Recovery of massless Dirac fermions at charge neutrality in strongly interacting twisted bilayer graphene with disorder

Alex Thomson and Jason Alicea

Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA; Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, California 91125, USA; and Department of Physics, California Institute of Technology, Pasadena, California 91125, USA

(Received 15 November 2019; revised 23 October 2020; accepted 30 November 2020; published 17 March 2021)

Stacking two graphene layers twisted by the magic angle $\theta \approx 1.1^\circ$ generates flat energy bands, which in turn catalyzes various strongly correlated phenomena depending on filling and sample details. At charge neutrality, transport measurements reveal superficially mundane semimetallicity (as expected when correlations are weak) in some samples yet robust insulation in others. We propose that the interplay between interactions and disorder admits either behavior, *even when the system is strongly correlated and locally gapped*. Specifically, we argue that strong interactions supplemented by weak, smooth disorder stabilize a network of gapped quantum valley Hall domains with spatially varying Chern numbers determined by the disorder landscape—even when an entirely different order is favored in the clean limit. Within this scenario, sufficiently small samples that realize a single domain display insulating transport characteristics. Conversely, multidomain samples exhibit re-emergent massless Dirac fermions formed by gapless domain-wall modes, yielding semimetallic behavior except on the ultralong scales at which localization becomes visible. We discuss experimental tests of this proposal via local probes and transport. Our results highlight the crucial role that randomness can play in ground-state selection of twisted heterostructures, an observation that we expect to have further ramifications at other fillings.

DOI: 10.1103/PhysRevB.103.125138

I. INTRODUCTION

The discovery of superconductivity and correlated insulators in magic-angle twisted bilayer graphene (mTBG) [1,2] opened a fascinating new chapter in the field of strongly interacting quantum matter. The "magic" stems from the fact that upon twisting the two graphene layers by an angle $\theta \approx 1.1^\circ$ from one another, the bands immediately above and below the charge neutrality point become exceptionally flat [3,4]—bringing interactions center stage. Accounting for spin and valley degrees of freedom, each of these two flat bands is essentially fourfold degenerate. Correlated physics, including superconductivity, thus naturally arises when the number of charge carriers per moiré unit cell is between $\nu = -4$ (four holes) and $\nu = +4$ (four electrons).

The observed phenomenology of mTBG depends sensitively on sample details. Cao *et al.* [1,2] originally observed correlated insulating states at $\nu = \pm 2$ along with superconducting domes upon doping away from the $\nu = -2$ insulator. Near the charge neutrality point at $\nu = 0$, the conductance exhibited a V-shaped suppression indicative of semimetallicity. Noninteracting band-theory calculations [4] predict massless Dirac fermions at charge neutrality, provided the system preserves $\mathcal{C}_2\mathcal{T}$ symmetry, with $\mathcal{C}_2$ a twofold rotation and $\mathcal{T}$ time reversal [5–8]; the latter observation thus at first sight suggests weak correlations at $\nu = 0$. The magic-angle device examined by Yankowitz *et al.* [9] additionally exhibited superconductivity adjacent to the $\nu = +2$ insulator and a resistive correlated state at $\nu = +3$. Near charge neutrality, transport again appeared consistent with the semimetallic behavior expected from band theory.

A second class of mTBG systems arises upon aligning the hexagonal boron nitride (hBN) substrate with one of the graphene sheets [10,11]. The alignment appears to underlie strikingly different correlated physics: an absence of superconductivity, removal of the $\nu = -2$ insulator, weak resistive peaks at $\nu = +2$ instead of robust insulation, and a quantum anomalous Hall state at $\nu = +3$. Furthermore, at charge neutrality, the system becomes strongly insulating instead of semimetallic. The behavior at charge neutrality is, however, yet again consistent with band theory. Indeed, alignment-induced breaking of the $\mathcal{C}_2$ symmetry renders the Dirac fermions massive, yielding a band gap at $\nu = 0$. Explicit $\mathcal{C}_2$ breaking has also been proposed as a catalyst for the observed quantum anomalous Hall state [12,13].

Still different phenomenology emerges in the ultrahomogeneous samples studied by Lu *et al.* [14]. These samples featured resistive peaks evincing either well-developed or incipient insulators at *all* integer fillings $\nu = 0, \pm 1, \pm 2, \pm 3$, as well as additional superconducting domes beyond those reported previously. Notably, the strongest insulating state within the flat band manifold occurred at charge neutrality, naively suggesting alignment with the hBN substrate as in Refs. [10,11]. Several factors challenge this interpretation, however. First, Lu *et al.* made no attempt to align the hBN, and it is unlikely to occur at random. Second, hBN-aligned samples and those of Lu *et al.* realize a largely disparate set of phenomena, suggesting against a common microscopic origin.

Finally, the gap reported by Lu *et al.* [14] dwarfs by roughly an order of magnitude that measured in hBN-aligned mTBG [11], making its formation by explicit symmetry-breaking seem unlikely in comparison. Thus the insulating behavior observed at all of the fillings indicated above—including $\nu = 0$—seems most naturally rooted in strong correlations.

A conservative interpretation of the available charge-neutrality transport data is that greater inhomogeneity in the samples from Refs. [1,2,9] merely obliterates the strong correlations operative at $\nu = 0$ in the Lu *et al.* samples. Such a viewpoint is supported by the fact that significant "twist-angle disorder" has been observed by multiple groups [1,2,9,15–19]; moreover, deviations from the magic angle locally enhance the flat band dispersion [20], potentially diminishing correlation effects. Scanning tunneling microscopy (STM) measurements from Refs. [15–18], however, do not simply fit this picture. All these STM studies observed *local* correlation effects at charge neutrality, manifested by a pronounced splitting of the flat band van Hove peaks upon approaching $\nu = 0$ and, in Ref. [18], evidence of a hard gap at charge neutrality.[1] Much subtler signatures of correlated states were also seen at other integer fillings (typically most prominently at $\nu = +2$). From a local perspective, it therefore appears that correlations in the STM samples are actually *strongest* at $\nu = 0$.

In this paper, we propose a unifying explanation for the diverse phenomenology observed to date in mTBG at charge neutrality. Our scenario posits that strong correlations are ubiquitous—even in samples that observe the semimetallic behavior expected from band theory—with disorder playing a secondary but still crucial role. We specifically assume that in a perfectly clean infinite system, interactions favor, or very nearly favor, correlated states that spontaneously break $\mathcal{C}_2\mathcal{T}$ symmetry in a way that yields Chern number $C = \pm 1$ for a given spin/valley sector. This assumption is bolstered by existing numerical simulations [14,16,21,22] and justified further below. Among the many possible insulators, only two preserve translation symmetry, spin rotation symmetry, and time reversal: the pair of quantum valley Hall (QVH) states [23–31] with $C = +1$ for both spins in one valley and $C = -1$ for both spins in the other valley or vice versa. Note that $\mathcal{C}_2$ transforms the QVH states into one another; hence they are exactly degenerate provided $\mathcal{C}_2$ is not explicitly broken.

Imagine now turning on smooth, nonmagnetic disorder (arising from twist-angle inhomogeneity, strain, etc.) that explicitly violates the infinite system's $\mathcal{C}_2$ symmetry but preserves it in an average sense. Within the manifold specified above, QVH states are unique in that their order parameter directly couples to the disorder potential—allowing the system to efficiently gain energy by locally forming one of those two phases. We further assume that the energy gain outweighs any energy cost (should one exist) for forming QVH order in the clean limit. Under these circumstances, the infinite system exhibits a random tiling of the two QVH states, details of which are determined by the interplay between interactions and the disorder landscape; see Fig. 1(a) for an illustration. Similar domain structures have been discussed in several other contexts, e.g., in systems with valley Hall nematic order [32,33] or as a source of non-Abelian PH-Pfaffian topological order [34–36].

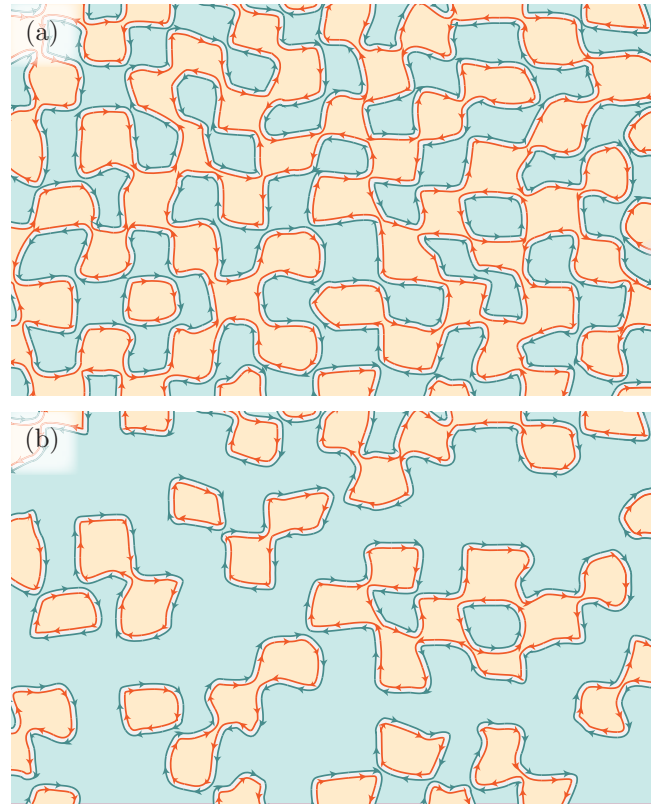

FIG. 1. Random tiling of quantum valley Hall states in a system where $\mathcal{C}_2$ is (a) preserved on average and (b) explicitly broken (e.g., by hBN alignment). For simplicity, we show the domain structure only for a single valley. Blue regions carry Chern number $C = +1$ for both spins, whereas orange regions carry Chern number -1. The arrows represent the two chiral edge modes (per spin) that traverse the domain boundaries. The domain structure corresponding to the valley sector not depicted here is simply obtained by exchanging the colors and reversing edge modes in (a) and (b).

Crucially, the infinite system is locally gapped within the QVH domains but is not entirely electrically inert. Each domain wall binds four right-moving and four left-moving charge-carrying modes, reflecting the fact that the Chern numbers for the spin/valley sectors change by ± 2 upon passing between adjacent domains. Smoothness of the disorder potential suppresses scattering among these modes and thus justifies treating the spin/valley sectors as decoupled (to a first approximation). In this limit, the system realizes four copies of a Chalker-Coddington network model [37] describing an integer-quantum-Hall (IQH) plateau transition at which the Hall conductivity changes by $\delta\sigma_{xy} = \pm 2e^2/h$ [38,39]. This set of plateau transitions can be described by eight massless Dirac fermions (two per sector) with disorder acting within each cone [40,41]. Hence, essentially the same low-energy physics expected from band theory emerges from a strongly

[1]The issue of hBN alignment is subtle given the different nature of these experiments. Nevertheless, alignment is expected to enhance the van Hove peak splitting *independent of filling*, whereas the splitting observed by STM is significantly larger at charge neutrality compared to when the bands are fully filled.

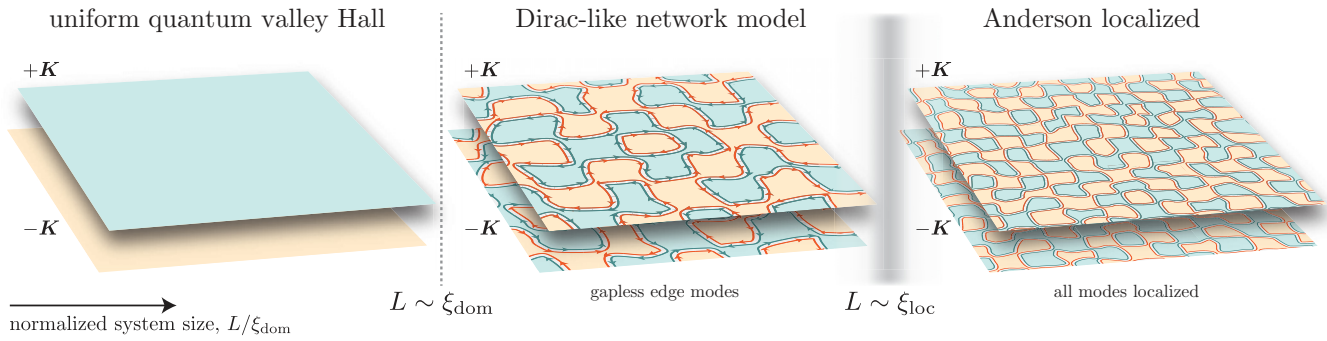

FIG. 2. Phase diagram versus system size L for a given disorder landscape. Left panel: With L below the typical domain size ξ_{dom}, a single domain is realized. Here each valley exhibits Chern number $C = +1$ (depicted in blue) or $C = -1$ (depicted in orange) throughout the entire sample, yielding insulating transport as observed by Lu *et al.* [14]. Central panel: When $L \gtrsim \xi_{\mathrm{dom}}$, multiple domains may be present within a single sample. The domain structure results in a percolating network of gapless edge modes that underlies the Dirac-like conductance seen by Cao *et al.* [1,2] and Yankowitz *et al.* [9]. Right panel: When L exceeds the localization length ξ_{loc} the sample localizes and ceases to conduct.

correlated framework! Residual scattering among the domain-wall modes generates intercone disorder that produces localization, but the localization lengths can be arbitrarily long.

We emphasize that strong correlations form the bedrock of the scenario outlined above. Without interactions and in the absence of explicit net $\mathcal{C}_2\mathcal{T}$-breaking, the fate of the system depends sensitively on the nonuniversal details of the disorder. The QVH state is but one among many potential phases, both gapped and gapless, that disorder could locally favor. Moreover, even if local QVH order happened to develop, the gap would be set by disorder and could be exceedingly small. In contrast, by inducing the spontaneous breaking of $\mathcal{C}_2\mathcal{T}$, interactions select a small subset of energetically competitive states in our scenario. Disorder then plays a subordinate role by favoring one of the two QVH orders in that set, thereby generating the domain structure. The local gap protecting the insulating domains is determined primarily by interactions rather than disorder.

Let us now revisit experiments in light of our proposed picture. Locally probing the QVH domains should reveal signatures of a correlation-driven gapped spectrum (possibly dressed with disorder-induced subgap states), consistent with STM experiments [15–18]. The outcome of global transport measurements depends on the ratio of sample size L to the typical domain size ξ_{dom} that would occur in an infinite system. For homogeneous systems such that $L/\xi_{\mathrm{dom}} \lesssim 1$, transport probes essentially a single domain, yielding insulating behavior as observed by Lu *et al.* [14]. (Strong intervalley scattering induced by the sample boundary is expected to suppress edge conduction). Conversely, for more-disordered samples with $L/\xi_{\mathrm{dom}} \gg 1$, transport probes many domains; here the massless Dirac fermions emerging from the gapless domain walls underpin semimetallic conduction as measured in Refs. [1,2,9]. See Fig. 2 for a summary. We can also make contact with the alignment-induced insulation observed in Refs. [10,11]. Turning on hBN alignment supplements the disorder landscape with a *uniform* $\mathcal{C}_2$-breaking potential that shrinks the area occupied by one of the QVH states and expands the area of the other, as shown in Fig. 1(b). Domain walls then no longer percolate, thereby gapping the re-emergent massless Dirac fermions and producing insulating transport when $L/\xi_{\mathrm{dom}} \gg 1$.

The arguments outlined above are justified through a Landau-Ginzburg theory describing the QVH order parameter. With the inclusion of disorder, we arrive at a classical 2D random-field Ising model (RFIM), which allows us to estimate the scaling of the typical domain size as a function of system parameters. Through a simple extension of this formulation, we can further study what occurs when a different phase that does *not* couple directly to disorder is energetically favored over the QVH state in the clean limit. As expected, when the (clean) ground-state energy splitting between the two states is sufficiently small—in a sense that we quantify with our Ising formulation—QVH order prevails throughout the majority of the sample.

Our scenario for ubiquitous strong correlations at charge neutrality is not only compatible with existing charge-neutrality data but further leads to falsifiable predictions both for STM and transport as described in Sec. VI. We also propose that two elements of this work may have broader applications in the study of mTBG. First, disorder can play a key role in discriminating among nearly degenerate correlated states. And second, disorder need not obliterate correlations but can mask them as seen by global transport experiments.

The rest of the paper is organized as follows. We begin by reviewing the low-energy theory and establishing our conventions for twisted bilayer graphene in Sec. II. Next, Sec. III describes the fate of noninteracting mTBG Dirac fermions at charge neutrality in the presence of disorder. We then discuss the clean interacting theory in Sec. IV. The interaction form is first outlined (Sec. IV A), and then used to argue that the QVH state is energetically competitive at charge neutrality [Secs. IV B and IV C]. Our main thesis is presented in detail in Sec. V, where each of the three regions illustrated in Fig. 2 is described in turn. We conclude in Sec. VI by summarizing and highlighting future directions. Supplemental details appear in numerous Appendices.

II. REVIEW OF LOW-ENERGY THEORY

In this section, we set the stage by reviewing the low-energy physics of mTBG at charge neutrality in the absence of interactions and disorder.

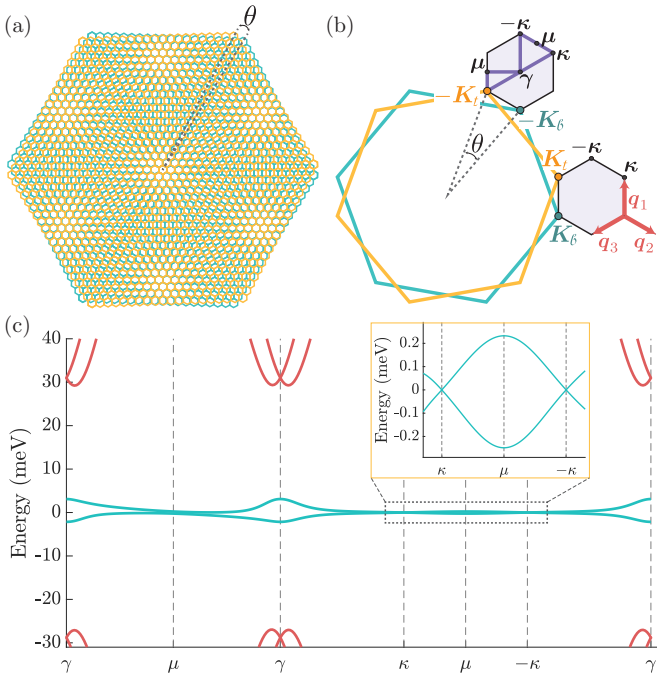

FIG. 3. (a) Cartoon representation of twisted bilayer graphene. The top and bottom graphene sheets are, respectively, represented by the orange and turquoise honeycomb lattices. The light, orange-tinted AA regions form a triangular superlattice, each of which is surrounded by a darker hexagonal rim whose vertices correspond to alternating AB and BA stacking regions. (b) Representation of the microscopic and moiré Brillouin zones. The large orange and turquoise hexagons represent the microscopic Brillouin zones of the underlying graphene layers. The moiré Brillouin zones, shown in purple, are defined by the distance between the the $\boldsymbol{K}$ points of the top and bottom layers. The size of the twist angles in both (a) and (b) has been exaggerated for clarity. (c) Flat bands corresponding to one $\boldsymbol{K}$-valley as calculated using the continuum model along the momentum line cut shown in purple in (b). Here, we took $\theta = 1.05^\circ$. Lattice relaxation is mimicked by decreasing the AA tunneling amplitude w_0 relative to the AB tunneling amplitude w_1. In particular, we have $w_0 = 85\,\text{meV}$ and $w_1 = 110\,\text{meV}$ [6,42,43]. The inset zooms in on the Dirac cones at $\boldsymbol{\kappa}$ and $-\boldsymbol{\kappa}$.

