PHYSICAL REVIEW B **96**, 075204 (2017)

Length scale of puddle formation in compensation-doped semiconductors and topological insulators

Thomas Bömerich, Jonathan Lux, Qingyufei Terenz Feng, and Achim Rosch[*]

Institute for Theoretical Physics, University of Cologne, D-50937 Cologne, Germany

(Received 30 March 2017; revised manuscript received 24 July 2017; published 17 August 2017)

In most semiconductors and insulators the presence of a small density of charged impurities cannot be avoided, but their effect can be reduced by compensation doping, i.e., by introducing defects of opposite charge. Screening in such a system leads to the formation of electron-hole puddles, which dominate bulk transport, as first recognized by Efros and Shklovskii. Metallic surface states of topological insulators (TIs) contribute extra screening channels, suppressing puddles. We investigate the typical length ℓ_P, which determines the distance between puddles and the suppression of puddle formation close to metallic surfaces in the limit where the gap Δ is much larger than the typical Coulomb energy E_c of neighboring dopants $\Delta \gg E_c$. In particular, this is relevant for three-dimensional Bi-based topological insulators, where $\Delta/E_c \sim 100$. Scaling arguments predict $\ell_P \sim (\Delta/E_c)^2$. In contrast, we find numerically that ℓ_P is much smaller and grows in an extended crossover regime approximately linearly with Δ/E_c for numerically accessible values $\Delta/E_c \lesssim 35$. We show how a quantitative scaling argument can be used to extrapolate to larger Δ/E_c, where $\ell_P \sim (\Delta/E_c)^2/\ln(\Delta/E_c)$. Our results can be used to predict a characteristic thickness of TI thin films, below which the sample quality is strongly enhanced.

DOI: 10.1103/PhysRevB.96.075204

I. INTRODUCTION

The formation of locally conducting puddles is a phenomenon caused by charged Coulomb disorder in insulators, semiconductors, and Dirac matter like graphene, topological surface states, or Weyl semimetals. Efros and Shklovskii [1] predicted that puddle formation is, in three dimensions (3D), an unavoidable consequence of the long-range nature of the Coulomb interaction. Puddles are formed to screen large potential fluctuations exceeding the size of the gap Δ.

In graphene it has been shown both theoretically and experimentally that puddles are necessary to understand most transport experiments close to charge neutrality [2–5]. Recently, they have been observed in the bulk of a three-dimensional topological insulator [6,7]. These materials, from the class of the $Bi_{2-x}Sb_xTe_{3-y}Se_y$ compounds [8], are almost perfectly compensated semiconductors with a band gap of order 250–300 meV almost 2 orders of magnitude larger than the typical Coulomb energy E_c of neighboring dopants [6]. The relatively high density ($>10^{19}$ cm^{-3}) of dopants implies a strongly fluctuating Coulomb potential in the bulk. This leads to band bending and eventually to the formation of electron and hole puddles [9,10]. The additional surface states in the topological materials induce an additional screening channel close to the surface. Here surface puddles form [10,11] which are akin to puddles that form in graphene on a substrate which has charged impurities.

As the puddles are separated by insulating regions, they do not directly contribute to the dc conductivity. However, they do contribute to the optical conductivity at finite frequencies, which has been used to detect their presence and to measure the effective charge density in conducting regions [6]. We have shown that screening from thermal excitations can efficiently suppress puddle formation leading to a characteristic temperature dependence of the optical response [6]. Furthermore, in similar compounds a giant negative magnetoresistance was found experimentally and explained by merging of puddles driven by the Zeeman effect [12].

Surface puddles and puddles in graphene can be observed directly in real space by scanning tunneling microscopy (STM) [4,5,13]. From the two-dimensional (2D) STM maps, the size of the potential fluctuations and the corresponding length scale can be directly read off. These agree well with theoretical results, where these quantities are calculated self-consistently [2,3,11]. However, nothing is known experimentally about the length scales of puddles in the bulk and the effect of surface screening on the bulk puddle formation.

In the following we demonstrate numerically that the length scales governing the distance of puddles, the suppression of (bulk) puddles close to surfaces of TIs, and the suppression of puddles in thin films grow much slower with Δ/E_c than expected from scaling arguments. First we introduce the model and consider the scaling behavior of the charge-charge correlation function. We show numerical results for the bulk, and demonstrate that the simple scaling theory fails. Then we additionally take into account the gapless surface states which provide an extra screening channel. The length scales governing the size of puddles on the surface is different, and independent of the bulk band gap [11]. The bulk length describes, however, the size of a region where surface screening suppresses the formation of bulk puddles and is therefore important to understand the properties of thin topological insulator samples. We use scaling arguments to extrapolate our numerical results for $\Delta/E_c \lesssim 35$ to the experimentally relevant regime of $\Delta/E_c \sim 100$.

II. MODEL AND SIMULATIONS

Bi-based topological insulators typically have a very large dielectric constant $\varepsilon \approx 200$. Electron binding energies are therefore small. Thus, the bare energies of the dopants are located very close to the band edges and can be approximated by $+\Delta/2$ for the donors and $-\Delta/2$ for the acceptors. To model the nonlinear screening of randomly placed charged impurities

[*] rosch@thp.uni-koeln.de

in such a system we use a simple classical model [6,9,10,14]:

$$H = H_n + H_C = \frac{\Delta}{2}\sum_i f_i n_i + \frac{1}{2}\sum_{i\neq j} V_{ij}\, q_i q_j, \qquad (1)$$

where $f_i = \pm 1$ are random numbers with $f_i = +1$ for a donor states and $f_i = -1$ for the acceptor state at position $\boldsymbol{r}_i$. V_{ij} denotes the Coulomb interaction between the dopants at positions $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$. $n_i \in \{0,1\}$ denotes the electronic occupation of the ith dopant and is determined by minimizing the Hamiltonian. It is related to its charge q_i (in units of $|e|$ where e is the electron charge) by

$$q_i = \frac{f_i + 1}{2} - n_i. \qquad (2)$$

A donor (acceptor) in its ground state is characterized by $f_i = 1$, $n_i = 0$, and $q_i = 1$ ($f_i = -1$, $n_i = 1$, and $q_i = -1$). Somewhat counterintuitively, screening occurs when the Coulomb interaction drives donors or acceptors into a neutral state with $q_i = 0$. Several neutral donor states close by form an electron puddle, while neighboring neutral acceptor states form hole puddles. The Coulomb energy is modeled by

$$V_{ij} = \frac{e^2}{4\pi\varepsilon\varepsilon_0\sqrt{|\boldsymbol{r}_i - \boldsymbol{r}_j|^2 + a_B^2}} = \frac{E_c}{\sqrt{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2 + 1^2}}. \qquad (3)$$

Here the short-distance cutoff $a_B = \frac{4\pi\varepsilon_0\varepsilon}{m^* e^2}$ was introduced by Skinner *et al.* [9,10] to take into account that the wave function of the bound state is smeared over a length scale set by the effective Bohr radius of the impurity state. Skinner *et al.* [9,10] argued that due to the large dielectric constant in Bi-based topological insulators, a_B is large and of similar size as the typical distance of dopants. We use $a_B = N^{-1/3}$ where $N = N_A = N_D$ is the density of dopants where we assume a perfectly compensated system where the density of donors equals the density of acceptors $N_A = N_D$. For the last equality in Eq. (3) we expressed all distance in units of the average dopant distance $N^{-1/3}$. Here

$$E_c = \frac{e^2 N^{1/3}}{4\pi\varepsilon\varepsilon_0} \qquad (4)$$

is the typical energy scale describing the Coulomb interaction of neighboring dopants. A large $\varepsilon \sim 200$ leads to a small energy scale $E_c \sim 3.3$ meV ~ 40 K (assuming a typical density $N = 10^{20}$ cm^{-3}), about 2 orders of magnitude smaller than typical band gaps Δ. Indeed, in Ref. [6] we used the temperature dependence of the optical response to determine E_c and found $\Delta/E_c \approx 150$, similar parameters have also been found in Ref. [12]. In the following we assume $T \ll E_c \ll \Delta$ and consider therefore only properties at $T = 0$.

The model (1) describes how donor and acceptor states interact with each other. It does not include the states in the electronic bands. This turns out [6] to be well justified in the limit $\Delta/E_c \gg 1$ as the density of the relevant electronic states is much smaller than the density of dopants in this limit.

