourier transform is applied only to the lower-field data—if we window out the double peak at high field—the high frequency peak disappears (not shown).

At high carrier density, every sample's Fourier transform consists of two comparable-magnitude peaks, at around 10 and 30 T. A reasonable question is whether the low-frequency (approx. 10 T) peak is spurious, i.e. introduced by an incomplete (or overzealous) background subtraction. One test of the background subtraction is whether the oscillations vary around zero, as they can be seen to in Figs. 4(a-d). The other test is the number: the Fourier transform of the subtracted polynomial has a low-frequency component if it oscillates (its derivative has zeroes) on the scale of the data. One can estimate that a fourth-order polynomial (the highest degree used here) has at most one full peak or dip in the positive half of the number line. If that were to fall in the range 2 T to 14 T (the plotted and Fourier-transformed range in Fig. 4), we would register a peak in the Fourier transform with a maximum of one half period per 12 T, i.e. a frequency in $1/B$ terms corresponding to about 6 T. By contrast, the lowest-frequency peaks seen here, at a frequency of 10 T, would register in the background-subtracted data as having two peaks separated by 0.1 T^{-1}, which is clearly a feature of the raw data, and not just the background-subtracted traces. Both these factors, the success of the background subtraction and the size of the frequency relative

to that characteristic of the background, lead us to conclude that the low frequency peaks measured are not spurious. The results and interpretation are further confirmed by direct fitting of the Shubnikov–de Haas oscillations, as shown in Figs. 4(i-l).

5. Discussion

The relevant question in applying 3D TI model discussed in Section 2 to these data is how the Landau level spectrum, which depends on Δ_h, Δ_i, and the carrier density n, should evolve under the influence of experimental parameters varied here, namely the thickness and gate voltage. The term Δ_h, which couples the two surfaces, is relevant when there is appreciable spatial overlap between the states on each surface, which we expect occurs only in very thin films. Heuristically, if the length scale for the Dirac state goes as $\hbar v_F/\Delta$, a Fermi velocity v_F of 8×10^5 m/s and a gap Δ of 100 meV suggest that hybridization of the surface states should occur in films thinner than about 6 nm, which is similar to the estimate in ref. [2]. The hybridization gap Δ_h, accordingly, should be negligible for films thicker than that. In other words, Δ_h may be a strong function of film thickness when the film is only a few nanometers thick, but, in the regime studied here, Δ_h is small and unchanging as the thickness is varied. In addition, as discussed in the exposition of Figure 1(e), the Landau level spectrum is essentially insensitive to modest changes in Δ_h as long as the Fermi energy lies outside of the gap, meaning that, even if Δ_h did vary substantially for films 12 to 24 nm thick, our experiments would likely not detect its influence.

The inversion-breaking term, Δ_i, we understand to be the energy difference between the Dirac nodes. Microscopically, Δ_i should be relevant when the confining potential is not symmetric about the center of the film, such as in the case, relevant here, when the thin film is surrounded on either side by different materials. Then it is the band alignments that cause the offset in energy

between the Dirac nodes of either surface state. According to that picture, Δ_i does not depend on the thickness of the Cd_3As_2 as long as it is great enough to separate the outer layers from each other, which, as for Δ_h, is the case for films more than a few nanometers thick.

What happens when a gate voltage is applied to the film? Applied to the top gate electrode, which sits on top of a dielectric (Al_2O_3), relative to the Cd_3As_2 film, the gate voltage varies the carrier concentration, though its effect is mitigated somewhat by the presence of the semiconductor cap layer. It also alters the band alignment, and so shifts the Dirac node of the top surface in energy. We thus expect Δ_i to be affected by the gate voltage, though this effect, too, is mitigated by the intervention of the cap layer.

We can check this reasoning by examining the frequency of the quantum oscillations (Fig. 4). The oscillating part of the magnetoresistance against $1/B$ has a frequency F that is proportional to the area of the orbit in reciprocal space, according to $F = (\hbar/2\pi e)A_k$, where e is the magnitude of the electron charge and A_k is the area of the orbit. A circular orbit, for example, has $A_k = \pi k_F^2$; k_F is the magnitude of the Fermi wavevector. Only extremal orbits contribute; if there are multiple extremal orbits with different areas, then multiple frequencies can be visible.

The Fourier transforms at low carrier density all resemble each other; those at high carrier density are likewise similar to each other. At low carrier density, the Fourier transform for all samples appears to have a single large peak between about 15 and 20 T. At high carrier density, the Fourier transforms all consist of two comparable-magnitude peaks, at around 10 and 30 T (all samples). In the TI surface state model presented in Section 2, finite values of Δ_i result in two frequencies for quantum oscillations, whose difference increases as a function of increasing E_F or carrier density. This can be seen heuristically by considering the case where $\Delta_h = 0$ and $\Delta_i > 0$. Then the dispersion looks like two offset Dirac cones [see Fig. 1(b)]. This results in two extremal

orbits: one around the higher-energy Dirac cone, which has a smaller radius, and the other around the lower-energy cone, which has a larger radius. Because the dispersion is linear, the difference in radius is constant. But, if the radii are $k_0 - \Delta k$ and $k_0 + \Delta k$, then the difference in area is $\Delta A = 4\pi k_0 \Delta k$. Since $k_0 \propto E_F$, as E_F increases, clearly $\Delta A \propto \Delta F$ increases, where ΔF is the difference in quantum oscillation frequency. Using the dispersion relation in Eq. (2), with $\Delta_h = 0$, one can calculate that the difference between the two frequencies for quantum oscillations is

$$\Delta F = \frac{E_F \Delta_i}{e\hbar v_F^2}.$$

As a sanity check, note that a difference in frequency of 20 T, suggests that the quantity $E_F\Delta_i \approx$ (60 meV)^2, assuming $v_F = 5 \times 10^5$ m/s [19]. This is indeed what is observed in Fig. 4. The 3D TI picture in the model therefore provides a satisfactory explanation for the essentially thickness-independent properties of the quantum Hall effect in these films.