A. Continuum model

Consider two monolayer-graphene sheets stacked such that they are twisted relative to one another by an angle θ, as shown in Fig. 3(a) (for an arbitrary angle θ). The twist dramatically reduces the system's translational symmetry. While true translational symmetry requires special commensurate angles, when the twist angle is small, an effective moiré translational symmetry emerges; the resulting triangular superlattice of orange AA regions, each surrounded by a hexagon of alternating AB and BA regions, is clearly visible in the cartoon of Fig. 3(a). In this case, the band structure at charge neutrality descends from the band structure of the individual graphene layers in a relatively straightforward manner when described in momentum space [3,4]. Figure 3(b) shows the Brillouin zones (BZs) of the top and bottom graphene monolayers after applying a rotation by an angle $+\theta/2$ and $-\theta/2$, respectively. The reciprocal lattice vectors of the resulting moiré pattern are given by $\boldsymbol{G}_\ell = \mathcal{R}_{\theta/2}[\boldsymbol{\mathcal{G}}_\ell] - \mathcal{R}_{-\theta/2}[\boldsymbol{\mathcal{G}}_\ell]$ where $\boldsymbol{\mathcal{G}}_{1,2}$ denote the reciprocal lattice vectors of the unrotated graphene sheets and $\mathcal{R}_\phi[\boldsymbol{v}]$ rotates a vector $\boldsymbol{v}$ by an angle ϕ. The length of the moiré reciprocal lattice vectors, $|\boldsymbol{G}_\ell|$, is therefore suppressed relative the graphene reciprocal lattice vectors by a factor of $2\sin(\theta/2) \sim \theta$, making it very small by assumption. Equivalently, the moiré lattice constant is enlarged by $\sim 1/\theta$ relative to the graphene lattice constant. We let $\pm\boldsymbol{K}_t = \mathcal{R}_{\theta/2}[\pm\boldsymbol{K}]$ and $\pm\boldsymbol{K}_b = \mathcal{R}_{-\theta/2}[\pm\boldsymbol{K}]$ denote the $\pm\boldsymbol{K}$ points of the top and bottom layers, respectively.

As tunneling between the two layers turns on, states at momentum $\boldsymbol{K}_t + \boldsymbol{k}$ on the top layer mix with those at momentum $\boldsymbol{K}_b + \boldsymbol{k} + \boldsymbol{q}_\ell + \boldsymbol{G}$ on the top layer, where $\boldsymbol{G}$ is a moiré reciprocal lattice vector and $\boldsymbol{q}_\ell = \mathcal{R}_{2\pi(\ell-1)/3}[\boldsymbol{K}_t - \boldsymbol{K}_b]$, $\ell = 1, 2, 3$ [see Fig. 3(b)]. Since the moiré BZ is much smaller than the BZ of monolayer graphene, mixing states proximate to $\boldsymbol{K}$ with those proximate to $-\boldsymbol{K}$ is an extremely high-order tunneling process; the two valleys of the original graphene monolayers thus effectively decouple. This decoupling is particularly convenient as it allows us to express the full band Hamiltonian H_{cont} as a sum of terms for the $\boldsymbol{K}$ and $-\boldsymbol{K}$ valleys: $H_{\text{cont}} = H_+ + H_-$. For convenience, we explicitly reproduce H_+ in Appendix A. Our ability to decompose the Hamiltonian into $\boldsymbol{K}$-valley sectors is equivalent to the emergence of a U(1) "valley" symmetry, which we denote $\text{U}(1)_v$.

In addition to moiré translations and $\text{U}(1)_v$, the continuum model preserves the $\text{SU}(2)_s$ spin rotation symmetry (neglecting spin-orbit coupling), time reversal $\mathcal{T}$, $\mathcal{C}_2$ rotations by π, $\mathcal{C}_3$ rotations by $2\pi/3$, and a mirror symmetry $\mathcal{M}_y$ that takes $(x, y) \to (x, -y)$ and interchanges the two layers. The latter three should be regarded as emergent symmetries similar to $\text{U}(1)_v$. In our conventions, the time-reversal operator $\mathcal{T}$ does *not* flip the electronic spins and accordingly obeys $\mathcal{T}^2 = +1$. Both $\mathcal{T}$ and $\mathcal{C}_2$ interchange the two valleys and hence are not symmetries of the individual single-valley Hamiltonians $H_\pm$. Rather than keep track of these two symmetries separately, it is therefore convenient to consider $\mathcal{T}$ along with the composite operation $\mathcal{C}_2\mathcal{T}$—which commutes with $\text{U}(1)_v$.

The momenta $\boldsymbol{K}_t$ and $\boldsymbol{K}_b$ map to the corners of the moiré BZ. In what follows, we denote these momenta by $\pm\boldsymbol{\kappa}$ to distinguish them from the $\pm\boldsymbol{K}$-valleys of the microscopic graphene layers. Provided the $\mathcal{C}_2\mathcal{T}$ and $\mathcal{C}_3$ symmetries are present, the massless Dirac cones at $\boldsymbol{K}_{t,b}$ for the microscopic graphene layers evolve into massless Dirac cones at $\pm\boldsymbol{\kappa}$ even once tunneling is turned on. This crucial property follows from the fact that the Berry phase enclosed within any loop is quantized to 0 or π (mod 2π) when $\mathcal{C}_2\mathcal{T}$ is preserved. Since a Dirac point necessarily exhibits Berry phase π, the Dirac cones at $+\boldsymbol{\kappa}$ and $-\boldsymbol{\kappa}$ are locally protected against a mass [5,6]. Breaking $\mathcal{C}_3$ can shift the location of the Dirac cones, but cannot gap them. Importantly, since both cones in H_+ (H_-) descend from the Dirac cones at $\boldsymbol{K}_{t,b}$ ($-\boldsymbol{K}_{t,b}$) in a continuous fashion, they possess the same chirality [7,8]—thereby obstructing the development of a two-band, single $\boldsymbol{K}$-valley tight-binding model in which all symmetries are realized in a local fashion [6,44,45].

B. Flat bands

The previous subsection highlighted generic features of small-angle twisted bilayer graphene. At the *magic* angle,

the velocity of the massless Dirac fermions becomes very small, and the bands immediately above and below the charge neutrality point separate from the remaining bands by a finite energy (provided lattice relaxation is incorporated [6,42,43]); see Fig. 3(c). The resulting energetically isolated "flat bands" are each (essentially) fourfold degenerate, reflecting spin and valley degrees of freedom. We now describe the flat band Hamiltonian by first focusing on the $+\boldsymbol{K}$ valley and subsequently incorporating the $-\boldsymbol{K}$ valley.

Let $c_{\alpha j}(\boldsymbol{k})$ denote momentum-space annihilation operators associated with the flat bands at valley $+\boldsymbol{K}$; here $\alpha = \uparrow, \downarrow$ is a spin index and $j = 1, 2$ is a band index. Reference [46] showed that these operators can be defined such that they transform under $\mathcal{C}_2\mathcal{T}$ via

$$\mathcal{C}_2\mathcal{T}: \quad c(\boldsymbol{k}) \to \eta^x c(\boldsymbol{k}), \quad i \to -i, \tag{1}$$

with Pauli matrices $\eta^{x,y,z}$ that act on the band indices. (Here and below we often suppress indices for notational simplicity). It follows that the $\mathcal{C}_2\mathcal{T}$-invariant flat band Hamiltonian takes the form

$$H_0 = \int_{\boldsymbol{k}\in \mathrm{BZ}} c^\dagger(\boldsymbol{k})[h_0(\boldsymbol{k}) + h_x(\boldsymbol{k})\eta^x + h_y(\boldsymbol{k})\eta^y]c(\boldsymbol{k}). \tag{2}$$

Next we project onto the massless Dirac fermions at $\pm\boldsymbol{\kappa}$ in the moiré BZ by defining Dirac spinors $\psi_{1\alpha j}(\boldsymbol{q}) \sim c_{\alpha j}(+\boldsymbol{\kappa} + \boldsymbol{q})$ and $\psi_{2\alpha j}(\boldsymbol{q}) \sim c_{\alpha j}(-\boldsymbol{\kappa} + \boldsymbol{q})$ and retaining only small $\boldsymbol{q}$ modes. The fact that the massless Dirac cones exhibit the same chirality at $\pm\boldsymbol{\kappa}$ implies that $h_x(\pm\boldsymbol{\kappa} + \boldsymbol{q}) \sim +q_x$ and $h_y(\pm\boldsymbol{\kappa} + \boldsymbol{q}) \sim +q_y$. Upon shifting the energy such that $h_0(\pm\boldsymbol{\kappa}) = 0$ and reverting to real space, the low-energy Hamiltonian becomes

$$H_D = -\int_{\boldsymbol{r}} v_F \psi^\dagger(i\partial_x\eta^x + i\partial_y\eta^y)\psi. \tag{3}$$

The Fermi velocity v_F has been assumed isotropic and identical for both $\pm\boldsymbol{\kappa}$ Dirac cones, which is guaranteed when all symmetries outlined earlier are present.

An insulating phase at charge neutrality may only be obtained by either breaking the $\mathcal{C}_2\mathcal{T}$ symmetry or by closing the gap separating the flat bands and the dispersing bands [5,6]. We focus entirely on the former scenario, which is straightforward to represent using the Dirac theory. Let $\tau^{x,y,z}$ and $\sigma^{x,y,z}$ denote Pauli matrices that respectively act on $\boldsymbol{\kappa}$-valley indices and spin indices. Mass terms then take the form $\psi^\dagger\eta^z M\psi$ with $M = \{\mathbb{1}, \sigma^i, \tau^i, \tau^i\sigma^j\}$. The Chern number for a given spin/valley sector depends on the relative sign of the masses gapping the $\boldsymbol{\kappa}$ and $-\boldsymbol{\kappa}$ Dirac cones. When both cones have the same-sign mass, the sector acquires Chern number $C = \pm 1$, whereas opposite-sign masses yield $C = 0$. Consequently, mass terms with $M = \mathbb{1}$, σ^i yield insulating bands with nonzero Chern number, while masses with $M = \tau^i$, $\tau^i\sigma^j$ yield trivial insulating bands.[2]

[2]Given the Hamiltonian H_D, the above conclusions regarding Chern number hold true regardless of the details of the high-energy theory from which it was derived. It is worth noting that for a single graphene sheet, expanding in small $\boldsymbol{q}$ the functions analogous to $h_{x,y}(\pm\boldsymbol{\kappa} + \boldsymbol{q})$ would not yield the low-energy Hamiltonian of Eq. (3). Instead, the two Dirac cones would possess opposite chirality.

We now restore the $-\boldsymbol{K}$ valley. In terms of the low-energy Dirac Hamiltonian, the chirality of the massless Dirac fermions in the $-\boldsymbol{K}$ valley is opposite that of the $+\boldsymbol{K}$ valley. Defining the spinor $\Psi = (\psi_+, \psi_-)^T$, where $\psi_\pm$ describe Dirac fermions in valley $\pm\boldsymbol{K}$, the full Dirac Hamiltonian may be written

$$H_{D,\mathrm{tot}} = -\int_{\boldsymbol{r}} v_F \Psi^\dagger(i\partial_x \mu^z\eta^x + i\partial_y \eta^y)\Psi, \tag{4}$$

where we introduced Pauli matrices $\mu^{x,y,z}$ that act on $\boldsymbol{K}$-valley indices. The presence of μ^z in the first term above implements the opposite-chirality requirement. Our discussion of the mass terms and the associated Chern numbers extends straightforwardly to the Ψ fermions. For details see Appendix B. A notable consequence of the opposing chiralities of the $\pm\boldsymbol{K}$ valleys is that a mass term $\Psi^\dagger\eta^z\Psi = \psi_+^\dagger\eta^z\psi_+ + \psi_-^\dagger\eta^z\psi_-$ generates an insulator with $C = +1$ for the $+\boldsymbol{K}$ valley and $C = -1$ for the $-\boldsymbol{K}$ valley (or vice versa depending on the overall sign of the mass term). These insulators correspond to the QVH states that play a prominent role in this paper.

In the following sections, we use the operator ψ when restricting our discussion to a single $\boldsymbol{K}$-valley. We suppress the $\pm$ indices in such cases but assume for concreteness that the $+\boldsymbol{K}$ valley is being considered (as in the beginning of this subsection). We reserve use of Ψ for occasions when both valleys are discussed simultaneously.

III. FREE FERMIONS WITH DISORDER

Next, we discuss the physics of noninteracting twisted bilayer graphene with disorder at charge neutrality.

A. Sources of disorder in twisted bilayer graphene

It is useful to review the specific types of disorder that are believed to be most relevant to experiments, though we attempt to keep the majority of our discussion as general as possible. Charge disorder appears to be quite low: Refs. [2,9,19] estimate charge-carrier inhomogeneity in the range $\delta n \sim 1 - 2 \times 10^{10}\,\mathrm{cm}^{-2}$. Yankowitz *et al.* [9] further considered the observation of fractional quantum Hall states at magnetic fields as low as 4 T as additional proof of the high purity of their sample.

Twist-angle disorder is perhaps the most prevalent type of inhomogeneity in mTBG systems. Due to strain, different regions of a given sample may correspond to different twist angles, as directly imaged in STM [15–17]. From topography, the AA regions of the moiré structure are very clear, allowing one to locally establish the moiré lattice constants and thus the twist angle. Twist-angle variations were more recently characterized by Uri *et al.* [19] using a superconducting quantum interference device on a tip. Under an applied magnetic field, these authors measured the electron density of the sample as a function of the tip location, which in turn allowed them to map out the twist angle throughout the entire sample. Such measurements indicated local twist angles varying within a range $\delta\theta \sim 0.1^\circ$. Both samples they studied developed correlated insulating states, but only the sample with a continuous magic-angle region percolating across the sample displayed clear signs of superconductivity.

While transport measurements cannot access such local information, by comparing two-terminal conductance measurements between different pairs of contacts, Cao *et al.* [1,2] and Yankowitz *et al.* [9] nevertheless note that some regions require different electron densities to achieve the band insulator at full-filling, again implying that unit cells differ between regions. Similar measurements by Lu *et al.* [14] returned a much more uniform signal across the sample. Disorder signatures are also observable from within the superconducting states. Both Cao *et al.* and Yankowitz *et al.* observe phase-coherent Fraunhofer interference, indicating the coexistence of superconducting and normal regions. Conversely, the interference patterns measured by Lu *et al.* are comparatively weak, which they take as further indication of the high degree of sample homogeneity.

The hBN substrate may serve as yet another source of disorder. When uniformly aligned with one of the graphene monolayers, $\mathcal{C}_2\mathcal{T}$ symmetry is explicitly broken and a gap at charge neutrality is opened [12,13]. While the explicit gapping naturally explains the $\nu = 0$ insulator and anomalous Hall effect observed by Sharpe *et al.* [10] and Serlin *et al.* [11] at $\nu = +3$, hBN-alignment is believed to be an otherwise small effect in the majority of samples studied. Nevertheless, it is possible that a *local* alignment of the substrate, differing between regions, could weakly break the $\mathcal{C}_2\mathcal{T}$ symmetry—just as for twist-angle disorder—even though it may be present on average.

B. Theoretical modeling of disorder

Motivated by the preceding discussion, we now incorporate weak, smooth disorder that preserves time-reversal and spin-rotation symmetries. We model such disorder by coupling spatially varying (but static) fields to fermion bilinears of the non-interacting Dirac theory reviewed in Sec. II B. The most relevant forms of disorder couple to bilinears that do not contain derivatives, and so we focus our study on this subset.[3] Time-reversal invariance and spin symmetry further reduce the number of bilinears capable of coupling to disorder; we enumerate all such symmetry-preserving terms in Appendix B. Collectively denoting the set of symmetry-allowed operators by $\{\Psi^\dagger T^i \Psi\}$, the most general disorder Hamiltonian takes the form

$$H_{\text{dis}} = \int_{\boldsymbol{r}} \sum_i R_i(\boldsymbol{r}) \Psi^\dagger(\boldsymbol{r}) T^i \Psi(\boldsymbol{r}). \tag{5}$$

We assume Gaussian-distributed $R_i(\boldsymbol{r})$ with zero mean and variance:

$$\overline{R_i(\boldsymbol{r}) R_j(\boldsymbol{r}')} = \delta_{ij} g_i^2 K_i((\boldsymbol{r} - \boldsymbol{r}')/\xi_i). \tag{6}$$

Here, g_i is the disorder strength with units of energy, ξ_i is the disorder correlation length, and K_i is a dimensionless function that characterizes the spatial correlations of the disorder and obeys $K_i(0) = 1$. We frequently specialize to the case where the spatial correlations are Gaussian, i.e.,

$$K_i(\boldsymbol{r}/\xi_i) = e^{-\boldsymbol{r}^2/(2\xi_i^2)}. \tag{7}$$

Weakness of the disorder implies that g_i are small relative to the other scales of the theory, enabling a perturbative treatment. Smoothness of disorder is imposed by requiring that $\xi_i \gtrsim a_M$, with a_M the moiré lattice constant. We assume that the correlation lengths corresponding to different forms of disorder do not differ substantially and simply set $\xi_i = \xi_{\text{dis}}$ for all i.

The smoothness condition is physically very natural given that the existence of the moiré superlattice and the resulting band structure is predicated on the absence of fluctuations on the scale of the graphene lattice constant a. In momentum space, smoothness implies the suppression of inhomogeneities mediating momentum exchanges of order $\sim|\boldsymbol{K}|$, i.e., disorder processes that couple to bilinears of the form $\Psi^\dagger \mu^{x,y} M \Psi$. In fact, we demonstrate in Appendix C that given Gaussian-correlated disorder [Eq. (7)], the disorder strengths corresponding to inter-$\boldsymbol{K}$-valley scattering are exponentially suppressed relative the intra-$\boldsymbol{K}$-scattering disorder strengths: if g is the magnitude of a typical intra-$\boldsymbol{K}$-valley disorder field, then

$$g_{KK'} \sim g\, e^{-\boldsymbol{K}^2 \xi_{\text{dis}}^2/4} = g\, e^{-4\pi^2 \xi_{\text{dis}}^2/a^2} \tag{8}$$

is the typical amplitude of an inter-$\boldsymbol{K}$-valley scattering event. Neglecting such exponentially suppressed events for now, we focus on a single $\boldsymbol{K}$-valley and couple disorder to the ψ fermions described by H_D.