To find the true ground state of the model in Eq. (1) is an exponentially hard problem, but there is an algorithm to find an approximate ground state, called a pseudoground state, in polynomial time [1,9]. The physical properties of a pseudoground state are expected to be indistinguishable from that of the true ground state. The single particle energies are defined as

$$\epsilon_j = \frac{\Delta}{2} f_j - \phi_j = \frac{\Delta}{2} f_j - \sum_{i\neq j} V_{ij}\, q_i. \qquad (5)$$

In a pseudoground state

$$\Delta E_{(\alpha,\beta)} = \epsilon_\beta - \epsilon_\alpha - V_{\alpha\beta} > 0 \qquad (6)$$

has to be fulfilled for all pairs with $n_\beta = 0, n_\alpha = 1$. This state can be reached by exchanging electrons between states where this condition is not met. The algorithm is described in detail in Refs. [9,10] (we use a version where sites α and β are randomly selected). Simulations are performed in a cubic volume $V = L^3$ with periodic boundary conditions with $2L^3$ dopants, typically we use $L = 50$ or $L = 60$ corresponding to 250 000 or 432 000 dopants. Numerical results shown below are averaged over 200–800 disorder realizations, i.e., random configurations of the dopant positions. We have checked [15] that our code reproduces published results (e.g., on the Coulomb gap in the density of states) from other groups [9] on the same model in all quantitative details. In the following we use dimensionless units where all length are measured in units of $N^{-1/3}$, and all energies are measured in units of E_c defined in Eq. (4). In these units the only free parameter of our model is Δ besides the (dimensionless) system sizes considered in Sec. V.

III. LENGTH SCALES AND SCALING

One of the main questions that we will address is the following: What is the typical distance between electron and hole puddles or, equivalently, on what length scale does the potential typically change by an amount of Δ? It turns out that this length scale also characterizes screening properties on average, as discussed in more detail below.

A simple scaling argument by Efros and Shklovskii [16] suggests that the corresponding length scales as $R_g \sim \Delta^2$. The argument is as follows: in a volume of size $V \sim R^3$ there are on average NR^3 positive and negative charges where N is the density of dopants. But these two numbers are not exactly equal, instead the typical charge of the region is (in the uncorrelated state) $Q_R \sim \pm\sqrt{NR^3}$. This implies a typical potential of order $\phi_R \sim Q_R/R \sim \sqrt{R}$ within that region. The fact that this potential diverges for $R \to \infty$ shows that this situation is unstable and the huge potential fluctuations have to be screened. The potential can be screened when the Coulomb potential is sufficiently strong to change the charging state of the dopant. This is possible for $\phi \sim \pm\Delta/2$. Using that $\phi \sim \sqrt{R}$, this strongly suggests that the typical length scale R_g, describing both the screening length and the length scale where the potential changes by $\pm\Delta$, is proportional to Δ^2. Accordingly, the typical charge density in a volume $V = R_g^3$ is $\rho_g \sim Q_{R_g}/V \sim \sqrt{R_g^3}/R_g^3 = R_g^{-3/2} \sim 1/\Delta^3$. To summarize, this scaling argument suggests

$$R_g \sim \Delta^2 \quad \text{and} \quad \rho_g \sim \Delta^{-3}. \qquad (7)$$

Restoring dimensionfull units, these equations read $R_g \sim N^{-1/3}(\Delta/E_c)^2$ and $\rho_g \sim \pm eN(E_c/\Delta)^3$. We will show below that our numerical results for $\Delta/E_c \lesssim 35$ show a much slower

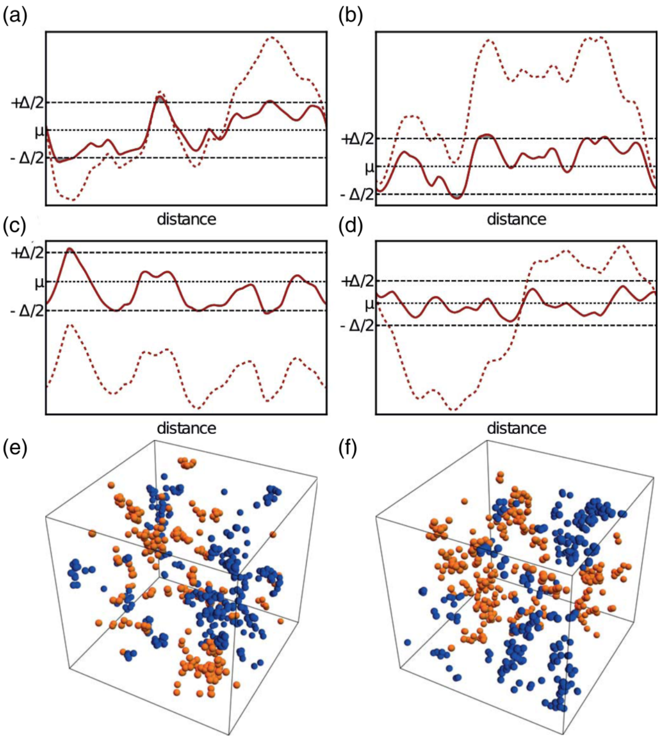

FIG. 1. Due to charged impurities, the potential fluctuates in space. Huge fluctuations of the potential in the uncorrelated state (dashed line), where all dopants are charged, are screened by the formation of electron-hole puddles. The potential $\phi(\boldsymbol{r})$ [solid line, (a)–(d)] obtained in the ground state is restricted to the range $[-\Delta/2 - E_c, \Delta/2 + E_c]$. Puddle formation occurs in tiny regions (gray shading) where $\phi(\boldsymbol{r})$ exceeds the band edges and is thus above $\Delta/2$ or below $-\Delta/2$. (a)–(d) Four one-dimensional cuts through the three-dimensional potential. (e) and (f) Two examples of three-dimensional configurations of electron (orange) and hole (blue) puddles. Only the neutral dopants within the puddles are shown (about 2% of the total number of dopants). The plots suggest that more than one length scale governs puddle formation, see text. Plots are taken from simulations with $\Delta = 10$, periodic boundary conditions, and $L = 50$ (250 000 dopants), the plots show boxes of size 25.

growth of length scales with Δ/E_c. We will attribute this to a huge crossover regime and the presence of logarithmic corrections obtained from a refined version of the scaling argument in Sec. V.

In Fig. 1 we compare several one-dimensional cuts of the potential in the uncorrelated state (dashed lines) and in the correlated ground state (solid lines) obtained from numerical simulations. The potential fluctuations of the uncorrelated state are much larger than $\pm\Delta/2$ (shown here for $L = 50$) triggering screening. For the correlated ground state, in contrast, the potential fluctuations are strongly reduced and lie within the band gap. Puddles are formed in the tiny regions (shaded in gray) where $|\phi|$ slightly exceeds $\Delta/2$. One finds that in these regions $|\phi| - \Delta/2 \sim E_c$.

The 3D plots in Figs. 1(e) and 1(f) show directly the puddles. Neutral donors constitute electron puddles and are colored in orange, while neutral acceptors are part of hole puddles, colored in blue. The snapshots and the cuts suggest that not only a single, but several length scales govern puddle formation [17–19]. The short one ℓ_P governs the closest distance of puddles and rapid fluctuations of the potential as shown in Fig. 1(b). Much longer length scales govern the formation of lengthy, anisotropic cluster structures and also regions without puddles, see Fig. 1(d).

To obtain more quantitative results, one can study the statistical properties of either the potential $\phi(\boldsymbol{r})$ or directly of the charge distribution $\rho(\boldsymbol{r})$, since both are related by the Poisson equation $\nabla^2\phi = -\rho$ (up to the short-distance cutoff a_B introduced above). In the following, we will mainly discuss the charge-charge correlation function $C_{\rho\rho}$. We split this into a local part $\sim\delta(\boldsymbol{r} - \boldsymbol{r}')$ and a nonlocal part $C^{\mathrm{nl}}_{\rho\rho}$:

$$C_{\rho\rho}(\boldsymbol{r},\boldsymbol{r}') = \langle\rho(\boldsymbol{r})\rho(\boldsymbol{r}')\rangle = Q_0\delta(\boldsymbol{r} - \boldsymbol{r}') + C^{\mathrm{nl}}_{\rho\rho}(\boldsymbol{r} - \boldsymbol{r}'), \quad (8)$$

where we used charge neutrality $\langle\rho\rangle = 0$. Here and in the following the expectation value $\langle\cdot\rangle$ denotes a disorder average. After disorder averaging all correlation functions only depend on the distance $r = |\boldsymbol{r} - \boldsymbol{r}'|$. Thanks to charge neutrality we know that $\int d^3\boldsymbol{r}\, C^{\mathrm{nl}}_{\rho\rho}(\boldsymbol{r}) = -Q_0$. The weight of the δ-peak Q_0 corresponds to $2N(1 - n_0)$ where n_0 is the fraction of neutral dopants.