It is instructive to examine what picture emerges from considering only the bands that form the 3D Dirac nodes. The Dirac nodes lie along the k_z axis at $k_z = \pm k_D$. Most first-principles calculations have found $k_D < 0.05$ Å^{-1} [2,20,21], consistent with several experimental studies [22-25], though there are some discrepancies—for a recent review, see ref. [4]. Since the length of the first Brillouin zone is 5 × or 10 × k_D, we consider a $k{\cdot}p$ approach to modeling the bulk band structure near the Dirac nodes to be accurate, as has been done elsewhere [2,26,27]. A naïve but effective way to model the thin film confinement is to treat an infinitely deep well, that is, quantize $k_z = n\pi/L$, with $n = 1, 2, 3\ldots$ and L the thickness of the film. By doing this, we have explicitly discarded surface states from our analysis. It is also worth noting that, though the confining potential $V(z)$ does not break the fourfold symmetry of the k_z axis, the bulk Dirac nodes are nevertheless destroyed. As remarked elsewhere [12] the agreement between this heuristic approach and more sophisticated ones [2] is nearly quantitative.

In-plane spectra, $E(k_x = k_y)$, are plotted in Fig. 5 across a range of thickness that includes the films studied here. Panels (a) through (f) show thicknesses from 6 nm (a) to 30 nm (f). (The numerical values of the $k \cdot p$ coefficients are taken from ref. [27].) The gap shrinks non-monotonically as thickness is increased: panel (g) shows the evolution of the gap at $k = 0$ as a function of thickness. Across the range of thickness studied in our experiments, the gap should decrease from a maximum of about 20 meV to as low as 5 or 10 meV; if some uncertainty is allowed in the correspondence between the model and reality, we should expect that the gap can take on arbitrarily small values near certain critical thicknesses. In any event, the prediction of the subband picture is that the thickness is a key parameter in determining the size of the gap. More qualitatively, as L increases, so does the number of subbands in any particular low-energy window. Comparing panels (b) through (e), which have thicknesses comparable to samples A through D, respectively, the number of conduction bands relevant to the transport increases from one to two or three (depending on E_F). As a result, the model predicts a commensurate increase in the complexity and/or apparent degeneracy of the Landau level spectrum. Since the essential feature of the experimental data, by contrast, is no change of the Landau level sequence with thickness, the subband picture cannot be said to agree with the experiment.

6. Conclusion and outlook

At fixed carrier density, the insensitivity of the quantum Hall effect to the film thickness as it is varied from 12 nm to 24 nm is problematic if the 2D states are thought to originate from the quantization of the bulk spectrum. The 3D TI picture in the model explored above is, by contrast, a satisfactory explanation for the essentially thickness-independent properties of the quantum Hall effect in these films. We emphasize that surface state transport is observed across an energy range

that is much larger than the comparatively small energy scale calculated for the overlap of the two As *p*-like bands that give rise to the bulk nodes, because of the much larger energy scale for the 5*s*-4*p* band inversion at the center of the Brillouin zone (hundreds of meV [2]). It remains an open question why bulk subband states do not give rise to observable magnetoresistance oscillations in these films. One possible explanation is that thin film strains may change the bulk band gap from those calculated in Fig. 5. A future direction for future research lies in dual-gated devices, which can disentangle tuning of the carrier density from that of Δ_i. That research will be enabled by optimization of the capping layer, gate dielectric [19,28], and device design, and is a critical step toward realizing the quantum spin Hall insulator state in cadmium arsenide.

Acknowledgments

The authors gratefully acknowledge support through the Vannevar Bush Faculty Fellowship program by the U.S. Department of Defense (Grant No. N00014-16-1-2814). The microscopy work was supported by the U.S. Department of Energy (Grant No. DEFG02-02ER45994). T. N. Pardue thanks the National Science Foundation Graduate Research Fellowship Program for support (Grant No. 1650114). This research made use of shared facilities of the UCSB MRSEC (NSF DMR 1720256).

Appendix A: π Berry phase from the fan diagram

Quantum oscillations from topological insulators are often analyzed in terms of a Berry phase, which is then used to support the topological nontrivial nature of their surface states. Extracting the phase from the fan diagram analysis is inherently fraught. At higher filling factors, the analysis suffers because non-ideality of the Dirac spectrum (i.e. a nonlinear dispersion) causes

a departure from the expected behavior, and there is a long lever. Furthermore, a large g factor causes deviations at low filling factor (the case here). These points are discussed in detail in ref. [29]. Others have identified difficulties arising from inhomogeneity (at the scale considered, not relevant here) and, more subtly, a constant density vs. constant chemical potential criterion, the applicability of which can in principle change as a function of e.g. gate voltage [30]. Additionally, particle-hole asymmetry jeopardizes the analysis of the Berry phase from the fan diagram for the 2D surface states of 3D TIs [31] as well as in bulk 3D Weyl and Dirac semimetals [32].

Besides these fundamental concerns, there are practical difficulties in applying the fan diagram analysis to our data, which fall into a combination of the quantum Hall regime and a transitional or incipient regime where the gaps between Landau levels are not fully established, there is significant parallel conductance, but the Shubnikov–de Haas oscillations are clear. In other words, these data bridge low Landau level and high Landau level regimes, and the crossover not only splits our data in two, but also adds a layer of ambiguity to the analysis.

Figure A1(a-d) shows a plot of maxima in $\sigma_{xx} = R_{xx}/\left(R_{xx}^2 + R_{xy}^2\right)$ vs. an integer n that simply indexes the counted maxima. These are used to construct the corresponding fan diagrams [Fig. A1(e-n)] as follows. First, the resistance (R_{xx} and R_{xy}) data are used to calculate the conductance $\sigma_{xx} = R_{xx}/\left(R_{xx}^2 + R_{xy}^2\right)$. Second, the background is subtracted by fitting $\sigma_{xx}(B) = \sigma_0 + \sigma_{1/2}/\sqrt{B} + \sigma_1/B + \sigma_2/B^2 + \sigma_3/B^3$ to the $B > 1$ portion of the data (the data are weighted like B^{-2} to counteract the influence of the widening quantum oscillations on the fit). Third, peaks in the subtracted data (i.e. the fit residuals), called $\Delta\sigma_{xx}$ in Fig. A1 above, are identified using a peak finding routine. At this stage, a couple of peaks are added and removed by hand (most low-field peaks are removed; double-peak features are added by hand on a maximum value criterion). After this, the identified peaks are plotted on top of the subtracted data in Figs. A1(a-d) as open

circles. Fourth, the field values where the peaks B_i have been found are used to assemble the fan diagrams, where the peak positions are plotted in inverse fashion ($1/B_i$) against an integer index, called *n*, which simply counts the number of maximums, from $n = 1$ at the highest-field peak, counting up through the lower-field ones. Last of all, we fit a line to the fan diagrams. In doing so we exclude peaks that are doubled: the highest-field peaks in panels (e), (h), (i), (j), and (k), and the three highest-field peaks in panels (m) and (n). We have also excluded the extreme low-field peak in panel (l).