Since time-reversal interchanges $\boldsymbol{K}$-valleys, it is *not* a symmetry of the single-$\boldsymbol{K}$-valley theory, implying that the system is described by the Wigner-Dyson class A [47,48]. Disorder can thus couple to all spin-rotation-invariant bilinears and takes the form

$$\begin{aligned} H_{\text{dis}}^{\text{smooth}} = \int_{\boldsymbol{r}} \psi^\dagger(\boldsymbol{r}) \Big\{ & \mathcal{M}_0(\boldsymbol{r}) \eta^z + \mathcal{M}_\ell(\boldsymbol{r}) \eta^z \tau^\ell \\ & + \sum_{i=x,y} \Big[\mathcal{A}_{i,0}(\boldsymbol{r}) \eta^i + \sum_{\ell=x,y,z} \mathcal{A}_{i,\ell}(\boldsymbol{r}) \eta^i \tau^\ell \Big] \\ & + \mathcal{V}_0(\boldsymbol{r}) + \sum_{\ell=x,y,z} \mathcal{V}_\ell(\boldsymbol{r}) \tau^\ell \Big\} \psi(\boldsymbol{r}), \end{aligned} \tag{9}$$

where $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{V}$, respectively, represent various forms of mass, vector potential, and scalar potential disorder.

It is also useful to consider the limit where disorder is sufficiently smooth relative to the moiré lattice scale that inter-$\boldsymbol{\kappa}$-valley scattering may also be neglected. We can then further restrict our attention to one of the Dirac cones in the moiré BZ—say $+\boldsymbol{\kappa}$. Denoting the spinor describing the Dirac cone at $+\boldsymbol{\kappa}$ by $\chi(\boldsymbol{r})$, the disorder Hamiltonian becomes simply

$$\begin{aligned} H_{\text{dis}}^{\text{ultra-smooth}} = \int_{\boldsymbol{r}} \chi^\dagger(\boldsymbol{r}) \Big[& m(\boldsymbol{r}) \eta^z \\ & + \sum_{i=x,y} a_i(\boldsymbol{r}) \eta^i + v(\boldsymbol{r}) \Big] \chi(\boldsymbol{r}), \end{aligned} \tag{10}$$

where the random mass m, vector potential $a_{x,y}$ and scalar potential v satisfy Eq. (6) with $R_i = m$, $a_{x,y}$, v. Since each

[3] In particular, we neglect disorder-induced variation in the Fermi velocity. This omission is supported by the numerics of Ref. [20], which show that the velocity remains largely unaffected by the presence of twist-angle disorder.

moiré unit cell encompasses ∼10 000 carbon atoms, distilling the disorder Hamiltonian down to Eq. (10) is significantly more suspect than merely omitting inter-$\boldsymbol{K}$-valley scattering terms. Moreover, though it might naively appear that inter-$\boldsymbol{\kappa}$-scattering should be suppressed in a manner analogous to Eq. (8), we only expect such an effect to be manifest for extremely large correlation lengths $\xi_{\rm dis}$ relative to a_M as discussed at the end of Appendix C. We nevertheless argue in Sec. V that interactions greatly enhance the validity of Eq. (10) over a broader parameter regime.

C. Free Dirac fermions coupled to disorder

While we are interested in the situation where the disorder strength is subleading relative to interactions, it is instructive to review the expected fate of the free Dirac theory at charge neutrality in several limits. Consider first the single-Dirac-cone theory with disorder described by $H_{\rm dis}^{\rm ultrasmooth}$. Having restricted to this minimal theory, it is convenient to abandon smooth disorder and instead take white-noise correlations such that $K_i(\boldsymbol{r}/\xi_{\rm dis}) = \xi_{\rm dis}^2\delta^2(\boldsymbol{r})$. Physically, this simplification implies that we are probing the system at long enough scales relative to $\xi_{\rm dis}$ that all correlations in R_i are washed away. The disorder correlation length is then encoded in the dimensionless (up to factors of $\hbar$ and v_F) disorder strength parametrized by $g_i^2\xi_{\rm dis}^2$.

Ludwig *et al.* [49] analyzed the effect of each of the three remaining disorder fields— $m(\boldsymbol{r})$, $a_{x,y}(\boldsymbol{r})$, and $v(\boldsymbol{r})$. In the absence of all other types of disorder, the random mass, vector potential, and scalar potential fields were individually found to be marginally irrelevant, exactly marginal, and marginally relevant in turn. Ludwig *et al.* further postulated that when all three disorder types are simultaneously present, the system flows to the IQH plateau transition fixed point.

The correspondence between Landau-level physics and disordered Dirac theories may be understood from the perspective of a Chalker-Coddington network model [37]. This model can be employed to efficiently study the transition between a trivial insulator with Landau-level filling $\tilde{\nu} = 0$ an IQH state with $\tilde{\nu} = 1$ (the tilde distinguishes Landau-level filling from the mTBG filling). The system is assumed to locally prefer either $\tilde{\nu} = 0$ or $\tilde{\nu} = 1$—thus forming domains of trivial and IQH states whose detailed structure depends on the total filling and the disorder potential. As the total filling varies, either the trivial state percolates, with small "lakes" of $\tilde{\nu} = 1$, or vice versa. At some critical value, the system transitions between these two limits and becomes gapless. The network model exploits the fact that each boundary between $\tilde{\nu} = 0$ and $\tilde{\nu} = 1$ regions binds a chiral edge mode, and maps the problem onto one of directed links scattering at different nodes.

The key observation for our purposes is that the network model may be directly mapped onto a massless Dirac cone coupled to random fields $m(\boldsymbol{r})$, $a_{x,y}(\boldsymbol{r})$, and $v(\boldsymbol{r})$ [40,41]. The correspondence between a lone disordered Dirac cone and the IQH plateau transition has been studied more recently in the context of *monolayer* graphene [50], where reducing the problem to that of a single Dirac cone only requires that disorder correlations are smooth on the scale of the microscopic lattice. When the effective time-reversal symmetry of the single Dirac cone is broken by strain [51], the appropriate nonlinear σ-model was shown to possess a topological term with $\theta = \pi$; consequently, the system exhibits universal conductivity, just as predicted for the IQH plateau transition by Pruisken [52,53].

Upon resurrecting inter-$\boldsymbol{\kappa}$-valley scattering in mTBG, disorder is instead described by $H_{\rm dis}^{\rm smooth}$. Here the theory localizes in the thermodynamic limit, and the conductivity accordingly approaches zero even at charge neutrality [54,55]. Nevertheless, the localization length is expected to be extremely long since the scaling theory of Anderson localization indicates a lower critical dimension of $d = 2$ [56,57]. The conductance thus only vanishes logarithmically with system size, suggesting that the localization length may be exponentially long, at least for a typical metal. Fradkin studied the fate of a system featuring N_f massless Dirac cones in the large-N_f limit [58,59]. Denoting the disorder strength for all processes simply by g, he obtained an exponentially large mean free path,

$$\ell_{\rm mfp} \sim a_M \exp\left[\frac{\pi}{2}\left(\frac{\hbar v_F}{g\,\xi_{\rm dis}}\right)^2\right], \tag{11}$$

and a still-larger localization length, $\xi_{\rm loc}\sim\ell_{\rm mfp}\exp(64N_f^2/9)$.

IV. INTERACTIONS IN THE CLEAN LIMIT

Turning away from the question of disorder, we now investigate the effect of interactions in a homogeneous sample (though we occasionally allude to disorder effects). We begin with a discussion of the general form and magnitude of the interactions. Drawing on numerical results, experimental observations, and symmetry considerations, we then argue that the QVH state is energetically competitive in interacting mTBG at charge neutrality.

A. Coulomb interaction

The Coulomb Hamiltonian $H_{C,\rm tot} = \frac{1}{2}\int_{\boldsymbol{q}} V(\boldsymbol{q})\rho(\boldsymbol{q})\rho^\dagger(\boldsymbol{q})$ encodes the leading interaction. Here $V(\boldsymbol{q})$ is the Fourier-transform of the long-range Coulomb potential (which technically depends on both the layer and sublattice, but these microscopic corrections can be ignored for the purpose of our discussion). The operator $\rho(\boldsymbol{q})$ represents the Fourier transform of the full microscopic density. Specifically, we write $\rho(\boldsymbol{q}) = \sum_\ell \int_{\boldsymbol{k}} \tilde{f}_\ell^\dagger(\boldsymbol{k})\tilde{f}_\ell(\boldsymbol{k}+\boldsymbol{q})$, where $\tilde{f}_\ell(\boldsymbol{k})$ denotes the annihilation operator corresponding to one of the decoupled graphene monolayers, with ℓ a combined index labeling both layer and sublattice and $\boldsymbol{k}$ taking values across the full microscopic BZ. As explained in Sec. II A, to a high degree of accuracy, the flat band wave functions are composed entirely of states originating proximate to the Dirac cones of the decoupled monolayers. We focus on these important momenta by introducing operators $f_{\ell,n=\pm}(\boldsymbol{k}) \equiv \tilde{f}_\ell(\boldsymbol{k}\pm\boldsymbol{K})$ that are defined for $|\boldsymbol{k}| \ll |\boldsymbol{K}|$; note that this "small $\boldsymbol{k}$" condition does not necessarily imply that $\boldsymbol{k}$ resides within the moiré BZ. It follows that only the density operators $\rho(\boldsymbol{q})$ and $\rho(\boldsymbol{q}\pm\boldsymbol{K})$ with $\boldsymbol{q}$ small are physically relevant to the flat band

physics:

$$\rho(\boldsymbol{q}) \cong \sum_{\ell,n} \int_{\boldsymbol{k}\ \text{small}} f^\dagger_{\ell,n}(\boldsymbol{k}) f_{\ell,n}(\boldsymbol{k}+\boldsymbol{q}),$$

$$\rho(\boldsymbol{q}+\boldsymbol{K}) \cong \sum_{\ell} \int_{\boldsymbol{k}\ \text{small}} f^\dagger_{\ell,+}(\boldsymbol{k}) f_{\ell,-}(\boldsymbol{k}+\boldsymbol{q}) = \rho^\dagger(-\boldsymbol{q}-\boldsymbol{K}). \tag{12}$$

Inserting these definitions into our expression for $H_{C,\text{tot}}$, we find $H_{C,\text{tot}} \cong H_C + H'_C$ where

$$H_C = \frac{1}{2}\int_{\boldsymbol{q}\ \text{small}} V(\boldsymbol{q})\rho(\boldsymbol{q})\rho^\dagger(\boldsymbol{q}),$$

$$H'_C = \int_{\boldsymbol{q}\ \text{small}} V(\boldsymbol{q}+\boldsymbol{K})\rho(\boldsymbol{q}+\boldsymbol{K})\rho^\dagger(\boldsymbol{q}+\boldsymbol{K}). \tag{13}$$

There is a vast separation of energy scales between H_C and H'_C. Since $V(\boldsymbol{q}) \propto 1/|\boldsymbol{q}|$, the largest contribution to H_C comes from momenta $\boldsymbol{q}$ within the moiré BZ, i.e., $|\boldsymbol{q}| \lesssim |\boldsymbol{\kappa}|$. On the other hand, in H'_C, the smallness of the internal momentum $\boldsymbol{q}$ implies $V(\boldsymbol{q}+\boldsymbol{K}) \approx V(\boldsymbol{K})$. It follows that the relative strength of H_C and H'_C is $V(\boldsymbol{K}+\boldsymbol{q})/V(\boldsymbol{q}) \lesssim V(\boldsymbol{K})/V(\boldsymbol{\kappa}) \sim |\boldsymbol{\kappa}|/|\boldsymbol{K}| \sim \theta \ll 1$. Hereafter, we focus our attention on the dominant term, H_C. Reverting to real space, the Coulomb potential is $V(\boldsymbol{r}) = e^2/(4\pi\epsilon|\boldsymbol{r}|)$. For graphene on hBN, we estimate the dielectric constant to be $\epsilon \sim 8\epsilon_0$ with ϵ_0 denoting the permittivity of free space. Using the moiré lattice spacing, $a_M = a/(2\sin(\theta/2))$, where $a = 0.246\,\text{nm}$ is the lattice constant of monolayer graphene as a typical length scale, one finds a characteristic interaction energy $V(a_M) \sim 14\,\text{meV}$ at the magic angle $\theta \sim 1.05^\circ$.

Theory estimates the bandwidth of the flat bands to be about 10 meV and the splitting between van Hove peaks within those bands to be $\sim$5 meV [43,60]. The above Coulomb-interaction scale thus raises natural questions regarding the validity of our expansion about the Dirac cones at $\pm\boldsymbol{\kappa}$ in Sec. II B. It appears that the entirety of the flat bands and perhaps even neighboring energy bands should be considered. Non-interacting simulations of mTBG systems with twist angle disorder, however, have been shown to increase the bandwidth with little change to the Dirac character at charge neutrality [20]. Moreover, STM measurements of the fully filled flat bands (i.e., in a regime where correlations are presumably less important) measure van Hove peak splittings of $\sim 10-20\,\text{meV}$ [15,16]—several times larger than the above theoretical estimate. The full bandwidth of the flat bands may therefore significantly exceed $V(a_M)$, supporting our use of the Dirac theory.

B. Preferred ground state of single-flavor theory

Before turning to the full theory, it is useful to examine interaction effects at charge neutrality in a minimal, single-flavor model that includes only one spin and one $\boldsymbol{K}$-valley. References [14,16,21,22] addressed this problem numerically via self-consistent Hartree-Fock calculations. Liu *et al.* [21] incorporated Coulomb interactions in the continuum model while Choi *et al.* [16] studied a ten-band lattice model [44] with a simplified local interaction. Both analyses find a $\mathcal{C}_2\mathcal{T}$-breaking gapped state with Chern number $C = \pm 1$ as the lowest-energy solution.

References [14,22] also predict an interaction-induced gapped phase at charge neutrality. However, while certain parameter regimes again return a $\mathcal{C}_2\mathcal{T}$-breaking state with nonzero Chern number, other regimes yield a $\mathcal{C}_2\mathcal{T}$-preserving, trivial insulator. The latter statement may seem at odds with our assertion in Sec. II A that $\mathcal{C}_2\mathcal{T}$ protects the masslessness of the Dirac cones, but this protection only holds when the flat bands are energetically isolated. In the calculations of Refs. [14,22], interactions close the gap separating the flat and dispersive bands, thus negating the protection conferred upon the Dirac cones by $\mathcal{C}_2\mathcal{T}$ symmetry. It is worth noting that STM measurements show that the flat bands indeed remain isolated as a function of filling [15–18], and yet still resolve correlation effects. We therefore view the formation of the $\mathcal{C}_2\mathcal{T}$-symmetric insulator as a less likely outcome.

Returning to the $\mathcal{C}_2\mathcal{T}$-breaking gapped states, we remark that from the perspective of the Dirac theory, it is natural to expect the phase with $C = \pm 1$ to be energetically favorable relative to the $\mathcal{C}_2\mathcal{T}$-breaking trivial insulator with $C = 0$. Recall from Sec. II B that the mean-field order parameter for the $C = \pm 1$ state is $\psi^\dagger\eta^z\psi$, which in principle can arise from a momentum-independent microscopic perturbation. The trivial $C = 0$ phase instead corresponds to an order parameter $\psi^\dagger\eta^z\tau^z\psi$ that yields opposite-sign masses for the Dirac cones at $\pm\boldsymbol{\kappa}$—and hence cannot arise from a momentum-independent microscopic perturbation. In monolayer graphene the converse situation arises: the momentum-independent staggered sublattice potential generates a trivial insulator whereas the relatively baroque, momentum-dependent Haldane mass [61] is instead required to enter a $C = \pm 1$ phase. (This distinction reflects the fact that the Dirac cones at $\pm\boldsymbol{\kappa}$ exhibit the same chirality in mTBG, while the Dirac cones at $\pm\boldsymbol{K}$ in monolayer graphene have opposite chirality [46]). Spontaneously generating a Haldane mass in monolayer graphene is thus unnatural—see, e.g., Ref. [62]—and it is analogously difficult to spontaneously enter the $C = 0$ phase in mTBG.

C. Inclusion of spin and $\boldsymbol{K}$-valley flavors

We have so far argued that the single-flavor version of interacting, charge-neutral mTBG prefers to enter a gapped phase with Chern number $C = \pm 1$. Inclusion of spin and $\boldsymbol{K}$-valley degrees of freedom not only allows for many distinct possible phases depending on the Chern numbers assigned to each sector, but further allows for additional phases that do not naturally descend from the single-flavor theory. Let us begin by discussing the former.

We specifically focus on four natural candidate insulators that we refer to as quantum valley Hall (QVH), quantum spin-valley Hall (QSVH), quantum Hall (QH), and quantum spin Hall (QSH) phases. Figure 4 depicts these states along with their corresponding mass terms. These insulators carry different symmetry properties as summarized in Table I. While all four phases break $\mathcal{C}_2\mathcal{T}$ symmetry, they do so in different ways: the QVH and QSVH states break $\mathcal{C}_2$ while preserving time reversal $\mathcal{T}$, whereas the converse is true of the QH and QSH states. They are further distinguished by the action of

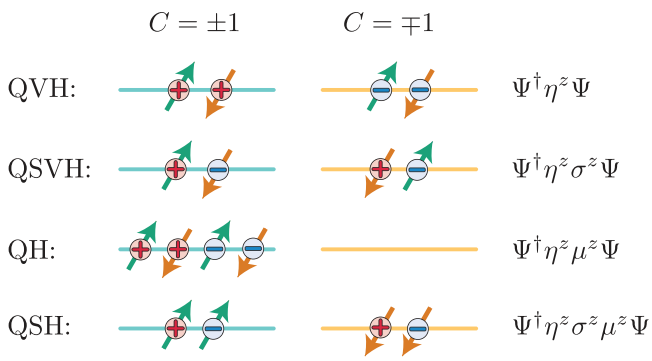

FIG. 4. Four natural $\mathcal{C}_2\mathcal{T}$-breaking insulators at charge neutrality. In order from top to bottom: quantum valley Hall (QVH), quantum spin-valley Hall (QSVH), quantum Hall (QH), and quantum spin Hall (QSH). The direction of the arrow indicates spin, while the sign, + or −, labeling the arrow indicates the $\boldsymbol{K}$-valley.

the $\mathrm{SU}(2)_s$ spin symmetry, which is preserved (broken) by the QVH and QH (QSVH and QSH) states. Note that because of the additional $\boldsymbol{K}$-valley flavor index, our QSH state differs from the 2D topological insulator realized, e.g., in the Kane-Mele model [63]. We nevertheless adopt this nomenclature since the state breaks spin-rotation symmetry and preserves the physical electronic time reversal operation $\mathcal{T}_{\mathrm{elec}} \equiv i\sigma^y\mathcal{T}$ that obeys $\mathcal{T}_{\mathrm{elec}}^2 = -1$.

It is useful to highlight some physical differences between these states and thus their compatibility with experimental observations. Neither the QVH nor QSVH insulator is expected to possess gapless edge modes at a sample boundary. Our discussion has made significant use of the approximate $\mathrm{U}(1)_v$ valley symmetry, but this symmetry is violently broken by the edge itself, which naturally occurs on the microscopic length scale a of the underlying graphene monolayers. As a result, the edge modes from the two valleys scatter strongly, resulting in a purely insulating state. By contrast, the QH state hosts robust gapless edge modes that are completely immune from scattering by virtue of their chirality. Edge modes of the QSH insulator, while nonchiral, are nevertheless also robust since backscattering at a sample boundary must be accompanied by a spin flip. The sample studied by Lu *et al.* [14] displayed insulating transport with no signs of edge conduction. Among the four insulators, QH and QSH states thus appear unlikely, at least in that platform.