IV. SCREENING IN THE BULK

Screening in insulating charged Coulomb systems is a highly nonlocal and nonlinear mechanism. Early work by Baranovskii, Shklovskii, and Efros [17] (see also a lucid discussion in Ref. [18]) pointed out that adding a single charge can trigger an avalanche of discrete changes of the charge of dopants not only in the neighborhood of the charge but also at large distances. This appears to be a highly anisotropic, nonlocal (and perhaps fractal [19]) process. The change of the potential at larger distances is random in sign but does not decay rapidly. In contrast to a metal, there is therefore no true screening (as is also obvious from the fact that the system is characterized by a Coulomb gap). In the following we will not track these changes but focus on the shorter length scale ℓ_P which governs the impurity-averaged charge correlations $C^{\mathrm{nl}}_{\rho\rho}(\Delta,s)$, but also controls the typical "nearest" distance of oppositely charged puddles. Later we will argue that the same length scale also governs the impact of metallic surface states on puddle formation.

Instead of studying directly the charge-charge correlation function $C^{\mathrm{nl}}_{\rho\rho}(\Delta,s)$, we find it more convenient to investigate the distance dependence of the "screening charge" defined by

$$Q_s(\Delta,r) = 4\pi \int_0^r ds\, s^2\, C^{\mathrm{nl}}_{\rho\rho}(\Delta,s). \quad (9)$$

The advantage of this quantity is that it has a direct physical interpretation: it describes the charge accumulated—on average—around a dopant within the radius r multiplied with the charge of that dopant and the density of dopants. As negative charges accumulate around a positive charge and vice versa, the screening charge is always negative. Total charge neutrality requires that around a positive (negative) charge exactly the charge -1 $(+1)$ accumulates for $r \to \infty$. As neutral dopants do not contribute, one therefore obtains $Q_s(\Delta, r \to \infty) = -2N(1 - n_0) = -Q_0$. This also follows directly by integrating Eq. (8) over $\boldsymbol{r}$ in a charge-neutral system. In our simulations we use boxes of size L with

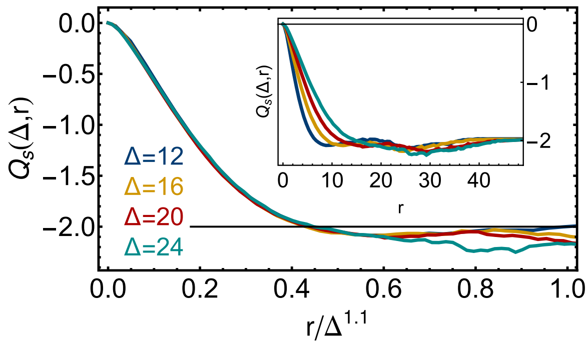

FIG. 2. Apparent scaling of the screening charge defined in Eq. (9) for different values of the band gap Δ. The best scaling collapse is found for an exponent $\gamma = 1.1$ characterizing an extended crossover regime. The inset shows the unscaled data. Deviations from the scaling behavior can be seen for $r > 0.6\,\Delta^{1.1}$. Parameters are $L = 50$ (250 000 dopants) for $\Delta = 12,16$ and $L = 60$ (432 000 dopants) for $\Delta = 20,24$, and we checked that there are no significant finite size effects.

periodic boundary conditions. For $r > L/2$ we therefore have to replace in the integral in Eq. (9) the factor $4\pi s^2$ by $W(s) = \int \delta(s - |r|)d^3r$. This does not affect the scaling plots discussed below but is useful to check for overall charge neutrality.

We show numerical results for Q_s in Fig. 2. On a rather short length scale (see inset) the screening charge reaches the value $-Q_0 = -2N(1-n_0) \approx -2$ (the plot uses units where $N = 1$ and the fraction of neutral dopants n_0 is less than 2% for all shown values of Δ). The scaling plot (main figure) suggests that the length scale ℓ_P, on which the screening occurs, grows almost linear in Δ in the numerically accessible regime

$$\ell_P \sim \Delta^{\gamma}, \quad \gamma \approx 1.1 \pm 0.1. \tag{10}$$

Below we will argue that the exponent γ is not a "true" asymptotic exponent but only an effective parameter describing an extended crossover regime. For values of $r \gtrsim 0.4\Delta^{\gamma}$, the screening charge exceeds $-Q_0 \approx -2$. This implies that there is a substantial amount of overscreening in the system: on average too much charge of opposite sign accumulates around each charged dopant.

We have checked that other observables, for example the potential correlation function or the typical distance of neutral dopants of different type, show similar scaling behaviors, see Appendix B for an example. Most importantly, they all consistently show the importance of the length scale ℓ_P which governs not only screening but also the length scale on which the dominant short-distance fluctuations of the potential occur. ℓ_P therefore also determines the distance of puddles of opposite charge.

V. SCREENING BY METALLIC SURFACE STATES

Topological insulators differ from ordinary insulators or semiconductors because topology enforces the existence of conducting surface states. These states are of interest in the context of our discussion, because they provide an extra channel for screening. STM measurements of surface states can also be used to obtain quantitative information on the strength and length scale of potential fluctuation at the surface [13]. Most importantly, the suppression of puddles in thin slabs of topological insulators is expected to lead to a substantial reduction of the bulk conductivity and should therefore enhance the quality of devices based on topological insulators substantially. A major goal of this section is therefore to estimate how thin a topological insulator has to be so that puddle formation is effectively suppressed. Note that such a suppression can occur even in the absence of metallic surface states, as has been discussed heuristically by Mitin [20] for semiconductor heterostructures.

The surface states of a 3D topological insulator can, generically, be described by a Dirac equation, and thus have asymptotically a density of states proportional to the doping level. Their electronic properties can be characterized by the surface doping μ^S and the effective fine structure constant $\alpha = \frac{e^2}{4\pi\varepsilon_{\text{surf}}\varepsilon_0\hbar v_F}$, where, in vacuum, $\varepsilon_{\text{surf}} = \frac{\varepsilon_{\text{bulk}}+1}{2} \sim 100$. Typical values for α in Bi-based topological insulators are in the range of $\alpha \approx 0.1,\ldots,0.2$ (using, e.g., v_F taken from ARPES data [21]). In Ref. [11], Skinner, Chen, and Shklovskii develop a detailed analytic theory on how bulk impurity states affect the surface. We will instead investigate the question how the screening from surface states feeds back on bulk properties using some of their results.

If the surface possesses a finite doping, described by a finite chemical surface potential μ^S, it can screen charges on a length scale described by the surface screening length $\ell_s^S \sim v_F/(\alpha|\mu^S|)$. We first consider the limit that ℓ_s^S is smaller than the distance of bulk impurities $\ell_s^S \lesssim N^{-1/3}$ or, equivalently, $|\mu^S| \gtrsim E_c/\alpha^2$. In this case the surface state of the topological insulator acts effectively like a perfect metal. Then, screening of a dopant with charge q_i at distance z from the surface is described by positioning a mirror charge with charge $-q_i$ at the same distance on the opposite side of the surface. This simple screening mechanism can be implemented in a straightforward way into the model described in Sec. II. To model a thin slab of a topological insulator with two metallic surface states, one formally needs an infinite sequence of mirror charges. As described in Appendix D, an accurate and numerically efficient description is obtained by using just a single mirror charge and a linear correction term setting the potential to zero at both surfaces. Besides its importance for applications of TI materials, the problem of an infinitely large TI slab of finite thickness L_z has also a technical advantage which we will use in the following: the "bare" potential $\Phi_0(\boldsymbol{r})$ arising from randomly placed impurities remains finite for finite L_z even in the thermodynamic limit (while it would diverge in the absence of surface screening). This is the potential one obtains in the absence of puddle formation when all donors (acceptors) have charge $+1$ (-1). One can easily calculate the distribution $p[\Phi_0(z)]$ of this potential by averaging over the position of dopants, here z is the coordinate perpendicular to the surfaces. In the following we will focus our discussion for simplicity on the distribution in the middle of the sample $z = L_z/2$. Due to the central limit theorem, this initial distribution (before puddle formation) is Gaussian

$$p^0_{L_z/2}(\Phi) = \frac{1}{\sigma(L_z)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\Phi}{\sigma(L_z)}\right)^2\right], \tag{11}$$