The fitted *x*-intercepts accompany the fan diagrams in Figs. A1(e-n). In most cases we see values near to 0.5 (after shifting *n* by the appropriate integer), corresponding to a Berry phase of π. For the reasons enumerated at the beginning of this section, we present this as tentative but not necessarily determinant support that reflects the nature of the topological surface states.

The exclusion of some of the peaks in the fitting of the data in panels A1(e-n) affects the fitted intercept. The rationale for these exclusions is the same as for avoiding the filling factor plot (Fig. A2) to find the Berry phase: were one to make a full accounting of all the filled Landau levels, transforming the index *n* into something like ν, one would simply recover the Landau level degeneracy formula. The analogy to graphene is perhaps the clearest way of looking at it (see Fig. 1(c) in ref. [30]). Another way of saying this is written above: the data fall in the regime where some of the data crosses between an incipient regime (the fitted data in the fan diagrams here) and the deep quantum Hall regime (partially excluded here), which can be seen from the low-*n* kinks in the fan diagrams.

Figure A2 shows nearly the same plots as Fig. A1(e-n), with a crucial difference: ν is plotted instead of the arbitrary index *n*. Each $R_{xx}(B)$ minimum is indexed by the concurrent value of $\nu = R_K/R_{xy}$, where $R_K = h/e^2$ is the von Klitzing constant. (Since there are fewer points where

ν can be identified, the fans here are somewhat sparser than in Fig. A1.) Note that, unlike in the *n*-indexed plots, the expected value for the *y* intercept is zero. This is because, regardless of the zero-field spectrum (linear vs. parabolic dispersion, presence or absence of Berry monopole), this plot reflects only the Landau level degeneracy, i.e., for the Nth Landau level, $1/B_N = Ne/hn_{2D}$, see, e.g., discussion of Fig. 1(c) in ref. [30]. This is in contrast to the plots in Fig. A1, which have intercepts that are approximately an integer-and-a-half. But this plot, in which the Berry phase does not appear as an intercept, demonstrates that, even with unambiguous peak indexing as in the quantum Hall regime, there is some amount of error (see the nonzero intercepts for the high-density sample C and D traces). With that caveat, we observe as well that, along the lines of the discussion in ref. [30], the discrepancy between the ν- and *n*-indexed plots indicates against a reservoir of bulk states pinning the Fermi level of the surface states.

References

[1] S. M. Young, S. Zaheer, J. C. Y. Teo, C. L. Kane, E. J. Mele, and A. M. Rappe, Phys. Rev. Lett. **108**, 140405 (2012).

[2] Z. J. Wang, H. M. Weng, Q. S. Wu, X. Dai, and Z. Fang, Phys. Rev. B **88**, 125427 (2013).

[3] N. P. Armitage, E. J. Mele, and A. Vishwanath, Rev. Mod. Phys. **90**, 015001 (2018).

[4] I. Crassee, R. Sankar, W.-L. Lee, A. Akrap, and M. Orlita, Phys. Rev. Mater. **2**, 120302 (2018).

[5] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).

[6] B.-J. Yang and N. Nagaosa, Nat. Comm. **5**, 4898 (2014).

[7] M. Kargarian, M. Randeria, and Y. M. Lu, Proc. Natl. Acad. Sci. **113**, 8648 (2016).

[8] X. G. Wan, A. M. Turner, A. Vishwanath, and S. Y. Savrasov, Phys. Rev. B **83**, 205101 (2011).

[9] B. A. Volkov and O. A. Pankratov, JETP Letters **42**, 178 (1985).

[10] O. A. Pankratov, S. V. Pakhomov, and B. A. Volkov, Solid State Commun. **61**, 93 (1987).

[11] S. Tchoumakov, V. Jouffrey, A. Inhofer, E. Bocquillon, B. Placais, D. Carpentier, and M. O. Goerbig, Phys. Rev. B **96**, 201302 (2017).

[12] D. A. Kealhofer, L. Galletti, T. Schumann, A. Suslov, and S. Stemmer, Phys. Rev. X **10**, 011050 (2020).

[13] C. Brüne, C. X. Liu, E. G. Novik, E. M. Hankiewicz, H. Buhmann, Y. L. Chen, X. L. Qi, Z. X. Shen, S. C. Zhang, and L. W. Molenkamp, Phys. Rev. Lett. **106**, 126803 (2011).

[14] D. A. Kealhofer, H. Kim, T. Schumann, M. Goyal, L. Galletti, and S. Stemmer, Phys. Rev. Mater. **3**, 031201 (2019).

[15] T. Schumann, M. Goyal, H. Kim, and S. Stemmer, APL Mater. **4**, 126110 (2016).

[16] M. Goyal, S. Salmani-Rezaie, T. N. Pardue, B. H. Guo, D. A. Kealhofer, and S. Stemmer, APL Mater. **8**, 051106 (2020).

[17] L. Galletti, T. Schumann, T. E. Mates, and S. Stemmer, Phys. Rev. Mater. **2**, 124202 (2018).

[18] D. Shoenberg, Magnetic Oscillations in Metals (Cambridge University Press, Cambridge, 1984).

[19] O. F. Shoron, M. Goyal, B. H. Guo, D. A. Kealhofer, T. Schumann, and S. Stemmer, Adv. Electron. Mater. **6**, 2000676 (2020).

[20] M. N. Ali, Q. Gibson, S. Jeon, B. B. Zhou, A. Yazdani, and R. J. Cava, Inorg. Chem. **53**, 4062−4067 (2014).

[21] A. M. Conte, O. Pulci, and F. Bechstedt, Sci. Rep. **7**, 45500 (2017).

[22] S. Jeon, B. B. Zhou, A. Gyenis, B. E. Feldman, I. Kimchi, A. C. Potter, Q. D. Gibson, R. J. Cava, A. Vishwanath, and A. Yazdani, Nat. Mater. **13**, 851 (2014).

[23] A. Akrap, M. Hakl, S. Tchoumakov, I. Crassee, J. Kuba, M. O. Goerbig, C. C. Homes, O. Caha, J. Novak, F. Teppe, et al., Phys. Rev. Lett. **117**, 136401 (2016).