As a result of the separation of scales between $\boldsymbol{K}$ valleys and the $\mathrm{SU}(2)_s$ symmetry, all four insulating states have very similar energies. In Appendix D, we compare the QVH ground-state energy against the other three insulators using a simple Hartree-Fock variational approach. We show that all four states are exactly degenerate in the chiral model [64], a version of the continuum model that possesses an exact particle-hole symmetry that renders it exactly solvable. Nevertheless, for more realistic versions of the continuum model (where particle-hole symmetry is absent), we find that the QVH state is actually disfavored relative to the other insulators. However, when computed numerically, we find the energy difference to be extremely small, less than $\sim 10^{-5}$ meV per electron, implying that the explicit breaking of particle-hole symmetry has little effect.

TABLE I. Symmetry-breaking pattern of the four topological states. Note that the QSH phase violates $\mathcal{T}$ but preserves the physical electronic time reversal operation $\mathcal{T}_{\mathrm{elec}} = i\sigma^y\mathcal{T}$.

	$\mathcal{T}$	$\mathcal{C}_2$	$\mathrm{SU}(2)_s$
QVH	✓	✗	✓
QSVH	✓	✗	✗
QH	✗	✓	✓
QSH	✗	✓	✗

We turn now to alternative phases. Polarized phases—for which the flat bands of two flavors are fully occupied—represent one class of competing ground states. In general, both spin- and valley-polarized phases are degenerate at charge neutrality when H_C' [Eq. (13)] is neglected [65,66]. Liu *et al.* [21] find that, within the chiral model, these polarized states have identical Fock energies to the $\mathcal{C}_2\mathcal{T}$-breaking insulators with nontrivial Chern number in each flavor. They also obtained self-consistent versions of these solutions numerically using a more realistic version of the continuum model; while no longer exactly degenerate, these states remained close in energy. Adding explicit $\mathcal{C}_3$-breaking strain—as observed in multiple STM experiments [15–17]—was, however, found to promote the $\mathcal{C}_2\mathcal{T}$-breaking insulators over the polarized states. Another proposed state is the intervalley coherent phase (IVC) [6], which spontaneously breaks $\mathrm{U}(1)_v$ symmetry by coupling the $+\boldsymbol{K}$ and $-\boldsymbol{K}$ bands. General considerations [65] as well as calculations using the analytically tractable chiral model [21] indicate that IVC order is disfavored at the Hartree-Fock level. Other numerics nevertheless challenge these conclusions [67].

Importantly, among the gapped phases discussed here, only the QVH order parameter directly couples to disorder that is smooth and preserves $\mathcal{T}$ and $\mathrm{SU}(2)_s$ spin symmetry. Time reversal and spin symmetry forbid coupling to the order parameters for QSVH, QH, QSH, and polarized phases, whereas smoothness of disorder prohibits coupling to an IVC order parameter. Hence, even if one of the latter states is energetically favorable in a perfectly clean system, the unavoidable presence of inhomogeneity in any physical sample may nevertheless stabilize the QVH phase, a possibility that we explore in Sec. V C.

V. INTERPLAY OF INTERACTIONS AND DISORDER

We are now in a position to explore the fate of charge-neutral mTBG in the presence of interactions *and* smooth disorder. Let us first recapitulate the expected behavior in the disordered, noninteracting limit (Sec. III) and in the clean but strongly interacting regime (Sec. IV):

(1) In the absence of interactions, disorder localizes the massless Dirac fermions when any form of inter-$\boldsymbol{\kappa}$-valley scattering is present in a manner that is formally analogous to physics of monolayer graphene. However, while monolayer graphene only requires that disorder be long ranged on the scale of the microscopic lattice to avoid localization, disorder must be long ranged on the scale of the moiré lattice to suppress localization in twisted bilayer graphene.

(2) We have argued that in the strongly interacting, clean limit, the QVH phase that spontaneously breaks $\mathcal{C}_2$ symmetry constitutes (at the very least) an energetically competitive state that is compatible with experimental observations. Moreover, we observed that among various other candidate ground states, QVH order uniquely couples to smooth disorder respecting spin and time-reversal symmetries.

To simultaneously incorporate interactions and disorder below, we start with the assumption that the QVH state is the true ground state of the clean, interacting Hamiltonian. We construct an Ising formulation of the system in the presence of disorder, which allows us to systematically consider the crossover between the first and second panels of Fig. 2. We discuss the titular recovery of the massless Dirac cones before showing that even when the QVH insulator is not the true ground state in the clean theory, disorder may nevertheless tip the balance back in its favor. We close with some comments on the eventual localization of the Dirac fermions, as illustrated in the final panel of Fig. 2.

A. Ising model formulation and domain formation

Suppose that the interaction energy scale dominates the physics, preferring to spontaneously break the $\mathcal{C}_2$ symmetry and form a QVH insulator. Disorder terms that do not couple to the QVH order parameter can then be neglected, leaving only the random field $\mathcal{M}_0(\boldsymbol{r})$ that couples to $\psi^\dagger\eta^z\psi$ in Eq. (9) (or, in the full theory, a random scalar field that couples to $\Psi^\dagger\eta^z\Psi$). A random mass cannot produce localization but does compete against long-range order. In fact, we show that even when disorder is weak and uncorrelated, the system always loses long-range order in the thermodynamic limit due to the formation of domains, as sketched in the central panel of Fig. 2. Destruction of long-range order only becomes observable, however, once the linear extent of the system, L, exceeds the typical domain size, ξ_{dom}. The goal of this subsection is to demonstrate that ξ_{dom} is finite and to determine its size as a function of the physical parameters of the theory.

We approach the problem in the standard fashion, via the formulation of a Landau-Ginzburg theory. The order parameter for the $\mathcal{C}_2$ symmetry breaking is simply an Ising field ϕ obtained by coarse graining the bilinear $\psi^\dagger\eta^z\psi$, i.e.,

$$\phi(\boldsymbol{r}) \sim \int_{\boldsymbol{r}'\in R(\boldsymbol{r})} \psi^\dagger\eta^z\psi(\boldsymbol{r}') \sim \ell_{\text{UV}}^2\psi^\dagger\eta^z\psi(\boldsymbol{r}), \tag{14}$$

where $R(\boldsymbol{r})$ is a spatial region centered at $\boldsymbol{r}$ of typical size ℓ_{UV}^2 and ℓ_{UV} is an ultraviolet cutoff quantified below. Since we are interested in the physics deep within the ordered phase with $\langle\phi\rangle \neq 0$, a *classical* Ising model suffices:

$$H_{\text{Ising}} = \int_{\boldsymbol{r}}\left[\mathcal{K}(\nabla\phi)^2 + \frac{r}{2}\phi^2 + \frac{u}{4!}\phi^4\right]. \tag{15}$$

The mass r is clearly assumed to be negative.

The scales of the original fermionic Hamiltonian ultimately determine parameters of the Ising model, though this assignment is not necessarily straightforward. Consider first $\mathcal{K}$. Since ϕ is dimensionless, $\mathcal{K}$ has units of energy, and hence $\mathcal{K} \sim U$, with U a characteristic energy scale of the system. Both the rough estimate for the Coulomb potential, $V(a_M) \approx 14$ meV, given at the end of Sec. IV A, and the experimentally measured transport gap at charge neutrality, $\Delta_{\text{CNP}} \approx 1$ meV [14], provide natural candidates for U. Given uncertainties in our calculation of $V(a_M)$ related to screening from other bands, we view the latter option as a more reasonable and conservative estimate. We stress, however, that this choice has little direct bearing on the discussion that follows.

It is also important to assign a length scale to the interactions and hence the Ising theory. Since our primary goal is to describe domain-wall physics, the most natural scale is

$$\xi_{\text{int}} \sim \frac{\hbar v_F}{\Delta_{\text{CNP}}}, \tag{16}$$

which corresponds to the decay length of a Dirac fermion of mass $\Delta_{\text{CNP}}/v_F^2$. In our context, these fermions are the chiral modes that bind to the domain walls at which the Chern numbers for each flavor change signs, identifying ξ_{int} as the domain boundary width. Any physics occurring on scales smaller than ξ_{int} necessarily includes these fermionic degrees of freedom, and hence lies outside our Ising formulation's regime of validity. The interactions length scale therefore defines a UV cutoff.[4] As a consistency check, we must verify that ξ_{int} exceeds the moiré lattice constant, $a_M \approx 12.8$ nm. Inserting $v_F \approx 0.15 \times 10^6$ m/s [2] and $\Delta_{\text{CNP}} \approx 1$ meV [14] into Eq. (16), we indeed find $\xi_{\text{int}} \approx 100\,\text{nm} \sim 10\,a_M$. We are therefore permitted to set $\ell_{\text{UV}} \sim \xi_{\text{int}}$. In turn, dimensional analysis gives $r,\ u \sim U/\xi_{\text{int}}^2$.

Because disorder breaks $\mathcal{C}_2$, it should couple to the Ising field in a manner that breaks the $\mathbb{Z}_2$ Ising symmetry. In other words, disorder appears as a random "magnetic" field,

$$H_{\phi,\text{dis}} = \int_{\boldsymbol{r}} \mathcal{B}(\boldsymbol{r})\phi(\boldsymbol{r}), \tag{17}$$

where $\mathcal{B}(\boldsymbol{r}) \sim \int_{\boldsymbol{r}'\in R(\boldsymbol{r})} \mathcal{M}_0(\boldsymbol{r}')/\xi_{\text{int}}^4$. The random field $\mathcal{M}_0$ is defined by the disorder strength δm, correlation length ξ_{dis}, and correlation function $K(\boldsymbol{r}/\xi_{\text{dis}})$ (in the notation of Sec. III B, these quantities correspond to $g_{\mathcal{M}_0}$, $\xi_{\mathcal{M}_0}$, and $K_{\mathcal{M}_0}$, respectively). We focus on the situation where the disorder is Gaussian correlated: $K(\boldsymbol{r}/\xi_{\text{dis}}) = e^{-\boldsymbol{r}^2/(2\xi_{\text{dis}}^2)}$. Our assertion that the interaction energy scale dominates the disorder energy scale can now be more precisely stated as $\delta m/U \ll 1$.

In summary, the Hamiltonian controlling the ordering of ϕ is $H_{\text{RFIM}} = H_{\text{Ising}} + H_{\phi,\text{dis}}$, which is none other than the much-studied random-field Ising model (RFIM)[5] [69–71]. As claimed, the RFIM in 2D is generically disordered [72,73], and so ξ_{dom} is finite. The mechanism of domain formation depends largely on the magnitude of the ratio:

$$\alpha = \frac{\delta m}{U}\frac{\xi_{\text{dis}}}{\xi_{\text{int}}}. \tag{18}$$

This result and the scenarios we outline below are derived and further explained in Appendix E.

[4] The definition of ξ_{int} and ℓ_{UV} is largely independent of our choice of U.

[5] This theory and its derivation should not be confused with the fact that a *free* Dirac fermion with random mass disorder maps onto the random *bond* Ising model [49,68].

We first examine what occurs when $\alpha \gtrsim 1$. Since $U/\delta m$ is already presumed large, for α to be larger than unity, this limit corresponds to that of extremely smooth disorder: $\xi_{\rm dis}/\xi_{\rm int} \gg 1$. In this scenario, the energy gained by having ϕ align in the direction preferred by $\mathcal{B}(\boldsymbol{r})$ is larger than the interaction energy cost associated with the misalignment of ϕ along the domain boundary. The Ising field therefore directly tracks the disorder potential, implying that

$$\xi_{\rm dom} \sim \xi_{\rm dis}, \quad \alpha \gtrsim 1. \tag{19}$$

The situation is more subtle when $\alpha \lesssim 1$. With stronger interactions, we naturally expect larger domains. At some point, the domains are large enough that the correlated nature of the disorder is washed away, allowing us to treat it as white noise: $\overline{\mathcal{M}_0(\boldsymbol{r})\mathcal{M}_0(\boldsymbol{r}')} \cong \delta m^2 \xi_{\rm dis}^2 \delta^2(\boldsymbol{r}-\boldsymbol{r}')$. In this case, the destruction of long-range order occurs through the condensation of domain walls. An evaluation of the domain-wall roughening yields a lower bound for their size of [72]

$$\xi_{\rm dom} \lesssim \max(\xi_{\rm int}, \xi_{\rm dis})\, e^{c/\alpha^2}, \quad \alpha \lesssim 1, \tag{20}$$

where $c \sim \mathcal{O}(1)$ is a nonuniversal constant. We can verify that when $\alpha \ll 1$ the domain length scale is indeed far greater than the disorder correlation length, i.e., $\xi_{\rm dom} \gg \xi_{\rm dis}$.

B. Recovery of massless Dirac fermions

Next we discuss the physical consequences of the Ising model outlined above in the regime where the system size L exceeds the typical domain size $\xi_{\rm dom}$. For now, we continue to assume suppression of both inter-$\boldsymbol{\kappa}$- and inter-$\boldsymbol{K}$-valley scattering. At least close to the crossover scale $\xi_{\rm dom}$, the Ising formulation should remain valid: the system is characterized by multiple domains of opposing Chern numbers with typical size $\xi_{\rm dom}$, as the central panel of Fig. 2 illustrates. In this regime, the system can be described by eight independent Chalker-Coddington network models [37]—one for each of the two Dirac cones within the four spin/valley sectors. As mentioned briefly in the Introduction and more fully in Sec. III C, each network model may be mapped directly onto that of a single gapless Dirac cone [40,41], thus giving the promised restoration of massless Dirac fermions from a strongly correlated starting point.

We can alternatively motivate the recovery of massless Dirac cones without relying on network models. Let us return to the full disordered Dirac theory described by Eqs. (4) and (5), which includes all spin and valley degrees of freedom. Notably, here we additionally allow for weak intervalley scattering terms. Upon including strong correlations at a mean-field level, interactions dramatically enhance the effective strength of the random field that couples to the QVH order parameter $\Psi^\dagger \eta^z \Psi$. All other disorder fields, by contrast, remain weak and can be neglected to a first approximation. The problem then reduces to a set of independent Dirac cones, each governed by the far simpler disorder Hamiltonian in Eq. (10) with *only* random mass disorder. As noted earlier, the random mass is a marginally irrelevant perturbation to the clean Dirac theory when it is the sole source of disorder [49]. Massless Dirac fermions thus naturally re-emerge from this viewpoint as well.

At sufficiently low energy scales, however, the additional disorder fields neglected above eventually kick in. The dominant corrections are expected to arise from *intra*-$\boldsymbol{\kappa}$-valley scattering processes, encoded by the vector- and scalar-potential terms in Eq. (10). When these terms are also present, the theory is believed to flow to the IQH plateau transition, which is characterized by a finite a density of states with both universal longitudinal and Hall conductances (here, valley Hall). At still lower energy scales, inter-$\boldsymbol{\kappa}$-valley scattering is expected to produce localization, as Sec. V D discusses in more detail.

Nevertheless, the perspective just outlined should be viewed as a consistency check and not a proof of concept. Crucially, it cannot account for the energy scales separating the Dirac fermions of the clean, noninteracting mTBG system from the recovered Dirac cones of the interacting, disordered network model.

C. Competing phases

So far in this section, the QVH insulator has been taken as the true ground state of mTBG at charge neutrality, even in the absence of disorder. We now address the possibility that interactions prefer a different state. To simplify the problem, we consider the situation in which a single competing phase is energetically favorable relative to the QVH insulator. In accordance with the conventions of Sec. V A, this competition can be quantified through the energy difference $\delta\epsilon$ in an area of size $\ell_{\rm UV}^2 = \xi_{\rm int}^2$:

$$\frac{\delta\epsilon}{\xi_{\rm int}^2} \equiv \mathcal{E}_{\rm QVH} - \mathcal{E}_{\rm C} \geqslant 0, \tag{21}$$

where $\mathcal{E}_{\rm QVH}$ and $\mathcal{E}_{\rm C}$, respectively, denote the ground-state energy densities of the QVH state and competing phase. We further assume that the competing order may be described by an Ising field, Φ, that does not linearly couple to any disorder field; recall the discussion at the end of Sec. IV C. Generalizing our arguments to include continuous order parameters (as needed for the QSH and QSVH insulators) is straightforward, and we therefore leave the competing phase's identity unspecified. Note, however, that properties of domain walls separating the QVH order and the competing phase may depend on the precise nature of the latter.

We again work in a regime where strong interactions obviate the need to include all disorder fields save for the random mass $\mathcal{M}_0$ that linearly couples to the QVH order parameter via Eq. (17). Importantly, this type of disorder locally promotes the QVH state by lowering its energy relative to the competing phase, even though—as we saw earlier—it generally destroys true long-range order. When $\delta\epsilon$ is small enough, we expect the majority of the sample to realize the QVH phase and the scenario outlined in the previous section to hold. In terms of the Ising theory devised in the previous section, we can express this condition as

$$\left[\frac{1}{\rm vol}\int_{\boldsymbol{r}} \langle \phi^2(\boldsymbol{r})\rangle\right]^{1/2} \gtrsim \frac{1}{2}, \tag{22}$$

where vol denotes the sample volume. When this equation holds, we say the system is ϕ ordered; otherwise, the system is Φ ordered.

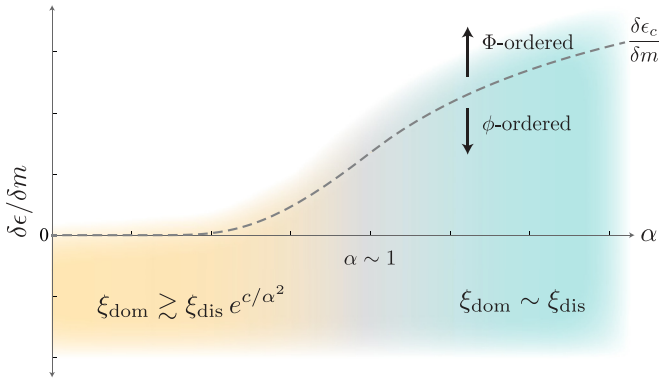

FIG. 5. Schematic phase diagram as a function of disorder, α, and ground-state energy difference. The dashed line indicates the "critical" energy difference $\delta\epsilon_c(\alpha)$ that characterizes the crossover from samples that are primarily ϕ-ordered to those that are primarily Φ-ordered. There are two distinct ϕ-ordered regimes. In the blue region, $\alpha \gtrsim 1$, $\langle\phi\rangle$ tracks the disorder so that domains are of the same size as the disorder correlation length, $\xi_{\rm dom} \sim \xi_{\rm dis}$. Conversely, in the orange region, $\alpha \lesssim 1$, the correlated nature of the disorder is unimportant, and domains are exponentially large, $\xi_{\rm dom} \gtrsim \xi_{\rm dis} e^{c/\alpha^2}$ [here, we assume that $\xi_{\rm dis} > \xi_{\rm int}$; see Eq. (20)]. In the white region above the dashed line, the competing phase prevails and $\langle\Phi\rangle \neq 0$ throughout most of the sample.