where the width of the distribution $\sigma(L_z)$ can simply be computed from $\sqrt{\langle \Phi_0(z=L_z/2)^2\rangle}$ and, in units of E_c, is given by

$$\begin{aligned}\sigma(L_z) &= \left(2\int_0^{L_z} dz \int_{-\infty}^{\infty} dx\,dy\;V^s[\boldsymbol{x},(0,0,L_z/2)]^2\right)^{1/2}\\ &\approx 2.41\sqrt{L_z} - \frac{7.75}{\sqrt{L_z}} \quad \text{for } L_z \gg 1,\end{aligned} \tag{12}$$

where the potential V^s is defined in Appendix D and the second line is a fit to the numerical integral valid for large L_z. The prefactor of the leading term depends only on the geometry of the setup but is otherwise universal, the subleading term is linear in the chosen cutoff a_B defined in Eq. (3). For small L_z, the width $\sigma(L_z)$ is much smaller than the gap, implying that puddle formation cannot take place. As discussed in the Introduction, Efros and Shklovskii [16] estimated the length scale triggering puddle formation (for a different geometry) from the condition $\sigma(L_z) \sim \Delta$, leading to a length scale proportional to Δ^2 as discussed in Eq. (7). We will need in the following a more quantitative version of this argument. We will take into account, that for the successful screening of the potential in the limit $L_z \to \infty$, it is not necessary to redistribute $O(1)$ charges. Instead we can use the total density of neutral dopants in the thermodynamic limit n_0 as an estimate of the volume fraction where the bare potential triggers puddle formation by becoming larger than $\Delta/2$. The characteristic width of the slab L_c^0 below which puddle formation is suppressed, is therefore estimated from the condition

$$p_0(|\Phi| > \Delta/2) \sim n_0, \tag{13}$$

with

$$p_0(|\Phi| > \Delta/2) = 2\int_{\Delta/2}^{\infty} p^0_{L_z/2}(\Phi)\,d\Phi, \tag{14}$$

solved by

$$\sigma(L_c^0)^2 \sim \frac{\Delta^2}{8\ln\left[1/\left(n_0\sqrt{\pi\ln\left[2/\left(\pi n_0^2\right)\right]/2}\right)\right]} \tag{15}$$

for small n_0. Using Eq. (12), we find for $\Delta \to \infty$ as an estimate for the characteristic width L_c^0,

$$L_c^0 \sim \frac{\Delta^2}{46.5\ln[1/n_0]} \approx \frac{\Delta^2}{139\ln[\Delta]}, \tag{16}$$

with (sizable) relative corrections of the order of $\ln[\ln\Delta]/\log[\Delta]$. In the last line we use that in the asymptotic regime $n_0 \sim 1/\Delta^3$, see Eq. (7). This formula misses logarithmic corrections to n_0 which give however only subleading terms beyond the precision of Eq. (16). While our equation can only be an order-of-magnitude estimate, we have kept multiplicative numerical prefactors to indicate their rather large numerical value. Note that Eq. (13) and therefore also (16) was obtained only by considering properties of the bare potential [by evaluating the integral in Eq. (12)], not including any self-consistent screening effects. The formulas can therefore only be viewed as a crude estimate of the relevant length scale of the problem obtained by extrapolating from the bare potential. The result clearly suggests the presence of logarithmic corrections to scaling but we cannot exclude that a resummation of logarithms leads to a modification of the logarithm, e.g., $L_c \sim \Delta^2/\ln^\alpha[\Delta]$. A scaling $L_c \sim \Delta^\gamma$ for $\Delta \to \infty$ with an exponent γ smaller than 1 can, however, be excluded as in this case the regions where the bare potential can trigger puddle formation are exponentially suppressed. In the following we will compare the estimate from condition Eq. (13) to the full numerical solution obtained for moderately large values of $\Delta \lesssim 35$ and find that the formula nevertheless reproduces the approximately linear Δ dependence ($\gamma \approx 1.1 \pm 0.1$) in the regime $\Delta \lesssim 35$.

The surface screening will suppress potential fluctuations and the formation of puddles close to the surface. Therefore, all donors will have charge +1, all acceptors have charge −1, and the density of neutral dopants vanishes close to the surface. This physics can be captured by computing the density of neutral dopants, having a charge 0, as a function of distance from the surface

$$n_0(d) = \left\langle \sum_i \delta(d - z_i)\delta_{q_i,0}\right\rangle, \tag{17}$$

where z_i is the (dimensionless) distance of dopant i from the surface and $\delta_{i,j}$ denotes the Kronecker delta. In contrast to charge and potential, the density of neutral dopants is a quantity not fluctuating in sign and therefore its average is both easier to compute (statistical fluctuations are much weaker) and to interpret. In Fig. 3 (upper panel) we show $n_0(d)$ for different values of L_z. As expected, the puddle formation and therefore $n_0(d)$ is suppressed close to the two metallic surfaces. For sufficiently thin L_z, $n_0(d)$ becomes small even in the center of the sample. The lower panel of Fig. 3 therefore shows the density of neutral dopants in the center of the slab $n_0(L_z/2)$, a function of L_z. In Fig. 4 we show four different length scales, ℓ_P, ℓ_c, ℓ_s, and L_c^0, as function of Δ. The first three length scales have been extracted from our numerics and the equations

$$Q_2(\Delta,\ell_P) = -1, \tag{18}$$

$$n_0(L_z/2) = \tfrac{1}{2}n_0^{\text{bulk}} \quad \text{for } L_z = \ell_c, \tag{19}$$

$$n_0(\ell_s) = \tfrac{1}{2}n_0^{\text{bulk}} \quad \text{for } L_z \gg \ell_c. \tag{20}$$

They describe the characteristic length scale ℓ_P on which—on average—a charge is screened in the bulk (see Fig. 2), the characteristic width ℓ_c of a slab of a topological insulator below which the density of puddles drops to half the bulk value, and the length scale on which puddle formation is suppressed close to the metallic surface of a thick slab of a topological insulator (see Appendix E). All three curves show an approximately *linear* behavior with Δ quite different from the $\ell \sim \Delta^2$ (with logarithmic corrections) expected from scaling arguments. All curves are well described by fits of the form $\ell_i = a_i + c_i\Delta^{1.1}$. Remarkably, the same behavior is also obtained from the estimate L_c^0, which was obtained from the condition in Eq. (13), i.e., from properties of the bare potential (before puddle formation). Therefore ℓ_i is well described by a linear fit to L_c^0 as shown by the solid lines in Fig. 4. The dashed red line in Fig. 4 shows a power-law fit $L_c^0 \sim \Delta^{1.1} + \text{const.}$, which works remarkably well. We therefore conclude that (i) the average bulk screening and the surface screening are

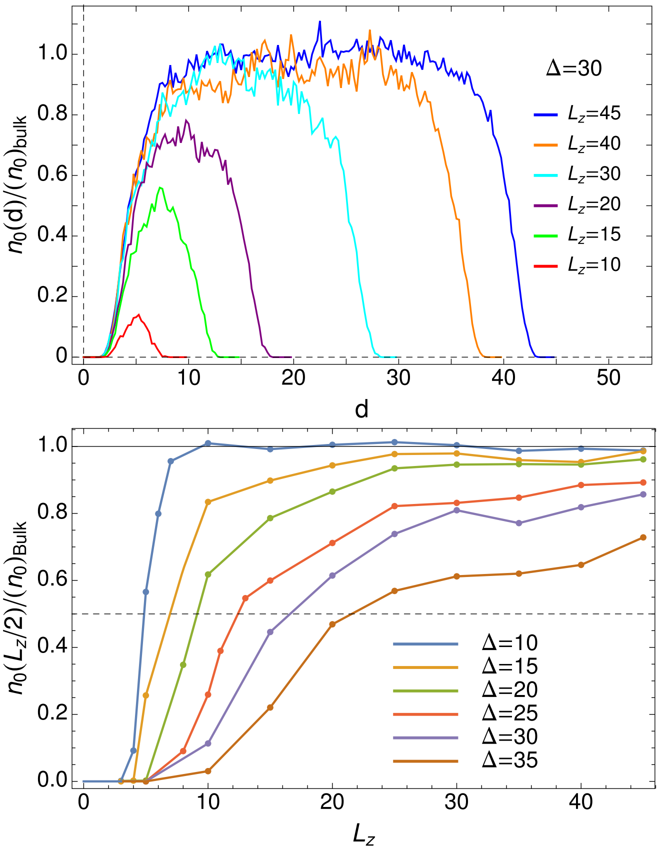

FIG. 3. Density of neutral dopants $n_0(d)$ in a slab of a topological insulator of width L_z, where d is the distance from one of the surfaces ($\Delta = 30$). The metallic surface states screen the potential, thus suppressing $n_0(d)$. For thick samples, $L_z \gtrsim 30$, $n_0(d)$ is only suppressed close to the boundaries while in the center one recovers the bulk puddle density. For $L_z \lesssim 20$, $n_0(d)$ also drops in the middle of the sample (simulations for $L_{x,y} = 50$, $2 \cdot L_x L_y L_z = 5000 L_z$ dopants, averaged over 300 disorder configurations). The lower panel show the density of defects in the center $n_0(L_z/2)$ as a function of L_z for various values of Δ.