[24] M. Hakl, S. Tchoumakov, I. Crassee, A. Akrap, B. A. Piot, C. Faugeras, G. Martinez, A. Nateprov, E. Arushanov, F. Teppe, et al., Phys. Rev. B **97**, 115206 (2018).

[25] G. Krizman, T. Schumann, S. Tchoumakov, B. A. Assaf, S. Stemmer, L. A. de Vaulchier, and Y. Guldner, Phys. Rev. B **100**, 155205 (2019).

[26] J. Bodnar, 3rd International Conference on Physics of Narrow-Gap Semiconductors (arXiv:1709.05845), Warsaw, p. 311 (1978).

[27] J. Cano, B. Bradlyn, Z. Wang, M. Hirschberger, N. P. Ong, and B. A. Bernevig, Phys. Rev. B **95**, 161306(R) (2017).

[28] O. F. Shoron, T. Schumann, M. Goyal, D. A. Kealhofer, and S. Stemmer, Appl. Phys. Lett. **115**, 062101 (2019).

[29] A. A. Taskin and Y. Ando, Phys. Rev. B **84**, 035301 (2011).

[30] A. Y. Kuntsevich, A. V. Shupletsov, and G. M. Minkov, Phys. Rev. B **97**, 195431 (2018).

[31] A. R. Wright and R. H. McKenzie, Phys. Rev. B **87**, 085411 (2013).

[32] C. M. Wang, H.-Z. Lu, and S.-Q. Shen, Phys. Rev. Lett. **117**, 077201 (2016).

Table I. Hall density and mobility extracted from the data shown in Figs. 2 and 3. The Hall mobility is calculated by fitting a line to the < 0.5 T antisymmetrized R_{xy} data to find the Hall coefficient, which is then divided by R_{xx}(0 T).

	Thickness (nm)	n_{2D} (10^{11} cm^{-2})	μ_H (cm^2/Vs)
Sample A	12	6.43	5,380
Sample B	14	6.92	3,260
		14.0	2,490
		15.0	2,420
Sample C	18	7.62	21,600
		14.5	17,700
		15.0	17,400
Sample D	24	6.52	21,700
		6.95	19,400
		13.6	10,800

Figures with Captions

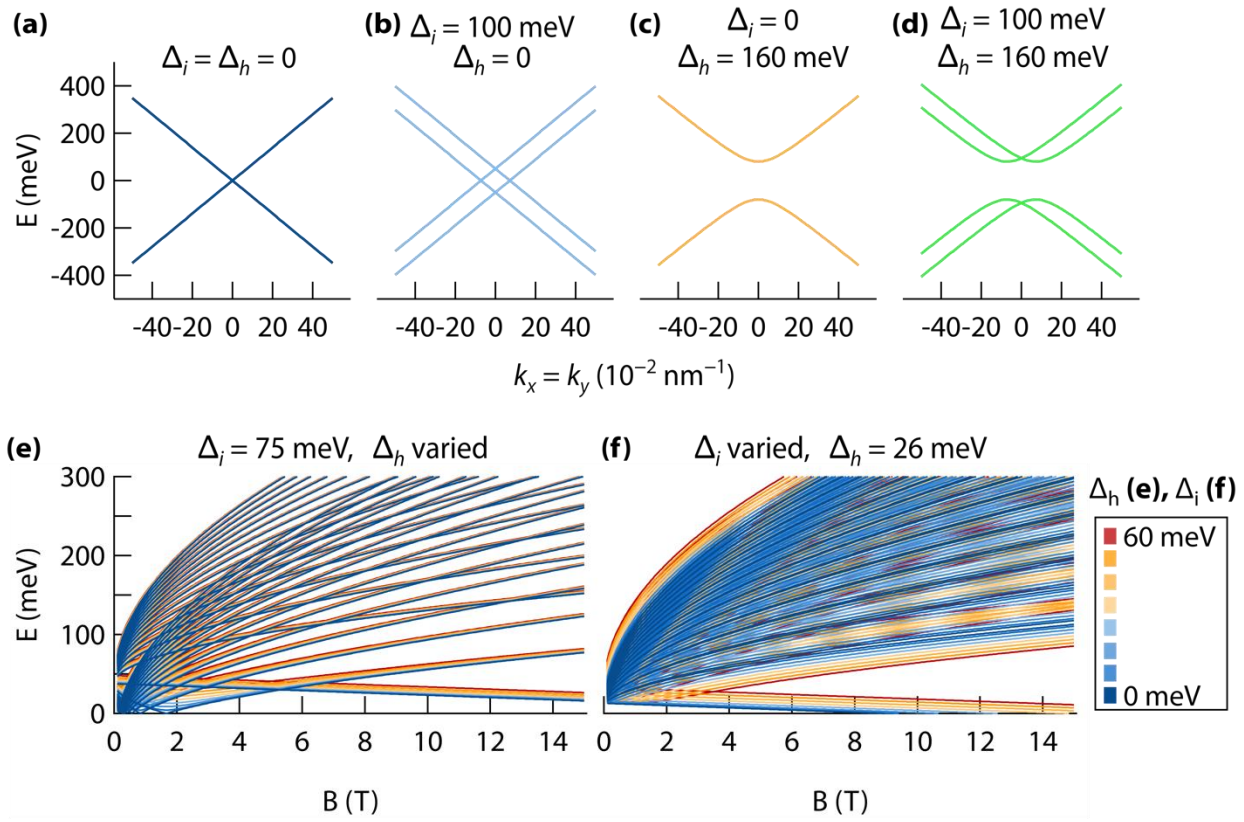

Figure 1: Model Hamiltonian spectra. (a)-(d) Zero-field spectra for small k and $v_F = 8 \times 10^5$ m/s. The parameter values are chosen for clarity. In panel (a), the cones are doubly degenerate everywhere because $\Delta_i = \Delta_h = 0$. (b) The effect of finite Δ_i is to displace the two cones in energy. The upper and lower cones are associated one to the eigenvalues of the τ_z operator. (c) By contrast, a finite Δ_h opens a gap in the spectrum, minimal at $k = 0$, which affects both surfaces in equal measure; the two bands are doubly degenerate everywhere. (d) When both Δ_h and Δ_i are appreciable, a gap opens and the two bands are in general nondegenerate. Whether away from $k = 0$ or $E = 0$, however, the spectrum is qualitatively and quantitatively similar to that in panel (b). (e) and (f) Landau level spectra as a function of Δ_h and Δ_i. Throughout g^* is set at +25. (The positive sign means that the zeroth Landau level disperses lower in energy with increasing field.)