Appendix F explores this problem in depth, ultimately deriving the schematic phase diagram shown in Fig. 5. We again find that the primary control parameter is the ratio α [recall Eq. (18)] corresponding to the horizontal axis. Motivated by the notion that Φ-ordered regions may be viewed as annealed vacancies, we begin with a dilute Ising-model description. At the lattice level, the theory is conveniently formulated by promoting the Ising variables $\sigma^z = \pm 1$ to three-state spin-1 variables s, where $s = \pm 1$ corresponds to the two QVH phases and $s = 0$ corresponds to the competing phase. We present a simple mean-field solution to the classical Blume-Capel model for these spin-1 degrees of freedom [74–77] in Appendix F 1. While the phase diagram we obtain resembles the one shown in Fig. 5 in many respects, it erroneously predicts long-range ϕ-order when $\delta\epsilon < 0$ and disorder is sufficiently small, $\alpha \lesssim 1$; as discussed in Sec. V A and Appendix E, in reality, long-range order is unstable to the addition of any finite disorder. This failure of mean-field theory is not unprecedented given the low dimensionality.

In Appendix F 2, we therefore return to the Imry-Ma-type arguments of Sec. V A (see also Appendix E), which allow us to derive a "critical" energy difference $\delta\epsilon_c(\alpha)$ that characterizes the crossover scale separating ϕ- and Φ-ordered regimes. We plot $\delta\epsilon_c(\alpha)$ with a dashed line in Fig. 5. In the white region above the line, $\delta\epsilon \gtrsim \delta\epsilon_c(\alpha)$, the competing phase is realized throughout the majority of the system, and the network picture we propose is no longer relevant. Conversely, Eq. (22) holds in regions where $\delta\epsilon \lesssim \delta\epsilon_c(\alpha)$ (including the trivial case, $\delta\epsilon < 0$, where QVH states minimize the energy in the clean limit). Just as we found above, depending on the strength of disorder, the destruction of long-range QVH order occurs in two fashions. In Fig. 5, the parameter regime where $\langle\phi\rangle$ tracks the disorder field is shown in turquoise. The orange area indicates the opposite limit, where long-range order is eliminated by domain-wall condensation. The intermediate regime where $\alpha \sim 1$ is shown in neutral grey.

Notably, these considerations imply that disorder may not only be responsible for selecting which QVH order is locally realized, but that it may also determine whether or not QVH order is realized at all. In particular, our proposal admits a scenario in which the clean samples of Lu *et al.* [14] are Φ-ordered, while the less homogeneous samples of Cao *et al.* [1,2] and Yankowitz *et al.* [9] realize the QVH network picture displayed in the central panel of Fig. 2—even supposing that the two sets of systems differ *solely* in the amount of disorder they present.

D. Localization

In the absence of any special symmetries, all two-dimensional systems are generically expected to localize in the thermodynamic limit, and our platform is no exception. Localization is likely irrelevant for the previously studied mTBG samples, whose linear dimensions are $\sim 2 - 8\,\mu$m $\sim 150 - 600 a_M$. It is nevertheless instructive to briefly discuss localization within our proposed scenario. The precise manner in which localization occurs in the presence of interactions poses a notoriously difficult and subtle problem that we will not wade into in detail. Rather, our goal is to discuss some general features of the problem that can be deduced given some reasonable simplifying assumptions.

When discussing localization, one can imagine either increasing the system size or increasing the disorder strength. In the latter case, the situation rapidly becomes unwieldy: as the disorder strength approaches the interaction energy ($\delta m/U \to 1$) or the disorder correlations become ultra-short-ranged ($a_M/\xi_{\rm dis} \to 1$), our Ising formulation breaks down. By contrast, the Ising-model perspective remains valid when we instead consider progressively larger samples with an otherwise identical set of parameters. Interactions can still, of course, pose complications; for instance, in the network-model picture, localization involves a network of gapless domain-wall modes that generically form Luttinger liquids [78]. We do not address such subtleties, instead postulating that the primary effect of interactions is to catalyze the spontaneous breaking of $\mathcal{C}_2\mathcal{T}$.

The most straightforward manner by which the re-emergent Dirac fermions can localize is through inter-$\boldsymbol{\kappa}$-valley scattering. Such scattering events can also localize the original Dirac cones that appear in the free-fermion band structure for mTBG, but the physics is not quite identical: the network picture underlying the re-emergent Dirac cones effectively postpones localization by renormalizing the UV scale at which it occurs. That is, if $\xi_{{\rm loc},fr}$ is the localization length in the free case, we have $\xi_{\rm loc} \sim \xi_{\rm dom}\xi_{{\rm loc},fr}/a_M$ with interactions. One can intuitively understand this rescaling from the perspective of the gapless domain-wall modes in the network model. As Fig. 6 illustrates, in a given $\boldsymbol{K}$ valley, the domain-wall modes corresponding to $\pm\boldsymbol{\kappa}$ copropagate, and hence non-forward-scattering processes can only occur at nodes where multiple domain walls intersect (see red arrows).

Inter-$\boldsymbol{K}$-valley scattering can also prompt localization [25]. Disorder coupling the two $\boldsymbol{K}$-valleys has so far been completely ignored since it is exponentially suppressed relative

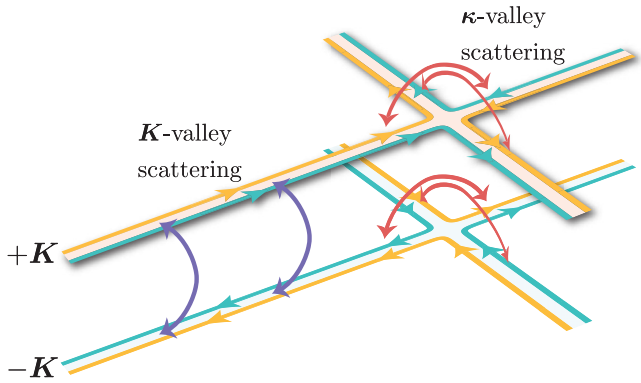

FIG. 6. Edge modes of the $+\boldsymbol{K}$ and $-\boldsymbol{K}$ valley sectors at a node connecting four domains. The orange and turquoise arrows represent the chiral modes at the domain boundaries. Red arrows at the node indicate $\mathrm{U}(1)_v$-preserving, inter-$\boldsymbol{\kappa}$-valley scattering processes, which result from inhomogeneities at the moiré lattice scale, $a_M \approx 12.8$ nm. The $\mathrm{U}(1)_v$-breaking inter-$\boldsymbol{K}$-valley scattering events are indicated by the purple arrows. While this type of scattering is exponentially suppressed [see Eq. (8)], it can occur at any point along a domain boundary.

to intra-$\boldsymbol{\kappa}$-scattering [see Eq. (8)]. However, inter-$\boldsymbol{K}$-valley scattering can occur at any point along the domain walls, as illustrated in Fig. 6, making it a fundamentally one-dimensional process. For very large domains, such intra-domain-wall scattering thus inevitably becomes the dominant localization mechanism. The localization length is then expected to be proportional to the mean free path of the domain-wall modes [79], which is $\xi_{\mathrm{loc}} \sim \hbar v_F / g_{KK'} \sim \hbar v_F e^{4\pi^2 \xi_{\mathrm{dis}}^2 / a^2} / g$, and hence an exponentially large function of the disorder correlation length.

VI. DISCUSSION

We have presented a theory that reconciles the seemingly conflicting experiments on charge-neutral mTBG by invoking a nontrivial interplay between strong interactions and weak disorder. In our proposed picture, uniform order (QVH or otherwise) is realized throughout ultrahomogeneous samples, like those of Lu *et al.* [14], whereas QVH domains with opposite spin/valley Chern numbers appear in systems with more disorder, like the experiments of Cao *et al.* [1,2] and Yankowitz *et al.* [9]. In the latter samples, gapless edge modes at domain boundaries form a network that may be mapped onto a theory of massless Dirac fermions, thereby explaining their semimetallic transport measurements. By contrast, since a physical sample boundary strongly breaks the $\mathrm{U}(1)_v$ symmetry protecting the edge modes, a uniformly ordered QVH state is an insulator at charge neutrality, in agreement with the observations of Lu *et al.* Both sample classes exhibit a local gap determined by the interaction strength—in harmony with STM experiments [15–18].

The network model outlined in this paper is somewhat reminiscent of proposals aimed at describing minimally twisted bilayer graphene (minTBG) [80–83]. When $\theta \lesssim 1^\circ$, it becomes energetically favorable for the microscopic lattices to distort such that the AB and BA regions occupying the moiré honeycomb sites enlarge at the expense of the AA regions situated at the center of each moiré hexagon [42,84–89]; see Fig. 3(a) for an illustration of the undistorted case. Under the application of a displacement field, AB and BA regions develop QVH order with opposing Chern numbers [29–31], yielding four edge modes per spin at the AB/BA boundaries. While both our theory for mTBG and the theory proposed for minTBG are built on network models comprised of QVH domains, there are key qualitative distinctions that we wish to underscore. The local QVH order in minTBG arises entirely as a single-particle effect, whereas the development of QVH order in our scenario relies principally on strong interactions. Moreover, the shape and size of the AB and BA regions in minTBG are fixed; together, they comprise a single moiré unit cell. The QVH domains discussed in this paper instead result from the smooth disorder background and typically extend over many moiré unit cells.

Our proposal is supported by available experimental data and crucially can be further tested in future experiments. One natural direction is to employ large-area STM scans to locally probe both gapped domains *and* gapless domain-wall modes. (To our knowledge, evidence of the latter in mTBG has not yet been reported in the literature). Samples that are simultaneously amenable to STM and transport would offer additional insight; for instance, the presence of gapless domain-wall modes should correlate with semimetallic transport, whereas such modes should be absent in homogeneous insulating samples. Some caveats are warranted, however. First, discussions of local phenomena in STM measurements are often complicated, e.g., by disorder- or tip-induced localized states, and it may be difficult to unambiguously distinguish the domain physics we propose from such effects. Additionally, the samples studied by Refs. [1,2,9,14] are enclosed on both sides by hBN, preventing STM study. The nature of transport in mTBG with hBN only on the bottom side, as in the samples studied by STM in Refs. [15–18], poses an interesting open question.

One can also investigate our scenario entirely within transport [90]. Consider a single mTBG sample etched into a series of strips of varying widths w, as shown in Fig. 7(a). Transport through a given strip depends sensitively on the value of w relative to the typical domain size, ξ_{dom}. When $w \gg \xi_{\mathrm{dom}}$—the limit presumably relevant to the experiments of Refs. [1,2,9]—semimetallic transport should occur. In the opposite limit, $w \ll \xi_{\mathrm{dom}}$, no edge modes connect the contacts and the strip should appear insulating. This experiment may be modified to preclude possible variations in the conductivity resulting from intrinsic variations between the strips, such as their local twist angle. Instead of physically cutting the sample, a strip can be electrostatically generated through spatially varying gate voltages: Within a channel of width w, the system is locally tuned to charge neutrality, whereas elsewhere the Fermi energy is tuned to lie within the gap separating the flat and dispersive bands. One could then study the conductivity as a function of width w for all regions within the sample. Figure 7(b) illustrates this refined version of the experiment.

Finally, although we have emphasized the validity of our proposal even in the eventuality that the clean system does not realize a QVH insulator, the observation of this state at charge neutrality in ultraclean, insulating samples would of course serve as support for our scenario. Unquantized valley Hall currents have been observed through nonlocal

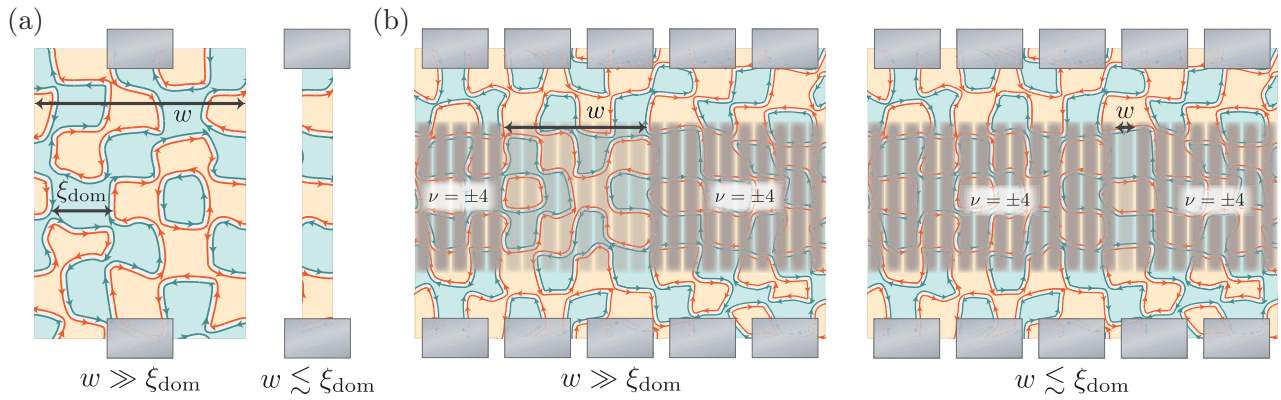

FIG. 7. Schematic illustration of a proposed transport experiment. The domain structure of one of the valley sectors is shown: regions carrying Chern number $C = +1$ are depicted in blue, whereas those carrying $C = -1$ are depicted in orange. The other valley sector is not shown explicitly. The grey rectangles above and below the sample indicate contacts through which the conductance is measured. (a) A single mTBG sample is sliced into multiple subsystems of varying width w. When $w \gg \xi_{\rm dom}$, as shown on the left, Dirac-like conductance is observed. Once the width is smaller than the typical domain size, $w \lesssim \xi_{\rm dom}$, the sample appears insulating, as shown on the right. (b) An alternate experiment in which the mTBG sample remains intact. The taupe rectangles aligned in a row along the center of the sample represent individually tunable gates through which the chemical potential may be locally varied. In regions where these gates are opaque, the chemical potential lies within the superlattice band gap, i.e., the flat bands are either completely empty or full ($\nu = \pm 4$). The system is tuned to charge neutrality in all other regions (either the gates are transparent or no gates are shown). On the left, $w \gg \xi_{\rm dom}$, and a semimetallic conductance should be observed. Conversely, since $w \lesssim \xi_{\rm dom}$, a large resistance is expected on the right.

resistance measurements [91] in hBN-aligned graphene [26] and Bernal stacked graphene with an out-of-plane electric field [27–29], and the realization of the QVH insulator in TBG at charge neutrality could therefore be demonstrated through an analogous set of measurements. The application of a small magnetic field parallel to the sample represents a more common probe of the correlated insulating states in twisted bilayer graphene. Because the QVH state is a spin singlet, this field is unlikely to have a substantial effect. Nevertheless, ignoring both interactions and modifications to the interlayer tunneling due to the magnetic field, a mean-field perspective would suggest that the gap at charge neutrality should decrease with increasing magnetic field.

Our proposal also spotlights various other avenues for future study. The fate of the network under an applied magnetic field poses a particularly interesting problem. One possibility is that the magnetic field simply stabilizes a different competing phase, thereby destroying the network. If this transition occurs at fields strengths close to or below 1 T, where quantum oscillations are first clearly resolved, the re-emergent Dirac theory is unlikely to produce observable Landau-fan phenomena. The occurrence of such a transition is neither necessary nor expected, however. In the case where the QVH network survives a broader magnetic-field window, there are two limits to consider. When the magnetic length ℓ_B far exceeds the typical domain size $\xi_{\rm dom}$, quantum oscillations are expected to be insensitive to the re-emergent nature of the fermions, implying that a Landau fan corresponding to massless Dirac fermions should be observed at low fields. Given that the magnetic length is already quite large at 1 T, $\ell_B \approx 25\,{\rm nm} \approx 2a_M$, this regime may be difficult to access experimentally (recall that the UV cutoff for our network model was $\xi_{\rm int} \sim 100\,{\rm nm}$). The opposite limit, $\ell_B \ll \xi_{\rm dom}$, appears to be more subtle. It is conceivable that the gapless edge modes do not affect the quantum oscillations, resulting in a Landau level spectrum similar to that of *massive* Dirac cones associated with the gapped QVH domains. Alternatively, the system could exhibit physics reminiscent of the Hofstadter butterfly [92], though it seems likely that nonuniformity of the domain sizes may hinder any clear signal. Quantifying these issues could shed additional light on the experimental relevance of our scenario.

The role of interactions at the edges of the domains is another topic that we have not touched on. The edge modes may display interesting interacting phenomena that could be studied through the well-controlled bosonization formalism. In fact, Wu *et al.* [81] analyzed this problem in the context of the minimal twist angle samples described above. Further, while disorder is generically expected to localize the edges, the inclusion of interactions may have nontrivial consequences [78].

We have said little regarding transport away from the charge neutrality point. While its semimetallic nature dictates that the conductivity σ increase with doping, it can do so in different ways. If transport is ballistic, far enough away from charge neutrality, the conductivity should essentially track the density of states: $\sigma \propto \sqrt{|n|}$, where n is the electron density [93,94]. Provided inter-$\boldsymbol{\kappa}$-scattering is the most important form of disorder, we expect the mean free path of the network model to be rescaled, implying that ballistic transport may not be unreasonable. That is, letting $\ell_{{\rm mfp},fr}$ be the mean free path of the noninteracting Dirac fermions, we may postulate that the mean free path of the recovered network Dirac fermions is $\ell_{\rm mfp} \sim \xi_{\rm dom}\ell_{{\rm mfp},fr}/a_M$. On the other hand, in monolayer graphene, the linear dependence of the conductivity on density away from charge neutrality, $\sigma \propto |n|$, is largely ascribed to long-range Coulomb scattering [95–102]. While it seems unlikely that a similar mechanism would play an important role in mTBG, it is possible that twist-angle disorder (which can also be long-range) could have a similar effect [103].

Finally, exploring the interplay between interactions and disorder at other integer fillings constitutes perhaps the most interesting future direction. The charge-neutrality regime that we examined here offers the virtue that the system is almost insulating even at the band-structure level—thereby facilitating the study of (at least certain) correlated insulators. Accessing correlated insulating states at other fillings requires a far more drastic modification of the band fillings. Generalizing our analysis to such cases could provide valuable insight into the observed phenomenology of mTBG.

ACKNOWLEDGMENTS

We are grateful to Cory Dean, Arbel Haim, Eslam Khalaf, Stevan Nadj-Perge, Felix von Oppen, Seth Whitsitt, and Andrea Young for illuminating discussions. This work was supported by the Army Research Office under Grant Award No. W911NF-17-1-0323; the NSF through Grant No. DMR-1723367; the Caltech Institute for Quantum Information and Matter, an NSF Physics Frontiers Center with support of the Gordon and Betty Moore Foundation through Grant No. GBMF1250; the Walter Burke Institute for Theoretical Physics at Caltech; and the Gordon and Betty Moore Foundation's EPiQS Initiative, Grant No. GBMF8682 to J.A. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation Grant No. PHY-1607611.