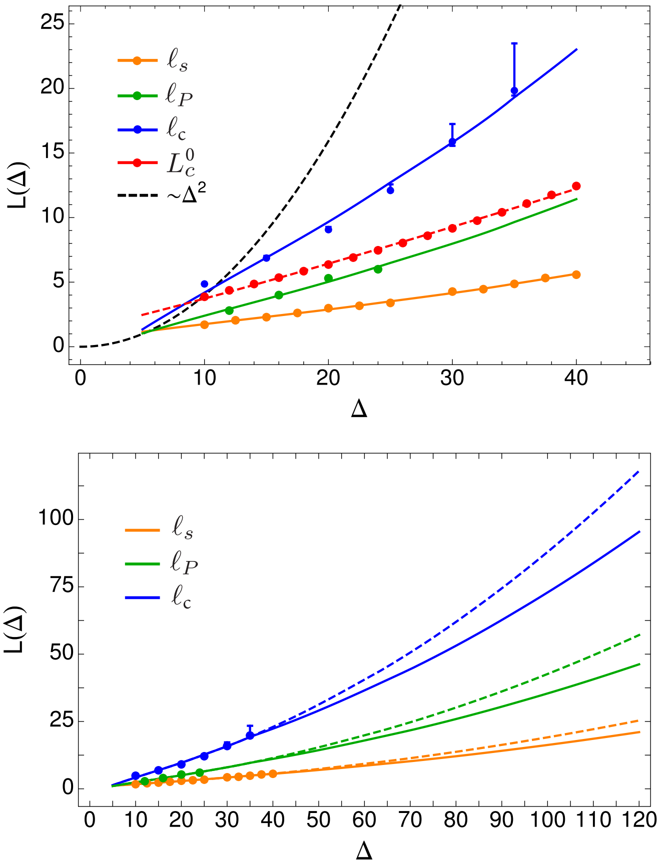

FIG. 4. Δ dependence of four different length scales ℓ_P, ℓ_c, ℓ_s, and L_c^0. Three of them, ℓ_P, ℓ_c, ℓ_s, have been obtained numerically, see Eqs. (18)–(20). ℓ_P characterizes the screening of charges (on average) in the bulk, ℓ_c is the suppression of puddles in a thin slab of a topological insulator, and ℓ_s is the suppression of puddles close to a metallic surface. The error in ℓ_c arises from the error in the determination of n_0 in the thermodynamics limit, see Appendix C. L_c^0 is an analytic order-of-magnitude estimate for l_c based on Eq. (13) and Eqs. (11) and (12) (using numerically determined values for n_0). The black dashed line in the upper panel is the curve $\Delta^2/(8\pi)$, see Ref. [10], which shows that all length scales rise much slower than Δ^2, the red dashed line is a power-law fit $1.36 + 0.19\Delta^{1.1}$ to L_c^0. To extrapolate to larger values of Δ, we use a linear fit of ℓ_i to L_c^0, $\ell_i = a_i' + c_i' L_c^0$ with $a_i' = -1.67, \ -4.29, \ -0.07$ and $c_i' = 1.05,\ 2.19,\ 0.46$ for $i = P,c,s$, respectively (solid lines in both panels). For $\Delta > 35$ (lower panel), L_c^0 was determined assuming $n_0(\Delta) = n_0(35)(35/\Delta)^3$ (solid lines), see text. The dashed lines, calculated from $n_0(\Delta) = n_0(35)(35/\Delta)^{1.62}$, are shown to indicate how sensitive the result is to a different extrapolation of n_0.

governed by the same length scale, and that (ii) one can use the "naive" scaling argument (13) to obtain this length scales [up to multiplicative factors of $O(1)$ and a small offset of $O(1)$]. As we have shown above, the asymptotic behavior for L_c^0 is according to Eq. (16) given by $\Delta^2/\ln[\Delta]$ and definitely not by an exponent close to 1. The apparent power-law behavior with an exponent close to one therefore reflects only an extended crossover regime: there is no "true" power law with an exponent smaller than 2 (see, e.g., Ref. [22] for a discussion on the definition and determination of exponents). The approximately linear behavior for $10 \lesssim \Delta \lesssim 50$ arises from the interplay of logarithmic corrections at large Δ and subleading corrections for small Δ, see Eq. (12). As also the numerically determined length scales ℓ_i show the same behavior, we conclude that also in this case the numerics probes the same extended crossover regime for numerically accessible vales of $\Delta \lesssim 35$. Below, we will argue that one can use the results for L_c^0 to estimate the value of ℓ_i for larger values of $\Delta \sim 100$, relevant for Bi-based topological insulators.

Above we assumed a perfectly metallic surface state $|\mu^S| \gtrsim E_c/\alpha^2$, which has a screening length that is short compared to the mean distance of dopants $N^{-1/3}$. Using the result given above, that the screening by bulk states sets in only at a parametrically larger scale ℓ_s, we can relax this requirement. Our results should be valid as long as the surface screening length $\ell_S^S \sim v_F/(\alpha|\mu^S|)$ is small compared to ℓ_s, or $|\mu^S| \gg v_F/(\alpha \ell_s)$.

Here μ^S denotes an effective chemical surface potential. Even if the chemical potential of the surface state is *exactly* at the Dirac point $\langle \mu^S \rangle = 0$, disorder will induce a finite density of states allowing for screening. Due to the charged dopants metallic puddles will form on the surface which can, in turn, screen bulk charges. To estimate the effect of these surface puddles (not to be confused with puddles in the bulk) we use the results of Ref. [11] (similar results in the context of graphene have, e.g., been obtained in Refs. [2,3]). For the computation of the resulting surface screening length ℓ_S^S for $\langle\mu^S\rangle = 0$ the authors of Ref. [11] did not take into account any bulk-screening effects, which is justified as long as $\ell_S^S \ll \ell_s$. Under these conditions, Skinner, Chen, and Shklovskii [11] found that $|\mu^S| \sim E_c/\alpha^{2/3}$ or $\ell_S^S \sim N^{-1/3}/\alpha^{4/3}$. From the condition $\ell_S^S \ll \ell_s$, we obtain (using our dimensionless units)

$$\ell_s > c\left(\frac{1}{\alpha}\right)^{4/3} \quad \text{for } \langle\mu^S\rangle = 0, \tag{21}$$

where $c \approx 0.6$ according to Ref. [11], where the authors estimate $\alpha \approx 0.24$ for a Bi-based topological insulator, which results in the condition $\ell_s > 4$ for this class of systems.

If the condition (21) is fulfilled, the surface state of a topological insulator provides sufficient screening to suppress efficiently the formation of puddles within the distance ℓ_s.