The changing parameter is colored according to the scale at the far right. (e) Effect of changing Δ_h with finite Δ_i. The Landau levels are shifted in energy, which is more noticeable for Landau levels with low indices. (f) Effect of changing Δ_i with finite Δ_h. Small changes to Δ_i cause large changes to the spectrum because the two fans [more visible in panel (e)] are pushed to higher and lower energies, respectively.

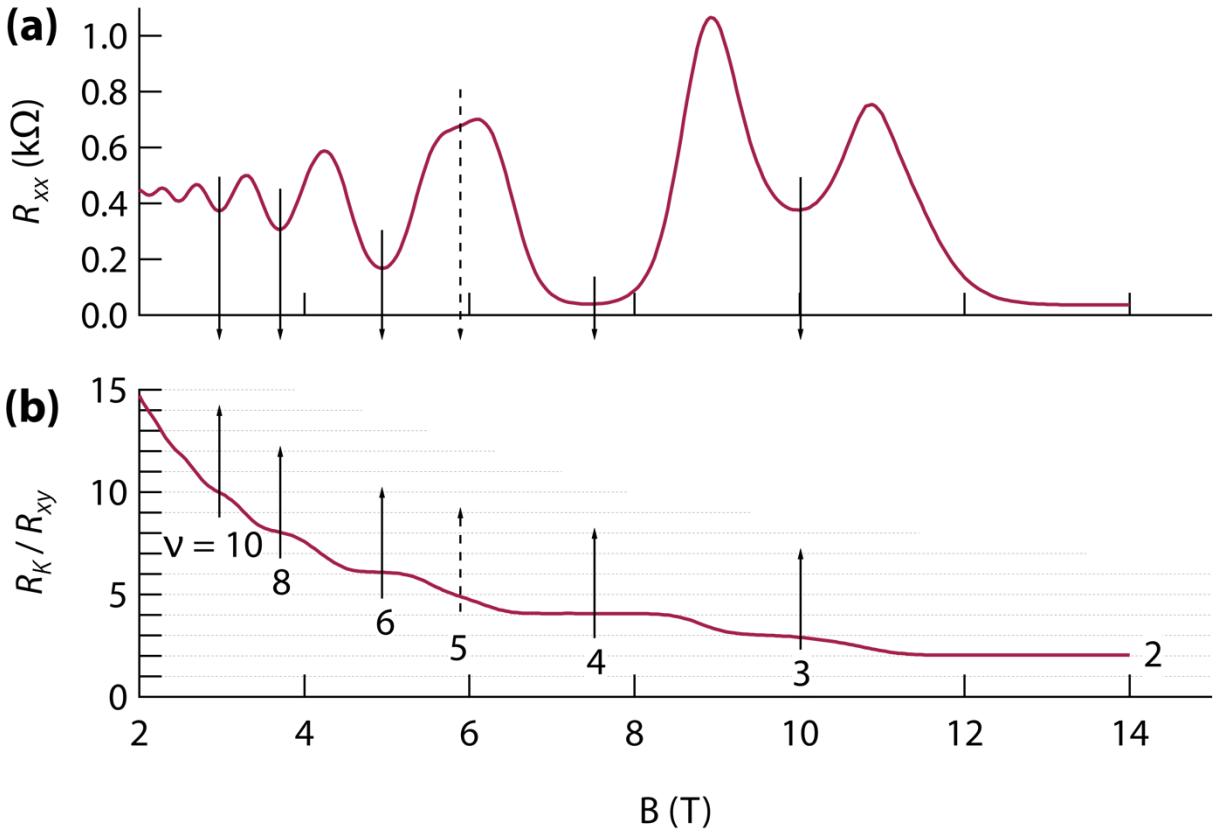

Figure 2: The quantum Hall effect in sample D ($n_{\mathrm{Hall}} = 6.95 \times 10^{11}$ cm^{-2}). (a) The longitudinal magnetoresistance, R_{xx}, acquired at 2 K, is plotted against magnetic field. Minima are indicated with arrows, matching those in the lower panel. (b) The von Klitzing constant ($R_K = h/e^2$) is divided by R_{xy}, the Hall resistance, to show that the plateaus match integer filling factors ν, indicated by the dashed lines and labels. Arrows correspond to the minima in R_{xx}, shown in panel (a). Though a plateau does not form with $\nu = 5$, a corresponding minimum in R_{xx} is nevertheless visible (dashed arrows).