APPENDIX A: CONTINUUM MODEL

We briefly outline the continuum model in this section. Spin indices are completely suppressed below. We first decompose the microscopic graphene operators as

$$\tilde{f}_\ell(\boldsymbol{r}) = e^{i\boldsymbol{K}\cdot\boldsymbol{r}} f_{+,\ell}(\boldsymbol{r}) + e^{-i\boldsymbol{K}\cdot\boldsymbol{r}} f_{-,\ell}(\boldsymbol{r}), \tag{A1}$$

where ℓ indicates both layer and sublattice. As discussed in Sec. II A, the continuum model Hamiltonian decouples into $\boldsymbol{K}$-valley sectors $H_{\text{cont}} = H_+ + H_-$, where $H_\pm$ act on $f_{\pm,\ell}$. For the moment, we consider H_+. We express $f_{+,\ell}$ as a vector $(f_{t,\text{A}}(\boldsymbol{r}), f_{t,\text{B}}(\boldsymbol{r}), f_{b,\text{A}}(\boldsymbol{r}), f_{b,\text{B}}(\boldsymbol{r}))$, where the $+$ has been dropped for convenience, t, b denote layer, and A, B denote sublattice. In this basis, H_+ acts as

$$H_+ = \begin{pmatrix} iv_0\boldsymbol{\eta}_{\theta/2}\cdot\boldsymbol{\nabla} & T(\boldsymbol{r}) \\ T^\dagger(\boldsymbol{r}) & iv_0\boldsymbol{\eta}_{-\theta/2}\cdot\boldsymbol{\nabla} \end{pmatrix}, \tag{A2}$$

where $\boldsymbol{\eta}_\phi = e^{-i\phi\eta^z/2}(\eta^x, \eta^y)e^{i\phi\eta^z/2}$ act on the sublattice space and $\boldsymbol{\nabla} = (\partial_x, \partial_y)$. The tunneling matrix $T(\boldsymbol{r})$ is given by

$$T(\boldsymbol{r}) = \sum_{\ell=1,2,3} t_\ell e^{-i\boldsymbol{q}_\ell\cdot\boldsymbol{r}}, \quad \boldsymbol{q}_\ell = \mathcal{R}_{2\pi(\ell-1)/3}[\boldsymbol{K}_t - \boldsymbol{K}_b] \tag{A3}$$

where $\mathcal{R}_\phi[\boldsymbol{v}]$ rotates the vector $\boldsymbol{v}$ by ϕ and the matrices t_ℓ are defined through

$$t_\ell = e^{2\pi i(\ell-1)\eta^z/3}\begin{pmatrix} w_0 & w_1 \\ w_1 & w_0 \end{pmatrix} e^{-2\pi i(\ell-1)\eta^z/3}. \tag{A4}$$

The physical parameters of the model are the twist angle θ, the velocity of the microscopic graphene layers v_0, and the tunneling amplitudes, w_0 and w_1. We take the angle to be close to the magic angle, $\theta = 1.05^\circ$, and the graphene velocity to be $v_0 = 9.1 \times 10^5$ m/s. The tunneling amplitudes are typically taken to be $(w_0, w_1) = (85, 110)$ meV [4]. However, for the chiral version [see Sec. D 5] of the model, we set $w_0 = 0$, keeping $w_1 = 110$ meV [64].

The Hamiltonian corresponding to the other valley, $\mathcal{H}_-$, may be obtained by acting time-reversal ($\mathcal{T}$) or by rotating by 180° ($\mathcal{C}_2$).

The continuum Hamiltonians maybe also be expressed in momentum space. Returning to second quantized notation, it may be written

$$H_\mu = \sum_{\boldsymbol{G},\boldsymbol{G}',\ell,\ell'} \int_{\boldsymbol{k}\in\text{BZ}} f^\dagger_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{G}) H^{(\mu)}_{\boldsymbol{G},\ell;\boldsymbol{G}',\ell'}(\boldsymbol{k}) f_{\mu,\ell'}(\boldsymbol{k}+\boldsymbol{G}'), \tag{A5}$$

where $\mu = +, -$ labels the $\boldsymbol{K}$ valley and the $\boldsymbol{G}$s are moiré reciprocal lattice vectors. Here, $H^{(\mu)}(\boldsymbol{k})$ may be thought of as an infinite matrix taking values within the moiré BZ with indices $(\boldsymbol{G}, \ell)$. It can be diagonalized through the unitary rotation

$$\begin{aligned} c^\dagger_{\mu,i}(\boldsymbol{k}) &= \sum_{\boldsymbol{G},\ell} u_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k}) f^\dagger_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{G}), \\ f^\dagger_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{G}) &= \sum_i u^*_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k}) c^\dagger_{\mu,i}(\boldsymbol{k}), \end{aligned} \tag{A6}$$

where i indexes the band. In terms of the $c_{\mu,i}(\boldsymbol{k})$ operators H_μ is

$$H_\mu = \sum_i \int_{\boldsymbol{k}\in\text{BZ}} c^\dagger_{\mu,i}(\boldsymbol{k})\epsilon_i(\boldsymbol{k}) c_{\mu,i}(\boldsymbol{k}). \tag{A7}$$

We note that invariance of $c_{\mu,i}(\boldsymbol{k})$ under shifts of $\boldsymbol{k}$ by a reciprocal lattice vector, $\boldsymbol{k} \to \boldsymbol{k} + \boldsymbol{G}$, implies $u_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k}+\boldsymbol{G}') = u_{\mu,i;\boldsymbol{G}+\boldsymbol{G}',\ell}(\boldsymbol{k})$.

APPENDIX B: SPIN AND TIME-REVERSAL SYMMETRIC BILINEARS

Here, we enumerate some of the symmetries of the Dirac theory. It is convenient to express them in terms of a large, unphysical, SU(8) symmetry generated by the $\boldsymbol{K}$-valley, $\boldsymbol{\kappa}$-valley, and spin symmetries. The generators of these symmetries are

$$\begin{aligned} &\text{SU}(2)_s: \quad (\sigma^x, \sigma^y, \sigma^z), \qquad \text{SU}(2)_\kappa: \quad (\tau^x, \tau^y, \tau^z), \\ &\text{SU}(2)_K: \quad (\mu^x\eta^y, \mu^y\eta^y, \mu^z). \end{aligned} \tag{B1}$$

Since the $\text{SU}(2)_K$ triplet does not take a particularly simple form, we define $\bar{\mu}^i = (\mu^x\eta^x, \mu^y\eta^x, \mu^z)$. Finally, the γ matrices are $\gamma^\mu = (\mu^z\eta^z, i\eta^y, -i\mu^z\eta^x)$. By combining the γ^μ, σ^i, τ^i, $\bar{\mu}^i$, we can generate all bilinears (pairing terms are not considered).

We are interested only in those bilinears that preserve the spin and time-reversal symmetries. Clearly, spin-conservation requires that disorder not couple to any bilinear containing σ^i, so we ignore it completely, treating Ψ as a spinless fermion. Time reversal then acts as

$$\mathcal{T}: \quad \Psi \to \mu^x\tau^x\Psi, \quad i \to -i. \tag{B2}$$

It follows that the $SU(2)_\kappa$, $SU(2)_K$ triplets and the γ matrices map as

$$\begin{aligned} &(\tau^x, \tau^y, \tau^z) \to (\tau^x, \tau^y, -\tau^z), \\ \mathcal{T}: \quad &(\mu^x\eta^y, \mu^y\eta^y, \mu^z) \to -(\mu^x\eta^y, \mu^y\eta^y, \mu^z), \\ &(\gamma^0, \gamma^x, \gamma^y) \to (-\gamma^0, \gamma^x, \gamma^y) \end{aligned} \tag{B3}$$

These transformation properties result in the following time-reversal invariant bilinears:

$$\begin{aligned} &\bar\Psi M\Psi, \quad M \in \{\bar\mu^i, \bar\mu^i\tau^{x,y}, \tau^z\}, \\ &\bar\Psi\gamma^0 M\Psi, \quad M \in \{\mathbb{1}, \tau^{x,y}, \tau^z\bar\mu^i\}, \\ &\bar\Psi\gamma^{x,y} M\Psi, \quad M \in \{\bar\mu^i, \bar\mu^i\tau^{x,y}, \tau^z\}, \end{aligned} \tag{B4}$$

where $\bar\Psi = \Psi^\dagger\gamma^0$. We are most concerned the mass bilinears, shown on the first line. We note that $\bar\Psi\bar\mu^i\Psi = \Psi^\dagger(\mu^y\eta^z, -\mu^x\eta^x, \eta^z)\Psi$. The last term, $\Psi^\dagger\eta^z\Psi$, is the order parameter for the QVH state. For completeness, we also list the bilinears that break time-reversal symmetry:

$$\begin{aligned} &\bar\Psi M\Psi, M \in \{\mathbb{1}, \tau^{x,y}, \tau^z\bar\mu^i\}, \\ &\bar\Psi\gamma^0 M\Psi, M \in \{\bar\mu^i, \bar\mu^i\tau^{x,y}, \tau^z\}, \\ &\bar\Psi\gamma^{x,y} M\Psi, M \in \{\mathbb{1}, \tau^{x,y}, \tau^z\bar\mu^i\}. \end{aligned} \tag{B5}$$

APPENDIX C: SUPPRESSION OF INTER-*K*-VALLEY SCATTERING

We briefly outline a schematic argument for the exponential suppression of inter-$\boldsymbol{K}$-valley scattering processes. We begin by considering the operators on the microscopic graphene lattice. Suppose disorder couples as

$$H_{\text{micro}} = \sum_{\ell,\ell'} \int_{\boldsymbol{r}} \mathcal{R}(\boldsymbol{r}) \tilde f^\dagger_\ell(\boldsymbol{r}) T_{\ell\ell'} \tilde f_{\ell'}(\boldsymbol{r}). \tag{C1}$$

Here, ℓ labels both the layer and sublattice of the fermion $\tilde f_\ell(\boldsymbol{r})$, $T_{\ell\ell'}$ is a matrix whose precise form is unimportant, and $\mathcal{R}(\boldsymbol{r})$ is the disorder field with values drawn from a Gaussian probability distribution:

$$\overline{\mathcal{R}(\boldsymbol{r})} = 0, \quad \overline{\mathcal{R}(\boldsymbol{r})\mathcal{R}(\boldsymbol{r}')} = g^2\, e^{-(\boldsymbol{r}-\boldsymbol{r}')^2/(2\xi_{\text{dis}}^2)}. \tag{C2}$$

In momentum space, we find

$$H_{\text{micro}} = \int_{\boldsymbol{k}} \mathcal{R}(\boldsymbol{q}) \tilde f^\dagger_\ell(\boldsymbol{k}) T_{\ell\ell'} \tilde f_{\ell'}(\boldsymbol{k}+\boldsymbol{q}), \tag{C3}$$

where

$$\overline{\mathcal{R}(\boldsymbol{q})\mathcal{R}^*(\boldsymbol{q}')} = \delta^2(\boldsymbol{q}-\boldsymbol{q}')\, g^2\xi_{\text{dis}}^2\, e^{-\boldsymbol{q}^2\xi_{\text{dis}}^2/2}. \tag{C4}$$

We now wish to expand about the $+\boldsymbol{K}$ and $-\boldsymbol{K}$ points. Letting $\tilde f_{n=\pm,\ell}(\boldsymbol{k}) = f_\ell(\pm\boldsymbol{K}+\boldsymbol{k})$, the Hamiltonian divides into two pieces:

$$\begin{aligned} H_{KK} &= \sum_{n=\pm} \int_{\boldsymbol{k},\boldsymbol{q}} \mathcal{R}(\boldsymbol{q}) f_n^\dagger(\boldsymbol{k}) T f_n(\boldsymbol{k}+\boldsymbol{q}), \\ H_{KK'} &= \int_{\boldsymbol{k},\boldsymbol{q}} \sum_{j=1}^{3} \mathcal{R}(\boldsymbol{q}+\boldsymbol{Q}_j) f_+^\dagger(\boldsymbol{k}) T f_-(\boldsymbol{k}+\boldsymbol{q}) + \text{H.c.} \end{aligned} \tag{C5}$$

where $\boldsymbol{Q}_j$ are the three (smallest) momenta such that $-\boldsymbol{K} + \boldsymbol{Q}_j = +\boldsymbol{K}$, each of which has magnitude $|\boldsymbol{K}| = 4\pi/3a$, where a is the lattice constant of monolayer graphene. We have also suppressed the summation over the ℓ indices of the fermions and matrix T. Letting $\mathcal{R}_{(+-)}(\boldsymbol{q}) = \sum_j \mathcal{R}(\boldsymbol{q}+\boldsymbol{Q}_j)$, we then see

$$\begin{aligned} &\overline{\mathcal{R}_{(+-)}(\boldsymbol{q})\mathcal{R}^*_{(+-)}(\boldsymbol{q}')} \\ &\quad = \delta^2(\boldsymbol{q}-\boldsymbol{q}')\xi_{\text{dis}}^2 g^2 \sum_j e^{-(\boldsymbol{q}+\boldsymbol{Q}_j)^2\xi_{\text{dis}}^2/2} \\ &\quad = \delta^2(\boldsymbol{q}-\boldsymbol{q}')\xi_{\text{dis}}^2 g^2 e^{-\boldsymbol{K}^2\xi^2/2} e^{-\boldsymbol{q}^2\xi_{\text{dis}}^2/2} \sum_j e^{-\boldsymbol{q}\cdot\boldsymbol{Q}_j\xi^2}. \end{aligned} \tag{C6}$$

Ignoring the anisotropic term on the right, the disorder field corresponding to $\boldsymbol{K} \to -\boldsymbol{K}$ scattering has the same correlation length, ξ_{dis}, but with an exponentially suppressed amplitude: $g_{KK'} \sim g e^{-4\pi^2\xi_{\text{dis}}^2/a^2}$.

These arguments may appear to carry over directly to the case of inter-$\boldsymbol{\kappa}$-scattering, i.e., we may wish to conclude that the typical inter-$\boldsymbol{\kappa}$ valley scattering amplitude $g_{\kappa\kappa'}$ is exponentially suppressed relative to the typical intra-$\boldsymbol{\kappa}$ scattering amplitude g: $g_{\kappa\kappa'} \sim g e^{-4\pi^2\xi_{\text{dis}}^2/a_M^2}$. However, in this case, there are additional subtleties to take into account. While the continuum Hamiltonian does not mix the f fermions on the scale of the large BZ, $\sim 1/a$, they *are* mixed on the scale of the moiré BZ, $\sim 1/a_M$. In particular, the flat band operator c (or, equivalently, the Dirac operator ψ) at a momentum quantum number $\boldsymbol{k}$ in the moiré BZ is composed of a superposition of f fermions with momenta $\boldsymbol{k}+\boldsymbol{G}$ (in the microscopic BZ), where the $\boldsymbol{G}$s are moiré reciprocal lattice vectors, as indicated in Eqs. (A6). With the exception of $\boldsymbol{G} = 0$, all such reciprocal lattice vectors are already of order $|\boldsymbol{\kappa}|$ or larger. Importantly, this mixing is responsible for the very flatness of the bands and therefore constitutes a nonnegligible effect. As a result, unless ξ_{dis} is much, much larger than a_M, these higher moments may nevertheless contribute substantially to the $\boldsymbol{\kappa} \to -\boldsymbol{\kappa}$ scattering processes. We therefore emphasize that the analysis and proposal presented in this paper is *not* predicated on the assumption that $g_{\kappa\kappa'}$ is small.

APPENDIX D: MEAN FIELD ANALYSIS OF INSULATING PHASES

Based on numerical results, we argued in Sec. IV B that the ground state of a single flavor theory with interactions is a Chern insulator. Upon including valley and flavor indices in Sec. IV C, we identified four natural insulating states distinguished by their symmetry action, as summarized in Table I. Further, we noted that only the order parameter for the QVH insulator could couple to disorder, which is vital for the scenario we propose.

Here, we discuss the circumstances under which the QVH insulator is or is not energetically preferred compared to the QSH, QH, and QSVH. We determine the band structure using the continuum model (see Appendix A), and, in spite of the concerns raised at the end of Sec. IV A, we model the interactions using H_C, as written in Eq. (13). Moreover, to further simplify the calculation, we project H_C onto the flat bands, a simplification that may admittedly neglect relevant contributions from the dispersive bands. We therefore view this exercise mainly as a guide intended to expose trends rather

than provide rigorous quantitative energetics. Nevertheless, we show that within a simple mean-field analysis, the Fock terms are not expected to distinguish these phases. While it appears that the Hartree terms favor the QSH, QSVH, and QH insulators over the QVH phase, we find that this preference is not the case for the chiral model [64]—they remain degenerate. We next calculate the energy difference between the QVH and other phases numerically for a more realistic set of parameters and demonstrate that while the energy difference is no longer zero, it remains negligibly small.

1. Flat band projection

The Hamiltonian H_C of Eq. (13) is still quite complicated: it includes all bands of the model, whereas we are only interested in what happens to the flat bands. Since these bands are separated from the dispersive bands by a gap E_g by assumption, the latter states can be integrated out to give an effective Hamiltonian acting only on the flat band subspace. The leading order contribution is obtained simply by projecting H_C to the flat bands:

$$H_{C,1} = \int_{\boldsymbol{q}\,\text{small}} V(\boldsymbol{q})\rho_{fl}(\boldsymbol{q})\rho_{fl}(-\boldsymbol{q}), \tag{D1}$$

where $\rho_{fl}(\boldsymbol{q})$ is the density operator projected onto the flat bands.

We show that the mean-field decoupling of $H_0 + H_{C,1}$ [where H_0 is given in Eq. (2)] is independent of the sign of the Dirac mass. To do so, we define the variational Hamiltonian $H_{\text{MF}}(\{M_\mu\}) = \sum_\mu H^{(\mu)}_{\text{MF}}(M_\mu)$ where $\mu = (n, \alpha)$ sums over both $\boldsymbol{K}$ valleys, $n = \pm$, and spin, $\alpha = \uparrow, \downarrow$. The individual mean-field Hamiltonians are

$$H^{(\mu)}_{\text{MF}}(M_\mu) = \int_{\boldsymbol{k}\in BZ} c^\dagger_\mu(\boldsymbol{k}) \underbrace{[h_\mu(\boldsymbol{k}) + M_\mu\eta^z]}_{\bar{h}_\mu(\boldsymbol{k};M_\mu)} c_\mu(\boldsymbol{k}),$$
$$h_\mu(\boldsymbol{k}) = h_{\mu,0}(\boldsymbol{k}) + h_{\mu,x}(\boldsymbol{k})\eta^x + h_{\mu,y}(\boldsymbol{k})\eta^y. \tag{D2}$$

We study the dependence of $\langle\{M_\mu\}|H_0 + H_{C,1}|\{M_\mu\}\rangle$ on the signs of M_μ, where $|\{M_\mu\}\rangle$ denotes the ground state of $H_{\text{MF}}(\{M_\mu\})$.