VI. DISCUSSION AND OUTLOOK

We have investigated the influence of charge dopants in (topological) insulators, focusing on the case of perfect compensation with equal densities of donors and acceptors. Motivated by the physics of Bi-based topological insulators, we studied the limit where the gap Δ is large compared to the Coulomb energy E_c of neighboring dopants with $\Delta/E_c \sim 100$ as a typical value [6]. In our numerical simulations we are not able to reach such large values of Δ. Therefore, analytical estimates are needed to extrapolate to larger values of Δ. Our main focus has been the investigation of the length scales governing the formation (and destruction) of puddles. As has been pointed out before in the literature [17,18], due to the long-ranged nature of the Coulomb interaction and the highly nonlinear screening effects, there is more than one such length scale. We have, however, found that the size of a (average) screening cloud around an impurity, the typical distance of electron and hole puddles, and most importantly the length scales governing the suppression of puddle formation in the bulk due to metallic surface states, are all similar and show an approximately linear increase with Δ for $\Delta \lesssim 35$ (a fit gives $\ell_i \sim \Delta^{1.1}$). We have found a simple analytic estimate of such length scales based on properties of the bare potential, which reproduces this behavior in an extended crossover regime but predicts $\ell_i \sim \Delta^2/\ln[\Delta]$ for $\Delta \to \infty$. One can use this analytic estimate to obtain a quantitative extrapolation of the numerically determined results to larger values of Δ. This is shown in the lower panel of Fig. 4. The fit $\ell_i = a_i' + c_i' L_c^0$ gives an excellent fit to the numerically determined length scales ℓ_i (see figure caption for details and fit parameters). Using this extrapolation, we can estimate the corresponding length scale for large values of Δ. Assuming, for example, $\Delta/E_c \sim 100$ and $N \approx 10^{19}$ cm^{-3}, our best estimates for the dimensionless length scales are $\ell_c \approx 72.9 \pm 15.0$, $\ell_s \approx 16.3 \pm 2.8$ where errors have been estimated based on the use of different extrapolations of n_0, see Fig. 4. In physical units this implies that the width of the region close to the metallic surface where puddle formation is inhibited is about 62.6–88.6 nm. Puddle formation in the center of a thin slab of a topological insulator is predicted to be suppressed by metallic surface states by at least a factor 2 if the slab is thinner than 268.7–407.9 nm. As puddles largely control the bulk conduction at low temperatures by reducing the energy gap for transport processes [1,9,10], the suppression of puddle formation in the bulk is expected to be accompanied by a strong suppression of bulk conduction. More precisely, at least three effects will contribute to the increase of bulk resistivity the screening from metallic surfaces, the suppression of Coulomb fluctuations due to the dimensional crossover (even without metallic surfaces), and the crossover from a 3D to a 2D percolation problem of electrons moving in a correlated potential [1,20]. As we have shown that surface and bulk effects are governed by similar length scales proportional to each other, we expect that all effects are of similar importance. In compensation doped Bi-based compounds the bulk conductance is expected to be suppressed considerably (i.e., much faster than to be expected from geometric factors) when the slab becomes thinner than, e.g., 270 nm. It will be interesting to develop a quantitative theory for transport in the future which combines numerical calculations for smaller values of Δ/E_c with analytic extrapolation schemes for large Δ similar to the ones used in this paper.

ACKNOWLEDGMENTS

We thank Y. Ando, O. Breuning, N. Borgwardt, M. Grüninger, B. Skinner, and, especially, B. Shklovskii for insightful discussions and useful comments. The numerical simulations have been performed on the CHEOPS cluster at RRZK Cologne funded by the DFG. We also thank the DFG for financial support within Project No. C02 of CRC 1238.

APPENDIX A: SUM RULES AND SCALING ARGUMENTS

From the definition of $C_{\rho\rho}^{\mathrm{nl}}$, Eq. (8), and the Poisson equation one can derive a set of exact sum rules

$$\frac{\langle H_{\mathrm{C}}\rangle}{V} = 2\pi \int ds\, s^1\, C_{\rho\rho}^{\mathrm{nl}}(\Delta,s), \tag{A1}$$

$$\frac{\langle H_{\mathrm{n}}\rangle}{V} = n_0 N\Delta = 2\pi\,\Delta \int ds\, s^2\, C_{\rho\rho}^{\mathrm{nl}}(\Delta,s) + \Delta, \tag{A2}$$

$$Q_0 = \frac{\langle Q\rangle}{V} = -4\pi \int ds\, s^2\, C_{\rho\rho}^{\mathrm{nl}}(\Delta,s), \tag{A3}$$

$$\langle\phi^2\rangle = -8\pi^2 \int ds\, s^3\, C_{\rho\rho}^{\mathrm{nl}}(\Delta,s). \tag{A4}$$

Here $\langle H_{\mathrm{C}}\rangle$ is the disorder average of the Coulomb energy, $\langle H_{\mathrm{n}}\rangle$ are the single particle energies of the dopants, see Eq. (1), $\langle Q\rangle$ is the number of ionized dopants (not counting the neutral ones), and $\langle\phi^2\rangle$ is the expectation value of the square of the potential (all expressed in our dimensionless units).

We can use these sum rules to obtain a more rigorous version of the scaling argument given above. We start from the assumption that the physics of the system is governed by a single length scale large compared to the average distance of impurities. In this case $C^{\mathrm{nl}}_{\rho\rho}(\Delta,s)$ can be written as

$$C^{\mathrm{nl}}_{\rho\rho}(\Delta,s) = \Delta^{-\beta}\,\overline{C^{\mathrm{nl}}_{\rho\rho}}(s/\Delta^{\gamma}). \tag{A5}$$

We will show that from this assumption alone Eq. (7) can be derived. Later, we will conclude that the scaling ansatz is *not* fully valid: there are substantial subleading corrections even for values of $\Delta \lesssim 35$ and logarithmic corrections in the $\Delta \to \infty$ limit. In this Appendix, we will, however, explore only the consequences of the scaling ansatz.

The bulk fluctuations of the potential are of the order of the band gap Δ, see Fig. 1, which implies $\langle\phi^2\rangle \sim \Delta^2$. Furthermore, as the fraction of neutral atoms vanishes for large Δ, $n_0 \to 0$ for $\Delta/E_c \to \infty$, the density of charged dopants Q_0 is of order Δ^0 with only subleading corrections. To leading order we therefore obtain from Eqs. (A3) and (A4)

$$\Delta^0 \sim \int ds\, s^2\, C^{\mathrm{nl}}_{\rho\rho}(\Delta,s) = \Delta^{-\beta+3\gamma}\int ds\, s^2\, \overline{C^{\mathrm{nl}}_{\rho\rho}}(s), \tag{A6}$$

$$\Delta^2 \sim \int ds\, s^3\, C^{\mathrm{nl}}_{\rho\rho}(\Delta,s) = \Delta^{-\beta+4\gamma}\int ds\, s^3\, \overline{C^{\mathrm{nl}}_{\rho\rho}}(s). \tag{A7}$$

Therefore, the scaling ansatz predicts $\beta = 3\gamma$ and $2 = -\beta + 4\gamma$ or, equivalently, $\gamma = 2$ and $\beta = 6$, implying a length scale $\sim\Delta^2$ and a typical charge density $\sim 1/\Delta^3$. This is just a refined version of the argument presented above in Eq. (7).

From Eqs. (A1) and (A6) we further deduce that the Coulomb energy density $\langle H_{\mathrm{C}}\rangle/V \sim -\Delta^{-\beta+2\gamma} = -\Delta^{-\gamma}$($C^{\mathrm{nl}}_{\rho\rho}$ is negative as it describes screening, for example the accumulation of negative charge around a positive one). This implies that the Coulomb energy is minimized by choosing the screening length $R_s \sim \Delta^{\gamma}$ as small as possible. However, this minimization competes with the increasing $H_{\mathrm{n}} \sim n_0 N \Delta$, see Eq. (A2), and of course has to respect the constraints, in particular Eq. (A7).

Our numerical results are in strong disagreement with the scaling result. Several factors play a role: First, even for $\Delta \sim 35$ the asymptotic scaling regime is not yet reached. It can be seen from Fig. 4, indicating that both ℓ_s and ℓ_P are well below 10 in this regime. Second, a more quantitative estimate based on properties of the bare potential strongly suggested the presence of logarithmic corrections, see Eq. (16). This may indicate that in Eq. (A7), $\int ds\, s^3\, C^{\mathrm{nl}}_{\rho\rho}(\Delta,s)$ obtains logarithmic corrections from a slow decay $\sim 1/s^4$ of $C^{\mathrm{nl}}_{\rho\rho}(\Delta,s)$.

APPENDIX B: CORRELATIONS OF THE POTENTIAL

As show in Fig. 1, the potential in the bulk of a compensation-doped insulator fluctuates in space. It is approximately restricted to the range $[-\Delta/2,\Delta/2]$ and exceeds $\pm\Delta/2$ by an amount of order E_c only in the region where puddles form. The correlation function $\langle\phi(r)\phi(0)\rangle$ shows on which length scale the characteristic potential fluctuations occur.