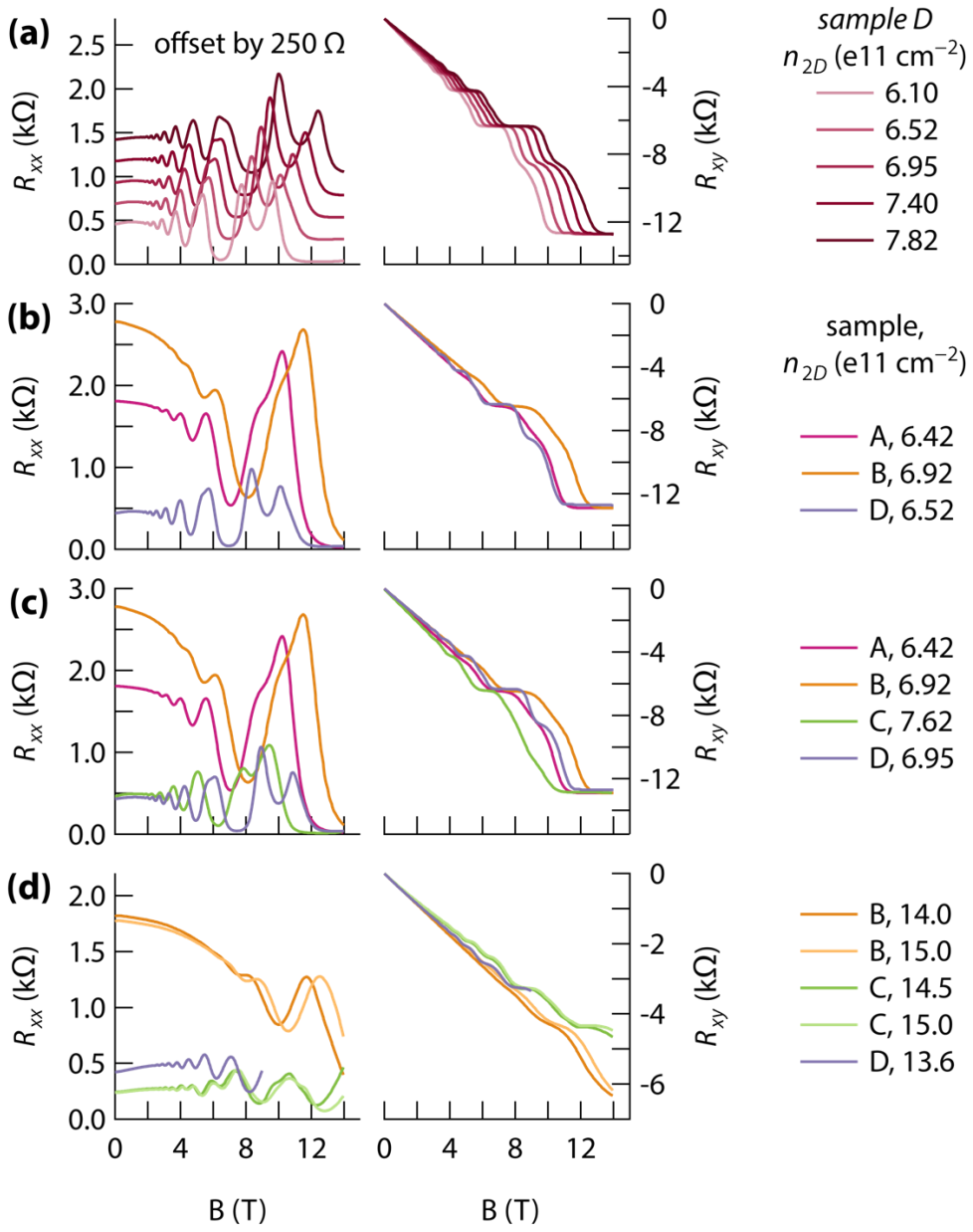

Figure 3: A comparison of the quantum Hall effect, measured at 2 K, in samples A, B, C, and D at various values of carrier density. (a) The quantum Hall effect in sample D. Each trace is recorded under a different top gate bias, corresponding to a different carrier density. For legibility, the traces in the longitudinal magnetoresistance, R_{xx}, are offset from each other by sequential multiples of 250 Ω. The true values are all comparable to the lowest resistance trace, which is not

offset. (b)-(d) A comparison of the quantum Hall effect in samples A, B, C and D at carrier densities of (b) about 6.5×10^{11} cm^{-2} (traces for samples A and B were acquired prior to gate deposition), (c) about 7×10^{11} cm^{-2} (traces for samples A and B were acquired prior to gate deposition), and (d) about 1.5×10^{12} cm^{-2} (the trace for sample D was acquired prior to gate deposition). No offsets have been added in panels (b)-(d).