2. Density operator and form factors

One complication of this calculation is the presence of form factors in the definition of the densities and thus $H_{C,1}$ as well. In particular, we have

$$\rho_{fl}(\boldsymbol{q}) = \sum_\mu \rho_\mu(\boldsymbol{q}),$$
$$\rho_\mu(\boldsymbol{q}) = \sum_{\mu,\ell}\int_{\boldsymbol{k}\,\text{small}} f^\dagger_{\mu,\ell}(\boldsymbol{k}) f_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{q}), \tag{D3}$$

where $f_{\mu,\ell}(\boldsymbol{k}) = f_{n=\pm,\alpha,\ell}(\boldsymbol{k})$ denotes the electron operator with spin $\alpha = \uparrow, \downarrow$ and total momentum $\pm\boldsymbol{K} + \boldsymbol{k}$. As in Sec. IV and Appendix A, ℓ labels both layer and sublattice. In what follows, we omit the label fl. Recall that neither the momentum of the density operator, $\boldsymbol{q}$, nor the momentum being summed over, $\boldsymbol{k}$, is required to lie within the moiré BZ. We therefore instead write

$$\rho_\mu(\boldsymbol{q}+\boldsymbol{G}') = \int_{\boldsymbol{k}\in\text{BZ}} \sum_{\boldsymbol{G},\ell} f^\dagger_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{G}) f_{\mu,\ell}(\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}+\boldsymbol{G}'), \tag{D4}$$

where $\boldsymbol{G}$ and $\boldsymbol{G}'$ are moiré reciprocal lattice vectors and both $\boldsymbol{k}$ and $\boldsymbol{q}$ lie within the moiré BZ. Using the definition of $c_{\mu,i}$ in relation to $f_{\mu,\ell}$ given in Eq. (A6), the density may now be expressed directly in terms of the flat band creation and annihilation operators:

$$\rho_\mu(\boldsymbol{q}+\boldsymbol{G}) = \sum_{ij\in fl} c^\dagger_{\mu,i}(\boldsymbol{k})\lambda_{\mu;ij} \times (\boldsymbol{k}, \boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}) c_{\mu,j}(\boldsymbol{k}+\boldsymbol{q}),$$
$$\lambda_{\mu;ij}(\boldsymbol{k}, \boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}) = \sum_{\boldsymbol{G}',\ell} u^*_{\mu,i;\boldsymbol{G}',\ell}(\boldsymbol{k}) u_{\mu,j;\boldsymbol{G}',\ell}(\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}). \tag{D5}$$

We frequently refer to the functions $\lambda_{\mu,ij}$ as form factors in what follows. We have used the fact that the band operators are invariant under reciprocal lattice translations up to a phase, $c_{\mu,j}(\boldsymbol{p}+\boldsymbol{G}) = e^{i\phi}c_{\mu,j}(\boldsymbol{p})$. From the fact that $u_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k}+\boldsymbol{G}') = u_{\mu,i;\boldsymbol{G}+\boldsymbol{G}',\ell}(\boldsymbol{k})$, we also have $\lambda_{\mu;ij}(\boldsymbol{k}, \boldsymbol{k}'+\boldsymbol{G}) = \lambda_{\mu,ij}(\boldsymbol{k}-\boldsymbol{G}, \boldsymbol{k}')$. Finally, with this notation, the flat band Coulomb interaction is

$$H_{C,1} = \int_{\boldsymbol{q},\boldsymbol{k},\boldsymbol{k}'} \sum_{\boldsymbol{G}}\sum_{\mu,\nu} c^\dagger_\mu(\boldsymbol{k})\lambda_\mu(\boldsymbol{k}, \boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}) c_\mu(\boldsymbol{k}+\boldsymbol{q}) \times \cdot c^\dagger_\nu(\boldsymbol{k}'+\boldsymbol{q})\lambda_\nu(\boldsymbol{k}'+\boldsymbol{q}+\boldsymbol{G}, \boldsymbol{k}') c_\nu(\boldsymbol{k}'). \tag{D6}$$

3. Symmetry constraints

We begin by discussing the symmetry properties of the mean-field kernel $\bar{h}_\mu(\boldsymbol{k}; M_\mu)$. We begin with the symmetry transformations

$$\mathcal{T}: \quad c(\boldsymbol{k}) \to \mu^x c(-\boldsymbol{k}),$$
$$\mathcal{C}_2\mathcal{T}: \quad c(\boldsymbol{k}) \to \eta^x c(\boldsymbol{k}), \tag{D7}$$

where μ^x acts on the $\boldsymbol{K}$-valley indices and η^x acts on the (flat) band indices. Both are anti-Hermitian, taking $i \to -i$. In terms of the mean-field Hamiltonian, they imply

$$\bar{h}_{+,\alpha}(\boldsymbol{k}; M) = \bar{h}^*_{-,\alpha}(-\boldsymbol{k}; M), \quad \bar{h}_\mu(\boldsymbol{k}; M) = \eta^x \bar{h}^*_\mu(\boldsymbol{k}; -M)\eta^x. \tag{D8}$$

Obviously, since $h_\mu(\boldsymbol{k}) = \bar{h}_\mu(\boldsymbol{k}; M = 0)$, these relations also hold for the noninteracting part of the flat band Hamiltonian.

We now define the projector:

$$P_{\mu;ij}(\boldsymbol{k}; M) = \left\langle c^\dagger_{\mu,j}(\boldsymbol{k}) c_{\mu,i}(\boldsymbol{k})\right\rangle_M. \tag{D9}$$

The subscript M is used as shorthand to denote which mean-field Hamiltonian the ground state being used to compute the expectation value is associated with. The equalities of Eq. (D8) then imply

$$P_{+,\alpha}(\boldsymbol{k}; M) = P^T_{-,\alpha}(-\boldsymbol{k}; M), \tag{D10a}$$
$$P_\mu(\boldsymbol{k}; M) = \eta^x P^T_\mu(\boldsymbol{k}; -M)\eta^x. \tag{D10b}$$

Note that $P_\mu^\dagger(\boldsymbol{k};M) = P_\mu(\boldsymbol{k};M)$. Similarly, we find that the form factors must satisfy

$$\lambda_{+,\alpha}(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}) = \lambda_{-,\alpha}^T(-\boldsymbol{k}-\boldsymbol{q},-\boldsymbol{k}), \tag{D11a}$$

$$\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}) = \eta^x\lambda_\mu^T(\boldsymbol{k}+\boldsymbol{q},\boldsymbol{k})\eta^x. \tag{D11b}$$

4. Evaluation of mean field Hamiltonian

We wish to compute the expectation value $\langle\{M_\mu\}\,|\,H_0+H_{C,1}\,|\,\{M_\mu\}\rangle$. This function may be separated into three pieces:

$$\begin{aligned}&\langle\{M_\mu\}\,|\,H_0+H_{C,1}\,|\,\{M_\mu\}\rangle\\ &\quad= \langle H_0\rangle_{\{M_\mu\}} + H_F(\{M_\mu\}) + H_H(\{M_\mu\}),\end{aligned} \tag{D12}$$

where H_F and H_H are the Fock and Hartree decouplings of the Coulomb interaction. These three terms are discussed in the following subsections.

a. Quadratic term: $\langle H_0\rangle$

We write the quadratic part of the Hamiltonian as a sum over the valleys and spins, $H_0 = \sum_\mu H_0^{(\mu)}$, where

$$H_0^{(\mu)} = \int_{\boldsymbol{k}} c_\mu^\dagger(\boldsymbol{k})h_\mu(\boldsymbol{k})c_\mu(\boldsymbol{k}). \tag{D13}$$

The kernel $h_\mu(\boldsymbol{k})$ is defined in Eq. (D2). Taking the expectation value, we find

$$\langle H_0^{(\mu)}\rangle_{M_\mu} = \int_{\boldsymbol{k}} \mathrm{tr}\big[P_\mu(\boldsymbol{k};M_\mu)h_\mu(\boldsymbol{k})\big]. \tag{D14}$$

Inserting the relations given in Eqs. (D10b) and (D11b), we arrive at

$$\begin{aligned}\langle H_0^{(\mu)}\rangle_{M_\mu} &= \int_{\boldsymbol{k}} \mathrm{tr}\big[\eta^x P_\mu^T(\boldsymbol{k};-M_\mu)\eta^x\eta^x h_\mu^T(\boldsymbol{k})\eta^x\big]\\ &= \langle H_0^{(\mu)}\rangle_{-M_\mu}.\end{aligned} \tag{D15}$$

Hence, we have verified that $\langle H_0\rangle$ is independent of the signs of the mass terms.

b. Fock term: H_F

The Fock term is

$$\begin{aligned}H_F(\{M_\mu\}) &= \sum_\mu H_F^{(\mu)}(M_\mu),\\ H_F^{(\mu)}(M_\mu) &= -\int_{\boldsymbol{k},\boldsymbol{p}}\sum_{\boldsymbol{G}} V(\boldsymbol{p}-\boldsymbol{k}+\boldsymbol{G})\mathrm{tr}\\ &\quad\times[\lambda_\mu(\boldsymbol{k},\boldsymbol{p}+\boldsymbol{G})P_\mu(\boldsymbol{p};M_\mu)\lambda_\mu\\ &\quad\times(\boldsymbol{p}+\boldsymbol{G},\boldsymbol{k})P_\mu(\boldsymbol{k};M_\mu)].\end{aligned} \tag{D16}$$

Inserting the relations from Eqs. (D10b) and (D11b), we find

$$\begin{aligned}H_F^{(\mu)}(M_\mu) &= -\int_{\boldsymbol{k},\boldsymbol{p}}\sum_{\boldsymbol{G}} V(\boldsymbol{p}-\boldsymbol{k}+\boldsymbol{G})\mathrm{tr}\\ &\quad\times[\lambda_\mu^T(\boldsymbol{p}+\boldsymbol{G},\boldsymbol{k})P_\mu^T(\boldsymbol{p};-M_\mu)\lambda_\mu^T\\ &\quad\times(\boldsymbol{k},\boldsymbol{p}+\boldsymbol{G})P_\mu^T(\boldsymbol{k};-M_\mu)]\\ &= H_F^{(\mu)}(-M_\mu).\end{aligned} \tag{D17}$$

TABLE II. Relative signs of the mass terms corresponding to the four phases depicted in Fig. 4.

	$M_{+,\uparrow}$	$M_{+,\downarrow}$	$M_{-,\uparrow}$	$M_{-,\downarrow}$
QVH	1	1	1	1
QSVH	1	−1	1	−1
QH	1	1	−1	−1
QSH	1	−1	−1	1

We again conclude that the Fock contribution is independent of the sign M_μ takes.

c. Hartree term: H_H

The Hartree term can be written

$$H_H(\{M_\mu\}) = \sum_{\boldsymbol{G}} V(\boldsymbol{G})\sum_{\mu,\nu}\langle\rho_\mu(\boldsymbol{G})\rangle_{M_\mu}\langle\rho_\nu(-\boldsymbol{G})\rangle_{M_\nu}. \tag{D18}$$

We therefore begin by calculating $\langle\rho_\mu(\boldsymbol{G})\rangle_M$:

$$\langle\rho_\mu(\boldsymbol{G})\rangle_M = \int_{\boldsymbol{k}}\mathrm{tr}[P_\mu(\boldsymbol{k};M)\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})]. \tag{D19}$$

We use the constraints imposed by time reversal [Eqs. (D10a) and (D11a)] to relate the expectation values of the densities of the two valleys to one another:

$$\begin{aligned}\langle\rho_{+,\alpha}(\boldsymbol{G})\rangle_M &= \int_{\boldsymbol{k}}\mathrm{tr}\big[P_{-,\alpha}^T(-\boldsymbol{k};M)\lambda_{-,\alpha}^T(-\boldsymbol{k}-\boldsymbol{G},-\boldsymbol{k})\big]\\ &= \int_{\boldsymbol{k}}\mathrm{tr}[P_{-,\alpha}(\boldsymbol{k};M)\lambda_{-,\alpha}(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})]\\ &= \langle\rho_{-,\alpha}(\boldsymbol{G})\rangle_M.\end{aligned} \tag{D20}$$

We see that the expectation value of the density operator is independent of the valley and spin degree of freedom, motivating us to define the function

$$R(M;\boldsymbol{G}) \equiv \langle\rho_\mu(\boldsymbol{G})\rangle_M. \tag{D21}$$

Note that the identity $\rho_\mu(\boldsymbol{G}) = \rho_\mu^\dagger(-\boldsymbol{G})$ implies $R(M;\boldsymbol{G}) = R^*(M;-\boldsymbol{G})$. The $\mathcal{C}_2\mathcal{T}$ symmetry [Eqs. (D10b) and (D11b)] then gives

$$\begin{aligned}\langle\rho_\mu(\boldsymbol{G})\rangle_M &= \int_{\boldsymbol{k}}\mathrm{tr}\big[P_\mu^T(\boldsymbol{k};-M)\lambda_\mu^T(\boldsymbol{k}+\boldsymbol{G},\boldsymbol{k})\big]\\ &= \int_{\boldsymbol{k}}\mathrm{tr}[P_\mu(\boldsymbol{k};-M)\lambda_\mu(\boldsymbol{k},\boldsymbol{k}-\boldsymbol{G})]\\ &= \langle\rho_\mu(-\boldsymbol{G})\rangle_{-M} = \langle\rho_\mu(\boldsymbol{G})\rangle_{-M}^*.\end{aligned} \tag{D22}$$

We conclude that $R(-M;\boldsymbol{G}) = R^*(M;\boldsymbol{G})$.

The Hartree term is therefore

$$H_H(\{M_\mu\}) = \sum_{\boldsymbol{G}} V(\boldsymbol{G})\Big|\sum_\mu R(M_\mu;\boldsymbol{G})\Big|^2. \tag{D23}$$

The relative signs of the mass terms of the four states under consideration are shown in Table II. Separating $R(M;\boldsymbol{G})$ into real and imaginary parts, $R(M;\boldsymbol{G}) = R'(M;\boldsymbol{G}) + iR''(M;\boldsymbol{G})$,

we conclude that

$$H_H^{\mathrm{QVH}} = 16\sum_{\boldsymbol{G}} V(\boldsymbol{G})[R'(M;\boldsymbol{G})^2 + R''(M;\boldsymbol{G})^2],$$
$$H_H^{\mathrm{QSVH}} = H_H^{\mathrm{QH}} = H_H^{\mathrm{QSH}} = 16\sum_{\boldsymbol{G}} V(\boldsymbol{G})R'(M;\boldsymbol{G})^2. \quad \text{(D24)}$$

It follows that the QVH state is *higher* in energy than the other three insulating states by $16\sum_{\boldsymbol{G}} V(\boldsymbol{G})R''(M;\boldsymbol{G})^2$.

We note that since $\lambda(\boldsymbol{k},\boldsymbol{k}) = \mathbb{1}$, for $\boldsymbol{G} = 0$ we necessarily have $R''(M;\boldsymbol{0}) = 0$, implying that for this term at least, there is no difference in energy between the QVH insulator and the other three. In a typical tight-binding model, the $\boldsymbol{G} = 0$ term accounts for the entirety of the Hartree energy. For the continuum model, however, the internal spatial structure of the wave functions also affects the Hartree energy. Nevertheless, the form factors $\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})$ decay quite quickly as a function of $\boldsymbol{G}$ [21]—implying that the spatial variation of the density within the unit cell is not too large. As we discuss in the next two sections, the contribution from $R''(M;\boldsymbol{G})$ is essentially negligible.

5. Chiral model

We show that in the chiral model [64], the functions $R(M;\boldsymbol{G})$ are purely real, implying that the Hartree terms are all degenerate. The chiral model is a particular case of the continuum model in which hopping only occurs between A and B sites both within and between graphene layers. This constraint is implemented by setting w_0 in Eq. (A4) to zero. The result is an exact particle-hole (chiral) symmetry Γ that interchanges positive and negative energy states. We follow the discussion in the Appendix of Ref. [21]. Γ may be assumed to act as

$$\Gamma: \quad c(\boldsymbol{k}) \to \eta^z c(\boldsymbol{k}). \quad \text{(D25)}$$

In fact, in this basis, the sublattice index of the $c(\boldsymbol{k})$'s can be identified with the A and B sublattices of the two layers. It's then convenient to reinterpret the wave functions written in Eq. (A6), $u_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k})$. We explicitly identify the index $i = \mathrm{A}, \mathrm{B}$ with the sublattice, leaving ℓ to denote the layer. It then follows that the form factor may be written

$$\lambda_{\mu,ij}(\boldsymbol{k},\boldsymbol{k}'+\boldsymbol{G}) = \left[\lambda_\mu^{(0)}(\boldsymbol{k},\boldsymbol{k}'+\boldsymbol{G})\mathbb{1}_{2\times 2} + i\lambda_\mu^{(z)}(\boldsymbol{k},\boldsymbol{k}'+\boldsymbol{G})\eta^z\right]_{ij}, \quad \text{(D26)}$$

where both $\lambda_\mu^{(0)}$ and $\lambda_\mu^{(z)}$ are real functions.

An additional symmetry allows one to rotate the two layers in opposite directions. The authors of Ref. [64] use this observation to simplify the problem substantially, resulting in an exact expression for the ground-state wave function at the magic angle. For any angle, however, it implies that the Hamiltonian of Eq. (A2) satisfies

$$\begin{pmatrix} iv_0\boldsymbol{\eta}_{\theta/2}\cdot\boldsymbol{\nabla} & T(\boldsymbol{r}) \\ T^\dagger(\boldsymbol{r}) & iv_0\boldsymbol{\eta}_{-\theta/2}\cdot\boldsymbol{\nabla}\end{pmatrix} = \eta^z\tau^z\begin{pmatrix} -iv_0\boldsymbol{\eta}_{\theta/2}\cdot\boldsymbol{\nabla} & T(\boldsymbol{r}) \\ T^\dagger(\boldsymbol{r}) & -iv_0\boldsymbol{\eta}_{-\theta/2}\cdot\boldsymbol{\nabla}\end{pmatrix}\tau^z\eta^z, \quad \text{(D27)}$$

where Pauli operators η^z and τ^z act on the sublattice (A,B) and layer (t,b) indices, respectively. The continuum representation of the wave function given in Eq. (A6) therefore satisfies

$$u_{\mu,i;\boldsymbol{G},\ell}(\boldsymbol{k}) = e^{i\varphi_{\boldsymbol{k}}}\sum_{i',\ell'}\eta^z_{ii'}\tau^z_{\ell\ell'}u_{\mu,i';-\boldsymbol{G}\ell'}(-\boldsymbol{k}), \quad \text{(D28)}$$

which in turn implies

$$\lambda_\mu(\boldsymbol{k},\boldsymbol{k}'+\boldsymbol{G}) = \lambda_\mu(-\boldsymbol{k},-\boldsymbol{k}'-\boldsymbol{G}). \quad \text{(D29)}$$

Similarly, the mean-field Hamiltonian must give $\bar{h}_\mu(\boldsymbol{k};M) = \bar{h}_\mu(-\boldsymbol{k};M)$ and therefore

$$P_\mu(\boldsymbol{k};M) = P_\mu(-\boldsymbol{k};M). \quad \text{(D30)}$$

These relations provide an additional constraint on the form of $\langle\rho(\boldsymbol{G})\rangle_M$:

$$\begin{aligned}\langle\rho_\mu(\boldsymbol{G})\rangle_M &= \int_{\boldsymbol{k}}\mathrm{tr}[P_\mu(\boldsymbol{k};M)\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})] \\ &= \int_{\boldsymbol{k}}\mathrm{tr}[P_\mu(-\boldsymbol{k};M)\lambda_\mu(-\boldsymbol{k},-\boldsymbol{k}-\boldsymbol{G})] \\ &= \int_{\boldsymbol{k}}\mathrm{tr}[P_\mu(\boldsymbol{k};M)\lambda_\mu(\boldsymbol{k},\boldsymbol{k}-\boldsymbol{G})] \\ &= \langle\rho_\mu(-\boldsymbol{G})\rangle_M = \langle\rho_\mu(\boldsymbol{G})\rangle_M^*. \end{aligned} \quad \text{(D31)}$$

That is, $R(M;\boldsymbol{G})$ is *real*: $R''(M;\boldsymbol{G}) = 0$. From Eqs. (D24), we conclude that the Hartree energies corresponding to all four insulating states are fully degenerate in the chiral limit:

$$H_H^{\mathrm{QVH}} = H_H^{\mathrm{QSVH}} = H_H^{\mathrm{QH}} = H_H^{\mathrm{QSH}}. \quad \text{(D32)}$$

6. Numerical evaluation of Hartree term

We now return to the nonchiral version of the model. In Fig. 8(a), we plot the energy difference per electron of the Hartree term for the model using the parameters given in Appendix A as a function of the Dirac mass M. Even for a mass $M = 3$ meV, the energy difference is as small as 2.5×10^{-6} meV — certainly our rough model is not expected to be reliable for such small energy differences.