In Fig. 5 we show $\langle\phi(r)\phi(0)\rangle$ normalized to $\langle\phi(0)\phi(0)\rangle \sim \Delta^2$. At short distances (of the order of the distance of impurities), this is governed by the autocorrelation of the potential of a single charge and decays on a length scale set by a_B. As can be seen in the upper panel of Fig. 5, the normalized correlation function is independent of Δ in this regime. The next-largest length scale, the screening length, on which the correlations decay is more interesting. As expected, we find that this length scale is governed by the bulk screening length $\ell_P \sim \Delta^{\gamma}$, see Eq. (10). This is shown by the scaling plot in the lower panel of Fig. 5. Clearly the same length scale $\approx 0.2N^{-1/3}(\Delta/E_c)^{1.1}$ (including prefactors) determines the screening radius and the dominant length scale of potential fluctuations. Note that scaling does not hold at the short distances ($\lesssim a_B$ and/or impurity distance $N^{-1/3}$) and that we had to subtract a short distance cutoff to obtain a reasonable scaling collapse.

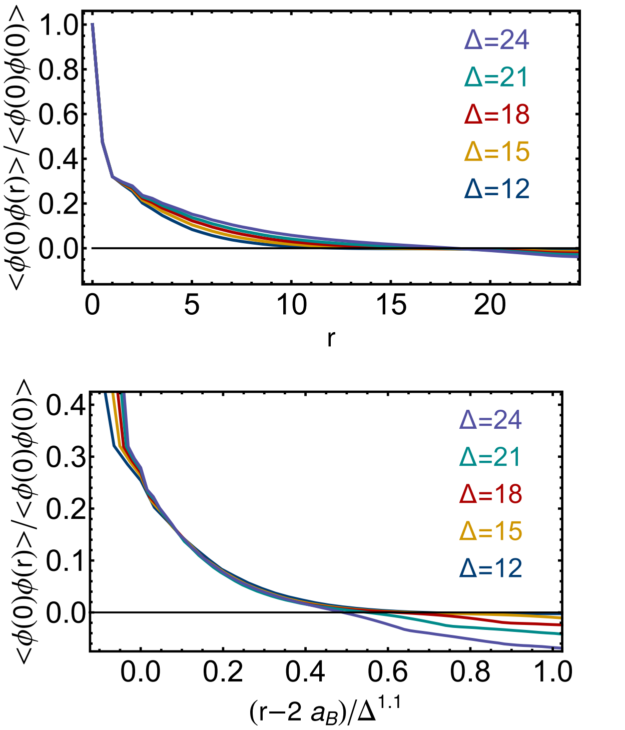

FIG. 5. The potential correlation function $\langle\phi(\boldsymbol{r})\phi(0)\rangle$ normalized to $\langle\phi(0)\phi(0)\rangle$ allows us to determine the length scale of fluctuations of the potential. Upper panel: Unscaled data. Lower panel: Scaling plot for $\Delta = 12,\ldots,24$. For the scaling of the horizontal axis, we first subtract a short-distance cutoff (see text) and then use the scaling of the bulk screening length $\ell_P \sim \Delta^{\gamma}$ where $\gamma \approx 1.1$, see Eq. (10), see text. Scaling breaks down both at short distances of the order of the cutoff and for larger distances, likely related to a second, longer length scale related to overscreening.

At larger length scales, the correlation function becomes negative. This physics is, however, *not* governed by ℓ_P as follows from the absence of a scaling collapse in this regime. As discussed in the main text, the physics in the second regime is related to overscreening and occurs on a length scale which we cannot resolve with our numerical simulations.

APPENDIX C: DENSITY OF NEUTRAL DOPANTS

To determine the density of neutral dopants in the thermodynamic limit, we have simulated boxes of size L^3 with periodic boundary conditions. The resulting density for $\Delta = 25$ is shown in the upper panel of Fig. 6 as function of $1/L$. As we do

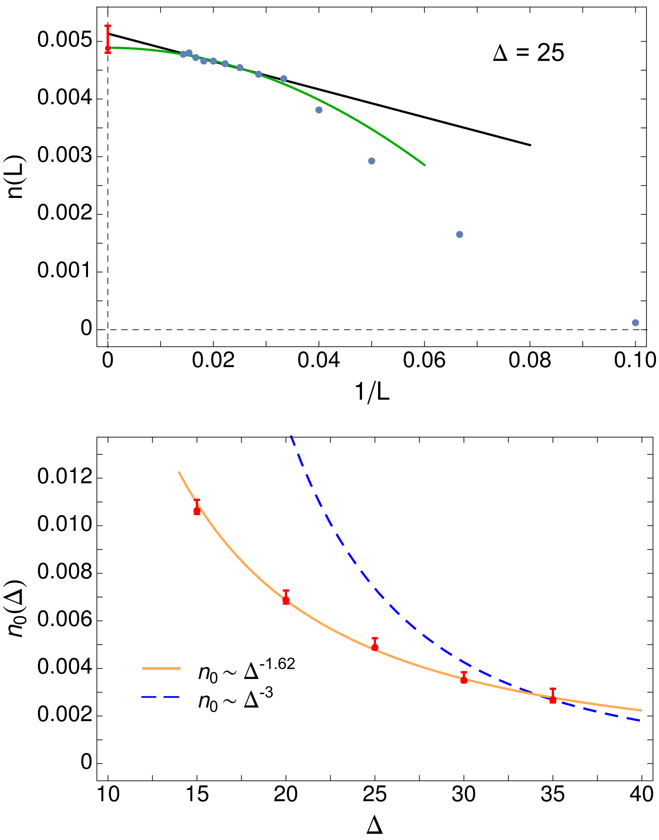

FIG. 6. Upper panel: Density of neutral dopants obtained from simulation of boxes of size L^3 with periodic boundary conditions. Largest simulations include $2 \times 70^3 \sim 700\,000$ dopants. To extrapolate to $L \rightarrow \infty$ we use both a quadratic (green) and linear (black) extrapolation schemes, assuming finite size errors of order $1/L^2$ and $1/L$, respectively. Error bars (red) are determined by combining the 1σ error intervals of both schemes. Lower panel: Δ dependence of the density of neutral dopants in the thermodynamic limit. For the range of Δ accessible to our numerics, the data is consistent with $1/\Delta^{1.6}$ (orange line) reflecting an extended crossover regime. A $1/\Delta^3$ dependence (dashed line) does not fit the data for $\Delta \lesssim 35$.

not know the analytic $1/L$ dependence of that quantity and as the numerical result is consistent with different interpolating functions, we use the following procedure. To estimate the density in the thermodynamics limit, we use a quadratic extrapolation scheme (green line), assuming that finite size effects are of order $1/L^2$. Within the statistical error bars this is approximately equivalent to the value obtained for the largest system size used in our numerics. A linear extrapolation in $1/L$ (black line) gives a higher value for $n_0(L \rightarrow \infty)$. The error bar is obtained from the largest and smallest one-standard-deviation values obtained from both interpolation schemes. It therefore reflects not only the statistical uncertainty of our data but also the much larger systematic error related to the unknown L dependence of finite size effects.

The lower panel of Fig. 6 shows how the density of neutral dopants drops for increasing Δ. In the crossover regime accessible to our numerics, $n_0(\Delta)$ decays much slower than the $1/\Delta^3$ law expected up to logarithmic corrections from the scaling arguments, see Eq. (7). This slow decay is, however, consistent with the slow rise of the length scales characterizing screening, which increase approximately linear instead of quadratically with Δ in the same parameter regime.

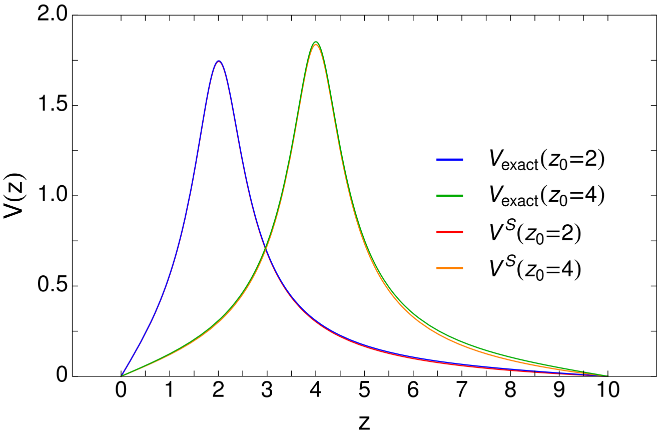

FIG. 7. Potential of a charged impurity located at $z = 2$ and at $z = 4$ in the presence of two metallic surfaces at $z = 0$ and $z = 10$. The plot compares the exact result (upper blue and green curves) to the approximation given by Eq. (D2) (lower red and orange curves).