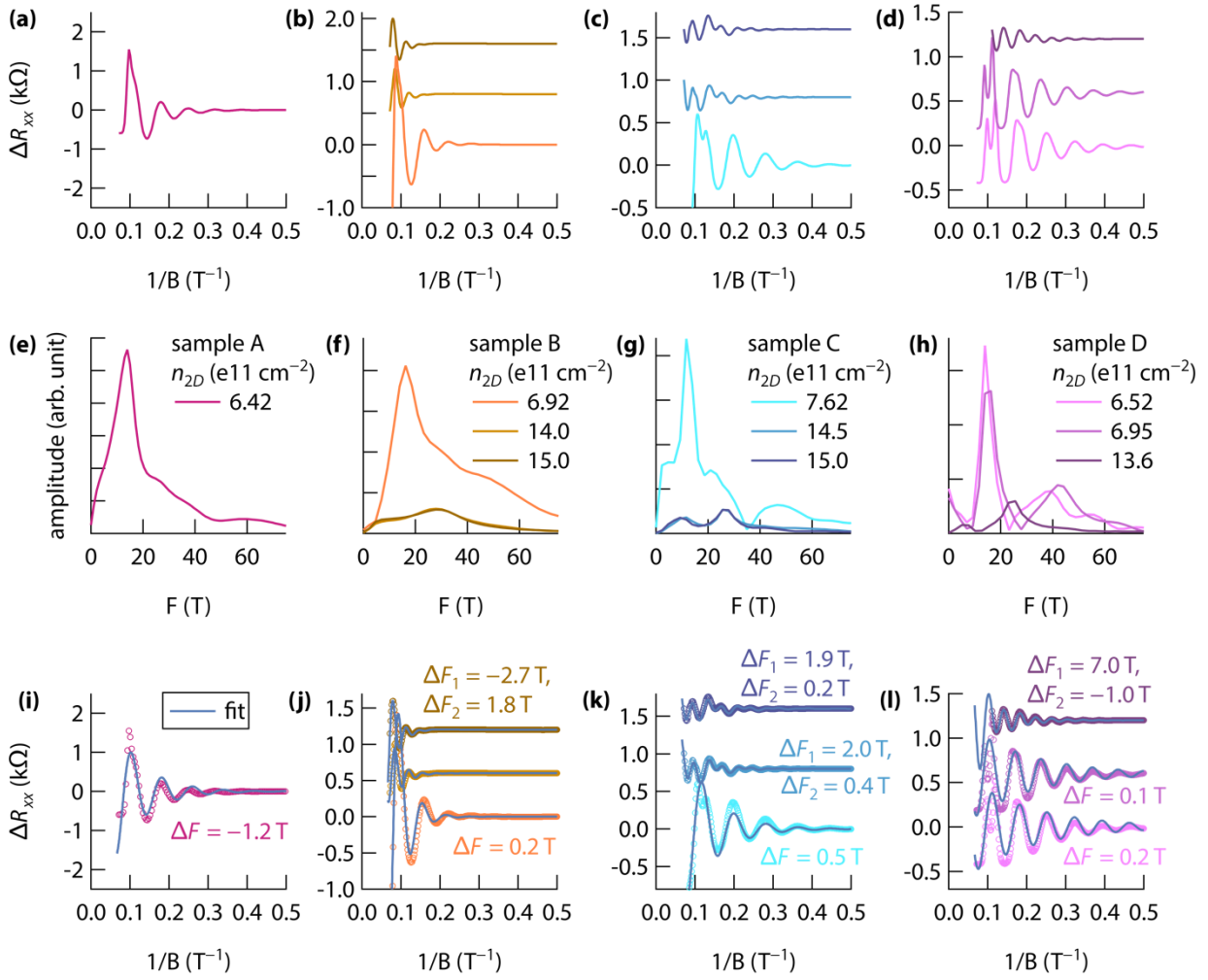

Figure 4: Frequency analysis of quantum oscillations. (a)-(d) Oscillations with background subtracted for samples A through D, respectively. (e)-(h) Fourier transform of the background-subtracted data from samples A through D, respectively. (i)-(l) Fits (lines) of the background-subtracted oscillations (markers) according to $\Delta R_{xx} = A\exp(-B_0/B)\cos(2\pi F/B + \phi)$ and $\Delta R_{xx} = A_1\exp(-B_{0,1}/B)\cos(2\pi F_1/B + \phi_1) + A_2\exp(-B_{0,2}/B)\cos(2\pi F_2/B + \phi_2)$, where $A, A_1, A_2, B_0, B_{0,1}, B_{0,2}, F, F_1, F_2, \phi, \phi_1,$ and ϕ_2 are fit parameters. The extracted oscillation frequencies F, F_1, and F_2, are compared to those derived from the Fourier transform. Note that the oscillation data are interpolated on even intervals in 1/B. Panels (i)-(l) are labeled by the difference $\Delta F = F_{\text{fit}} - F_{FT}$. Where two cosines are used to fit the data, two values are reported, the lower first.

The fit did not converge for the middle sample B trace, and no difference is reported. The agreement for the high-frequency peak is generally very good, less than 2 T—that is, ΔF and ΔF_2 are generally small.

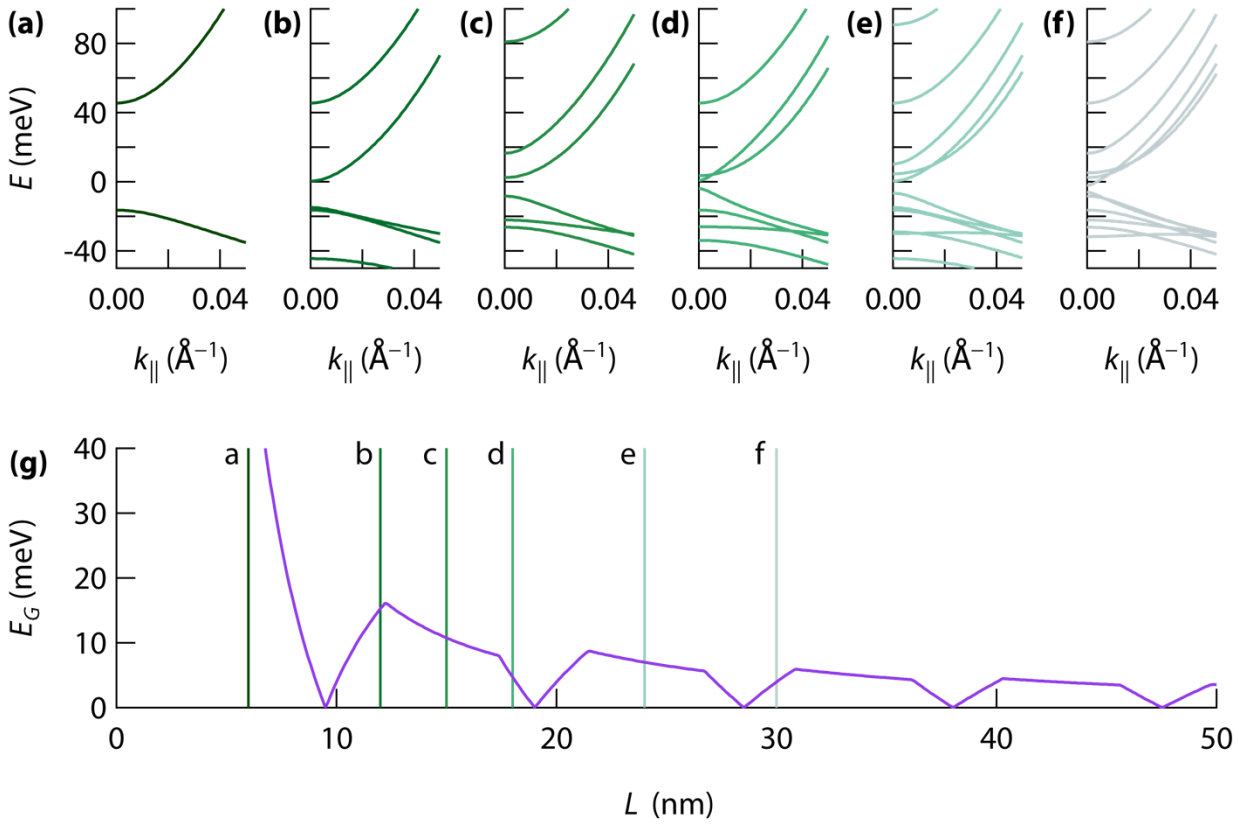

Figure 5: Thickness dependence in the subband picture. (The $k \cdot p$ model and coefficients are those from ref. [27]) In-plane spectra ($k_{||} = k_x = k_y$) are plotted for a film thickness L equal to (a) 6 nm, (b) 12 nm, (c) 15 nm, (d) 18 nm, (e) 24 nm, and (f) 30 nm. (g) Evolution of the gap E_G as a function of L. The thicknesses corresponding to panels (a)-(f) are marked with labeled, colored vertical lines.

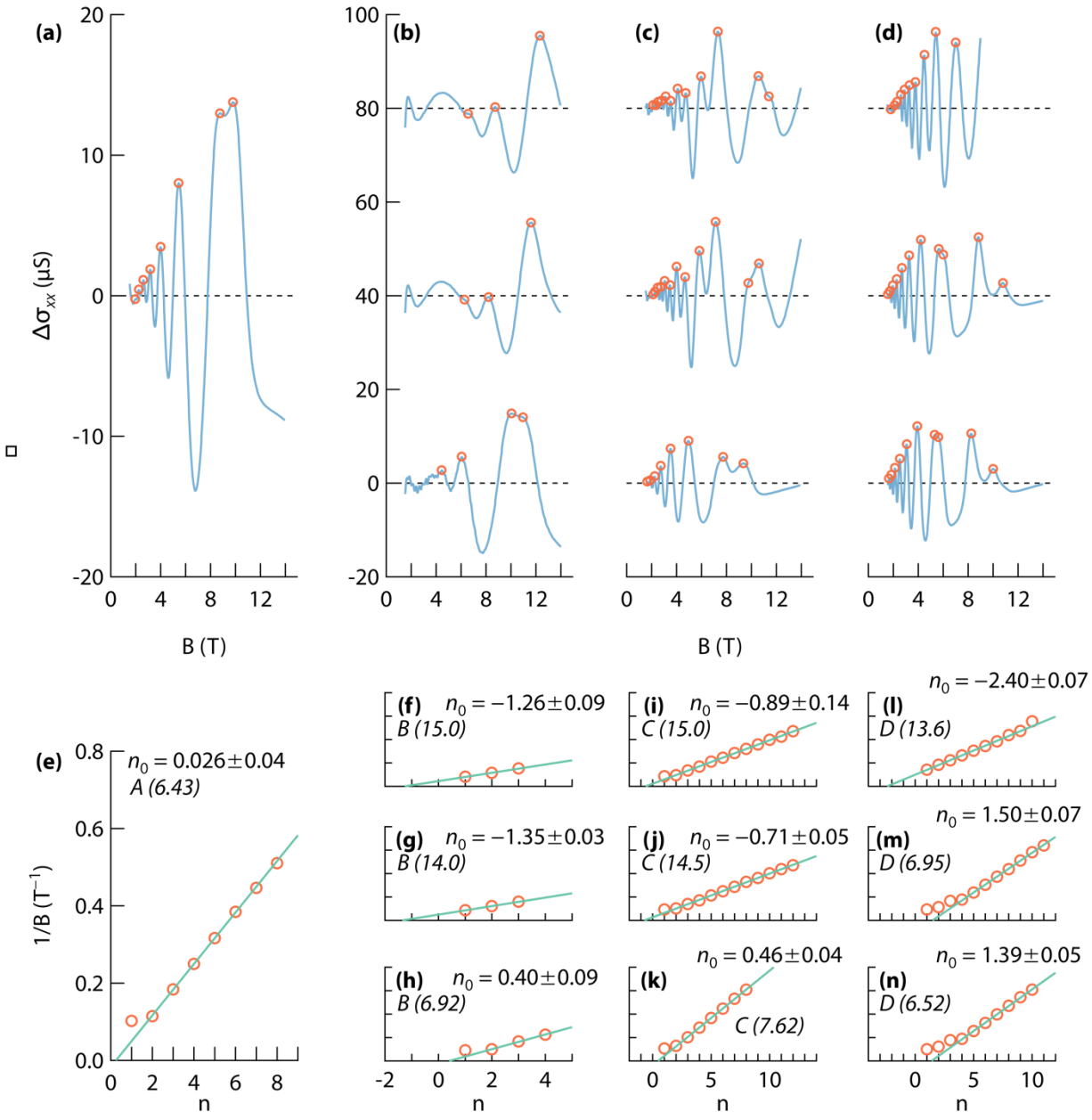

Figure A1: Background-subtracted oscillations in conductivity (a)-(d) and corresponding fan diagrams (e)-(n). Panels (a)-(d) show quantum oscillations, plotted against magnetic field, for samples A-D, respectively. For samples B, C, and D, traces are offset by 40 and 80 μS (zero is denoted by a horizontal dashed line). In all cases the higher density traces are on the top; the scheme follows the indexing of the fan diagrams in the bottom half of the figure. Peaks are identified in the conductivity. The magnetic field values for the peak centers are used to make the

fan diagrams in the bottom half of the figure, in panels (e) through (n). These fan diagrams are identified by the sample (a letter) and the carrier density (in units of 10^{11} cm^{-2}). Note that not all peaks (open circles) are fitted (solid lines). The *x*-intercept is identified as n_0 in each panel, along with an error (one standard deviation). Finally, note that the abscissa of the fan diagrams (e)-(n) is in all cases simply an un-adjusted integer index of the conductivity peaks, counting up from 1 for the peak at highest field.

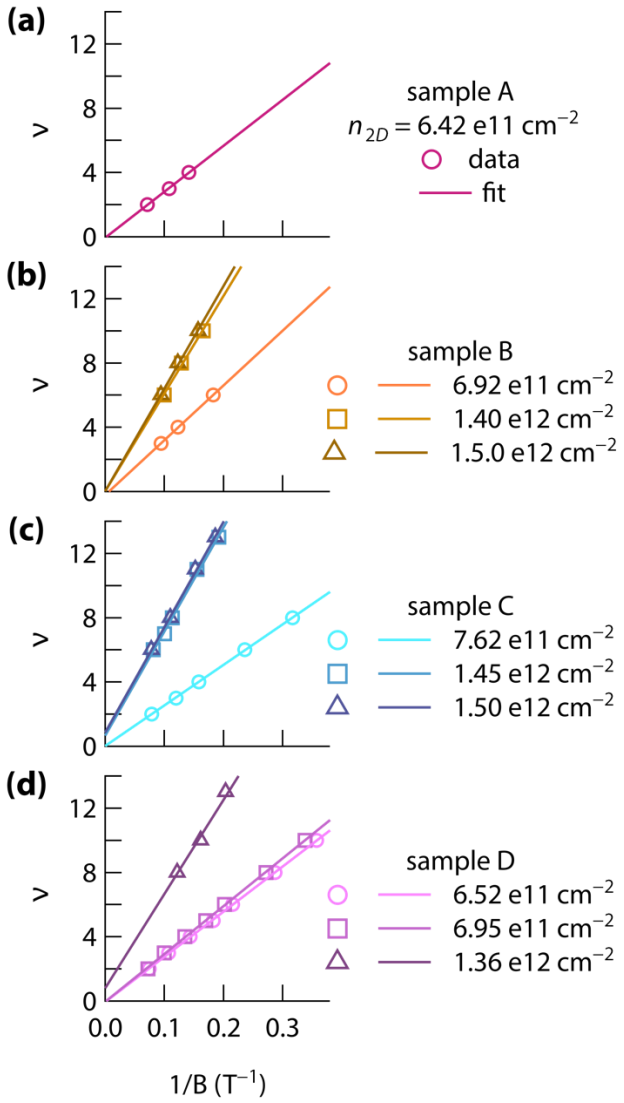

Figure A2: Fan diagrams extracted from the magnetoresistance data shown in Fig. 3. Each minimum in $R_{xx}(B)$ is indexed by the concurrent value of $\nu = R_K/R_{xy}$, rounded to an integer, which is plotted against the value of the inverse of the magnetic field $(1/B)$ where the minimum occurs. Linear fits are shown; *y*-intercepts are consistent with zero except for the high-density traces in samples C and D.