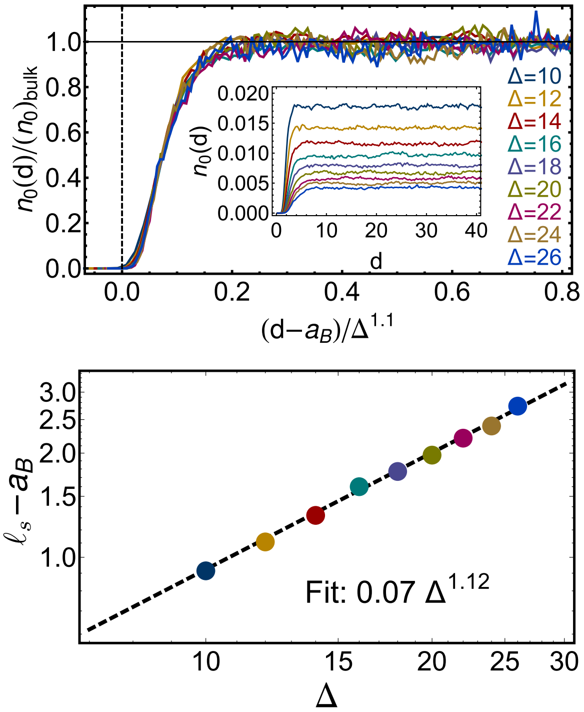

FIG. 8. Screening from surface states of a topological insulator suppresses puddle formation close to the surface. Upper panel: Scaling plot of the density of neutral dopants $n_0(d)$ as a function of the distance d to a metallic surface state ($|\mu^S| \gg E_c/\alpha^2$, $L = 50$). Curves are shifted by $a_B = 1$ since neutralization starts only for $d > a_B$. The inset shows the unscaled data. Lower panel: The width ℓ_s of the surface layer without puddles, defined by $n_0(\ell_s) = (n_0)_{\text{bulk}}/2$ as a function of Δ. A fit gives an exponent 1.12 very close to the bulk fit, see Eq. (10).

APPENDIX D: SCREENED COULOMB POTENTIAL IN THE PRESENCE OF TWO METALLIC LAYERS

The effective potential of a single impurity with (dimensionless) coordinate $\boldsymbol{x}_i$ with distance x_i^3 from a metallic layer at $x^3 = 0$ is described by [using the same conventions and cutoffs as in Eq. (3)]

$$V^m(\boldsymbol{x}_i,\boldsymbol{x}) = V\big[\boldsymbol{x} - \big(x_i^1,x_i^2,x_i^3\big)\big] - V\big[\boldsymbol{x} - \big(x_i^1,x_i^2,-x_i^3\big)\big]. \tag{D1}$$

A "mirror charge" guarantees that the potential vanishes on the metallic surface. In the presence of two metallic surfaces located at $z = 0$ and $z = L_z$, an infinite sequence of mirror charges is required. For efficient simulations it is mandatory that the screened potential can be computed rapidly. We have found that a potential containing only the mirror charge to the closest metallic surface in combination with a linear correction term which sets the potential to zero at both surfaces is sufficiently accurate and numerically efficient, see Fig. 7. We therefore use in our simulations for $z_i \leqslant L_z/2$,

$$V^s(\boldsymbol{x}_i,\boldsymbol{x}) = V^m(\boldsymbol{x}_i,\boldsymbol{x}) - \frac{z}{L_z} V^m[\boldsymbol{x}_i,(x_1,x_2,L_z)]. \tag{D2}$$

For $z_i > L_z/2$ we use a mirror image of the potential given above with mirror plane $z = L/2$. Due to this construction the derivative of the potential with respect to z_i has a small jump at $z_i = L_z/2$. We have found that this leads to a tiny, hardly visible bump at $z = L_z/2$ in the density of neutral dopants $n_0(z)$. When we determining the density of neutral dopants in the center of the slab, we fit a parabola to $n_0(z)$ for $0.4L_z < z < 0.6L_z$ omitting a tiny region of width $0.04L_z$ around $z = L_z/2$. In practice, the tiny bump (and the small corrections to the fit described above) does, however, have no qualitative or quantitative influence on our results.

APPENDIX E: SCREENING FROM A SINGLE SURFACE

In this Appendix we briefly discuss the suppression of puddle formation close to a metallic surface (for a system much thicker than ℓ_c). The inset of Fig. 8 shows $n_0(d)$ for values of Δ ranging from 10 to 26. On a relatively short length scale, the bulk value of $n_0(d)$ is reached. In the lower panel of Fig. 8 we plot the width ℓ_s of the zone, where surface screening suppresses puddle formation, defined by $n_0(\ell_s) = (n_0)_{\text{bulk}}/2$. After subtracting the offset $a_B = 1$ we obtain numerically an approximate power law relation in the numerically accessible regime

$$\ell_s \sim \Delta^\gamma, \quad \gamma \approx 1.1 \pm 0.2. \tag{E1}$$

We also performed simulation with several other values of a_B ($a_B = 0,0.5,1.5,2$) and have checked that subtracting a_B results in the same curve (for a fixed value of Δ). Also the scaling plot in the upper panel of Fig. 8 confirms that ℓ_d governs the size of "dead zone", where puddle formation is suppressed.

[1] A. L. Efros and B. I. Shklovskii, *Electronic Properties of Doped Semiconductors* (Springer, Berlin, 1984).

[2] E. H. Hwang, S. Adam, and S. Das Sarma, Phys. Rev. Lett. **98**, 186806 (2007).

[3] S. Adam, E. H. Hwang, V. M. Galitski, and S. Das Sarma, Proc. Natl. Acad. Sci. USA **104**, 18392 (2007).

[4] J. Martin, N. Akerman, G. Ulbricht, T. Lohmann, J. H. Smet, K. von Klitzing, and A. Yacoby, Nat. Phys. **4**, 144 (2008).

[5] Y. Zhang, V. W. Brar, C. Girit, A. Zettl, and M. F. Crommie, Nat. Phys. **5**, 722 (2009).

[6] N. Borgwardt, J. Lux, I. Vergara, Z. Wang, A. A. Taskin, K. Segawa, P. H. M. van Loosdrecht, Y. Ando, A. Rosch, and M. Grüninger, Phys. Rev. B **93**, 245149 (2016).

[7] C. W. Rischau, A. Ubaldini, E. Giannini, and C. J. van der Beek, New J. Phys. **18**, 073024 (2016).

[8] Z. Ren, A. A. Taskin, S. Sasaki, K. Segawa, and Y. Ando, Phys. Rev. B **84**, 165311 (2011).

[9] B. Skinner, T. Chen, and B. I. Shklovskii, Phys. Rev. Lett. **109**, 176801 (2012).

[10] B. Skinner, T. Chen, and B. I. Shklovskii, J. Exp. Theor. Phys. **117**, 579 (2013).

[11] B. Skinner and B. I. Shklovskii, Phys. Rev. B **87**, 075454 (2013).

[12] O. Breunig, Z. Wang, A. A. Taskin, J. Lux, A. Rosch, and Y. Ando, Nat. Commun. **8**, 15545 (2017).

[13] H. Beidenkopf, P. Roushan, J. Seo, L. Gorman, I. Drozdov, Y. S. Hor, R. J. Cava, and A. Yazdani, Nat. Phys. **7**, 939 (2011).

[14] S. A. Basylko, P. J. Kundrotas, V. A. Onischouk, E. E. Tornau, and A. Rosengren, Phys. Rev. B **63**, 024201 (2000).

[15] J. Lux, Ph.D. thesis, Universität zu Köln, 2016.

[16] B. I. Shklovskii and A. L. Efros, Zh. Eksp. Teor. Fiz. **62**, 1156 (1972) [Sov. Phys. JETP **35**, 610 (1972)].

[17] S. D. Baranovskii, B. I. Shklovskii, and A. L. Efros, Zh. Eksp. Teor. Fiz. **87**, 1793 (1984) [Sov. Phys. JETP **60**, 1031 (1984)].

[18] M. Lee, J. G. Massey, V. L. Nguyen, and B. I. Shklovskii, Phys. Rev. B **60**, 1582 (1999).

[19] T. Chen and B. Skinner, Phys. Rev. B **94**, 085146 (2016).

[20] V. F. Mitin, J. Appl. Phys. **107**, 033720 (2010).

[21] Y. Xia, D. Qian, D. Hsieh, L. Wray, A. Pal, H. Lin, A. Bansil, D. Grauer, Y. S. Hor, R. J. Cava, and M. Z. Hasan, Nat. Phys. **5**, 398 (2009).

[22] H. v. Löhneysen, A. Rosch, M. Vojta, and P. Wölfle, Rev. Mod. Phys. **79**, 1015 (2007).