We can understand the smallness in several ways. As mentioned at the end of Appendix D 4 c, the form factors $\lambda(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})$ decay quite quickly as a function of $\boldsymbol{G}$. We can further show that $R''(M;\boldsymbol{G}) = 0$ for all $\boldsymbol{G}$ such that $\boldsymbol{G} = \mathcal{M}_y[\boldsymbol{G}]$, $\boldsymbol{G} = \mathcal{C}_3\mathcal{M}_y[\boldsymbol{G}]$, or $\boldsymbol{G} = \mathcal{C}_3^2\mathcal{M}_y[\boldsymbol{G}]$. To do so, we start by using the fact that a basis exists in which Eqs. (D8) hold and the mirror symmetry acts as [46]

$$\mathcal{M}_y: \quad c(\boldsymbol{k}) \to \eta^x c(\mathcal{M}_y[\boldsymbol{k}]). \quad \text{(D33)}$$

Since $h_\mu(\boldsymbol{k})$ satisfies the symmetry whereas the mass term $M\eta^z$ does not (e.g., $\bar{h}_\mu(\boldsymbol{k};M) = \eta^x\bar{h}_\mu(\mathcal{M}_y[\boldsymbol{k}];-M)\eta^x$), we

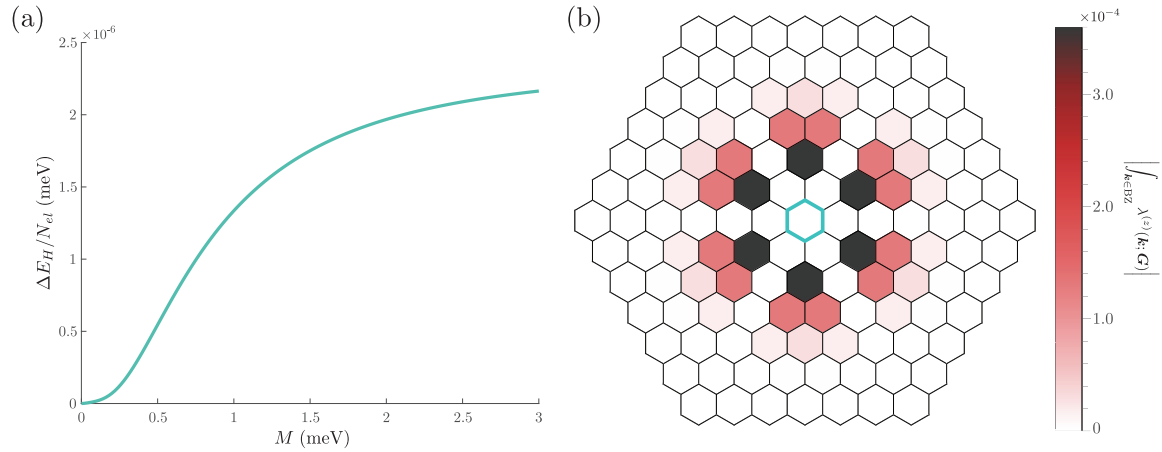

FIG. 8. (a) Energy difference as a function of the variational mass M between the QVH phase, H_H^{QVH}, and the other three phases, $H_H^{\mathrm{other}} = H_H^{\mathrm{QSH}} = H_H^{\mathrm{QH}} = H_H^{\mathrm{QSVH}}$, per electron at charge neutrality: $\Delta E_H/N_{\mathrm{el}} = (H_H^{\mathrm{QVH}} - H_H^{\mathrm{other}})/N_{\mathrm{el}}$. (b) Color plot of $|\int_{\boldsymbol{k}} \lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G})|$ as a function of the moiré reciprocal lattice vector $\boldsymbol{G}$. Each hexagon represents a different $\boldsymbol{G}$, with the central hexagon outlined in turquoise corresponding to $\boldsymbol{G} = 0$. Noticeably, $|\int_{\boldsymbol{k}} \lambda^{(z)}(\boldsymbol{k};\boldsymbol{G})| = 0$ along all mirror axes, as we showed in the main text.

must have

$$P_\mu(\boldsymbol{k};M) = \eta^x P_\mu(\mathcal{M}_y[\boldsymbol{k}];-M)\eta^x, \quad \lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G}) = \eta^x \lambda_\mu(\mathcal{M}_y[\boldsymbol{k}],\mathcal{M}_y[\boldsymbol{k}+\boldsymbol{G}])\eta^x. \tag{D34}$$

We therefore find

$$\begin{aligned}\langle \rho_\mu(\boldsymbol{G})\rangle_M &= \int_{\boldsymbol{k}\in \mathrm{BZ}} \mathrm{tr}[\eta^x P_\mu(\mathcal{M}_y[\boldsymbol{k}];-M)\eta^x \eta^x \lambda_\mu(\mathcal{M}_y[\boldsymbol{k}],\mathcal{M}_y[\boldsymbol{k}+\boldsymbol{G}])\eta^x] \\ &= \int_{\boldsymbol{k}\in \mathrm{BZ}} \mathrm{tr}[P_\mu(\boldsymbol{k};-M)\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\mathcal{M}_y[\boldsymbol{G}])] \\ &= \langle \rho_\mu(\mathcal{M}_y[\boldsymbol{G}])\rangle_{-M} = \langle \rho_\mu(\mathcal{M}_y[\boldsymbol{G}])\rangle_M^*. \end{aligned} \tag{D35}$$

It follows that $\langle \rho_\mu(\boldsymbol{G})\rangle_M$ is *real* for all moiré reciprocal lattice vectors such that $\boldsymbol{G} = \mathcal{M}_y[\boldsymbol{G}]$: $R''(M;\boldsymbol{G} = \mathcal{M}_y[\boldsymbol{G}]) = 0$. The reflection axis chosen for $\mathcal{M}_y$ was actually arbitrary—by $\mathcal{C}_3$ rotational symmetry, the same should hold for the two equivalent axes given by $\mathcal{C}_3\mathcal{M}_y$ and $\mathcal{C}_3^2\mathcal{M}_y$. Notably, this means that $R''(M;\boldsymbol{G}) = 0$ for the shortest set reciprocal lattice vectors.

We can quantify the size of $R''(M;\boldsymbol{G})$ for arbitrary $\boldsymbol{G}$ through the following set of observations. First, we note that the energies of the flat bands may be written as $E_{\mu,\pm}(\boldsymbol{k}) = h_{\mu,0}(\boldsymbol{k}) \pm \epsilon_\mu(\boldsymbol{k})$, where $\epsilon_\mu^2(\boldsymbol{k}) = h_{\mu,x}^2(\boldsymbol{k}) + h_{\mu,y}^2(\boldsymbol{k})$. This allows us to express the projection matrix as

$$P_\mu(\boldsymbol{k};M) = \frac{1}{2}\left(\mathbb{1} - \frac{1}{\sqrt{\epsilon_\mu^2(\boldsymbol{k})+M^2}}(h_{\mu,x}(\boldsymbol{k})\eta^x + h_{\mu,y}(\boldsymbol{k})\eta^y + M\eta^z)\right). \tag{D36}$$

It then follows that

$$\begin{aligned} R''(M;\boldsymbol{G}) &= \frac{1}{2}\left(\langle \rho_\mu(\boldsymbol{G})\rangle_M - \langle \rho_\mu(\boldsymbol{G})\rangle_{-M}\right) = -\frac{1}{2}\int_{\boldsymbol{k}\in \mathrm{BZ}} \frac{M}{\sqrt{\epsilon_\mu^2(\boldsymbol{k})+M^2}} \mathrm{tr}[\eta^z \lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})] \\ &= -\int_{\boldsymbol{k}\in \mathrm{BZ}} \frac{M}{\sqrt{\epsilon_\mu^2(\boldsymbol{k})+M^2}} \lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G}), \end{aligned} \tag{D37}$$

where we've defined

$$\lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G}) = -\frac{i}{2}\mathrm{tr}[\eta^z \lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})]. \tag{D38}$$

We can verify through Eq. (D11b) and the identity $\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G}) = \lambda_\mu^\dagger(\boldsymbol{k},\boldsymbol{k}-\boldsymbol{G})$ that $\mathrm{tr}[\eta^z\lambda_\mu(\boldsymbol{k},\boldsymbol{k}+\boldsymbol{G})]$ must be imaginary. In limit that M is large, Eq. (D37) implies

$$R''(M;\boldsymbol{G}) \to -\int_{\boldsymbol{k}\in \mathrm{BZ}} \lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G}). \tag{D39}$$

Assuming that $R''(M;\boldsymbol{G})$ is a monotonically increasing function of M (which Fig. 8(a) verifies at least for the parameters

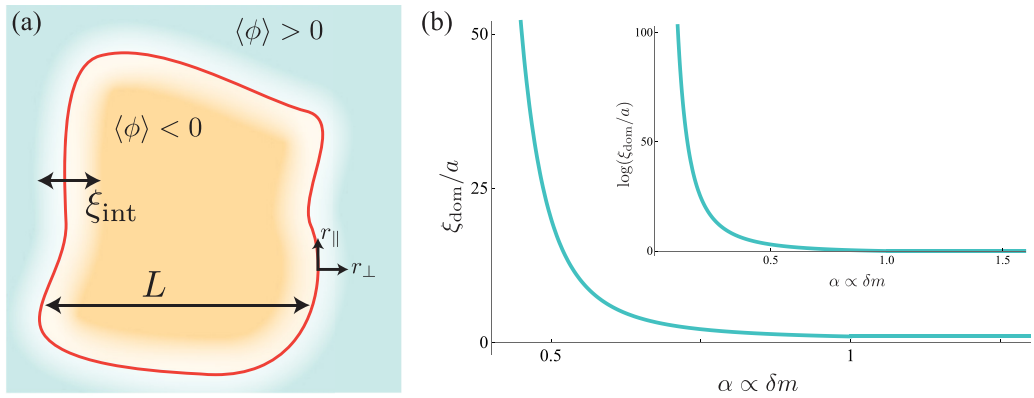

FIG. 9. (a) Illustration of a domain D of linear size $\sim L$ with $\langle\phi\rangle < 0$ (orange region) immersed within a region of $\langle\phi\rangle > 0$ (blue region). The boundary region of the domain, ∂D, is indicated in white. Its width, $\sim\xi_{\rm int}$, is shown with an arrow. The coordinates $(r_\perp, r_\parallel)$ used to estimate $\int_r \mathcal{K}(\nabla\phi)^2$ are shown to the right of the domain. (b) Schematic plot of domain size, $\xi_{\rm dom}$, as a function of α [Eq. (18)] for Gaussian-correlated disorder, Appendix E 3. The inset plots the logarithm of the domain size. In both, $a = \max(\xi_{\rm dis}, \xi_{\rm int})$. When $\alpha \lesssim 1$, the disorder is effectively local and the domains are exponentially large, as per Eq. (E25). On the other hand, for $\alpha \gtrsim 1$, the domain size is set by the disorder correlation length $\xi_{\rm dis}$. Coefficients of $\mathcal{O}(1)$ have been chosen by hand to smoothen the crossover between these two regimes. Since we assume that $\delta m \ll U$, $\alpha \gtrsim 1$ implies that $\xi_{\rm dis} \gg \xi_{\rm int}$.

considered), we expect $\lambda_\mu^{(z)}$ to supply an upper bound on R'':

$$|R''(M;\boldsymbol{G})| \leqslant \left|\int_{\boldsymbol{k}\in\mathrm{BZ}} \lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G})\right|. \tag{D40}$$

In Fig. 8(b) we plot the right hand side of the above equation as a function of $\boldsymbol{G}$. The fact that $\int_{\boldsymbol{k}} \lambda_\mu^{(z)}(\boldsymbol{k};\boldsymbol{G})$ vanishes for all $\boldsymbol{G}$ such that $\boldsymbol{G} = \mathcal{M}_y[\boldsymbol{G}]$, $\boldsymbol{G} = \mathcal{C}_3\mathcal{M}_y[\boldsymbol{G}]$, and $\boldsymbol{G} = \mathcal{C}_3^2\mathcal{M}_y[\boldsymbol{G}]$ follows from the symmetry analysis given at the beginning of this section—as we see, the reciprocal lattice vectors with the smallest amplitudes do not contribute to $R''(M;\boldsymbol{G})$.

More importantly, the largest value of $\int_{\boldsymbol{k}} \lambda_\mu^{(z)}(\boldsymbol{k},\boldsymbol{G})$ is already incredibly small—its maximum value is $\sim 3.6\times 10^{-4}$. Even when multiplied by the relatively large interaction scale $V(a_M)$, the energy difference between the QVH and the other insulating phases remains small, as evinced by Fig. 8(a). We conclude that, at least within the approximation considered here, the QVH insulator is indistinguishable from its cousins, the QSVH, QH, and QSH states.

APPENDIX E: RANDOM FIELD ISING MODEL DOMAIN ESTIMATES

In this Appendix, we discuss the Imry-Ma [69] arguments used in Sec. V A to obtain the estimates given in Eqs. (19) and (20) for the minimal domain size $\xi_{\rm dom}$. We consider the regime where the homogeneous system would like to order—in this sense, we are assuming that disorder is weak compared to the interaction energy: $\delta m \ll U$. We next estimate the energy cost $E_{\rm dom}(L)$ of changing the sign of ϕ within a domain D of linear extent $\sim L$, as depicted in Fig. 9(a). There are two contributions to $E_{\rm dom}$: one from the interaction energy, $E_{\rm int}(L)$, and another from the disorder potential, $E_{\rm dis}(L)$. As reasoned in the main text, we assume that $|\langle\phi\rangle| \sim \mathcal{O}(1)$. Since we are primarily interested in the relative scaling of the two energy terms, coefficients of $\mathcal{O}(1)$ are not be included.

The interaction energy of the domain is determined by the kinetic term of the Ising model:

$$E_{\rm int}(L) \sim \int d^2\boldsymbol{r}\, \mathcal{K}(\nabla\phi)^2. \tag{E1}$$

The coefficient $\mathcal{K}$ should have units of energy, and so we naturally set $\mathcal{K} \sim U$, as discussed in the main text. The Ising field ϕ changes only within the boundary region ∂D of the flipped domain D. Given our initial definition of ϕ [Eq. (14)], this change can only occur on the scale of $\xi_{\rm int}$ [Eq. (16)], implying that $(\nabla\phi)^2 \sim U\phi/\xi_{\rm int}^2 \sim 1/\xi_{\rm int}^2$. Integrating over ∂D, including its width, contributes a factor of $\xi_{\rm int}L$ so the total cost is

$$E_{\rm int}(L) = U\frac{L}{\xi_{\rm int}}. \tag{E2}$$

More concretely, this estimate can be obtained through the *ansatz* $\phi(\boldsymbol{r}) \sim \tanh(r_\perp/\xi_{\rm int})$, where $r_\perp$ is the direction perpendicular to the domain boundary, with the boundary itself occurring at $r_\perp = 0$ [see Fig. 9(a)]. Ignoring the effect of curvature, we again find

$$\begin{aligned} E_{\rm int}(L) &\sim U\int dr_\parallel \int dr_\perp \frac{1}{\xi_{\rm int}^2}\mathrm{sech}^4\left(\frac{r_\perp - r_0}{\xi_{\rm int}}\right) \\ &\sim U\cdot\frac{1}{\xi_{\rm int}^2}\cdot L\cdot\frac{4}{3}\xi_{\rm int} \sim U\frac{L}{\xi_{\rm int}}. \end{aligned} \tag{E3}$$

We now consider the contribution to the energy cost of the domain due to the random field $\mathcal{B}(\boldsymbol{r})$ [as defined in and below Eq. (17)]. For a given realization of disorder, we have

$$E_{\rm dis}(L) \sim \int_{\boldsymbol{r}\in D} \mathcal{B}(\boldsymbol{r}). \tag{E4}$$

Depending on where the domain is placed, disorder can either increase or decrease the domain energy. For an arbitrarily chosen D, $E_{\rm dis}$ will average to zero, with a standard deviation

given by

$$E_{\rm rms}^2 \sim \overline{\left[\int_{\boldsymbol{r}\in D} \mathcal{B}(\boldsymbol{r})\right]^2} = \frac{\delta m^2}{\xi_{\rm int}^4}\int_{\boldsymbol{r},\boldsymbol{r}'\in D} K\left(\frac{\boldsymbol{r}-\boldsymbol{r}'}{\xi_{\rm dis}}\right). \tag{E5}$$

Importantly, however, the location of the domain is not arbitrary. We can choose to place our domain in a region where this contribution is negative, taking the typical value

$$E_{\rm dis} \sim -\sqrt{E_{\rm rms}^2} \sim -\frac{\delta m}{\xi_{\rm int}^2}\left[\int_{\boldsymbol{r},\boldsymbol{r}'\in D} K\left(\frac{\boldsymbol{r}-\boldsymbol{r}'}{\xi_{\rm dis}}\right)\right]^{1/2}. \tag{E6}$$

The total cost of the domain is therefore

$$E_{\rm dom}(L) \sim U\frac{L}{\xi_{\rm int}} - \frac{\delta m}{\xi_{\rm int}^2}\left[\int_{\boldsymbol{r},\boldsymbol{r}'\in D} K\left(\frac{\boldsymbol{r}-\boldsymbol{r}'}{\xi_{\rm dis}}\right)\right]^{1/2}. \tag{E7}$$

If L_* exists such that $E_{\rm dis}(L_*) = 0$, the formation of the domain is energetically favorable and long-range order is destroyed. This destruction occurs in all of the examples we consider.

1. Long-range disorder

The simplest example actually turns out to be the case of long-range disorder [104]:

$$K\left(\frac{\boldsymbol{r}}{\xi_{\rm dis}}\right) = \frac{\xi_{\rm dis}}{|\boldsymbol{r}|}. \tag{E8}$$

We do not discuss this form of K in the main text since it is unlikely to describe the physical system; it nevertheless serves as a convenient example. We note that while $\xi_{\rm dis}$ is a length scale, it does not truly represent a correlation length in this context. Instead, it simply enters into the disorder strength as a multiplicative factor:

$$\overline{\mathcal{B}(\boldsymbol{r})\mathcal{B}(0)} = \left(\delta m^2 \xi_{\rm dis}\right)\frac{1}{\xi_{\rm int}^4}\frac{1}{|\boldsymbol{r}|}. \tag{E9}$$

Inserting this definition into Eq. (E7), we find that the change in energy expected for a (judiciously chosen) domain of size L is

$$E_{\rm dom}(L) \sim \frac{L}{\xi_{\rm int}}\left(U - \delta m\frac{\sqrt{\xi_{\rm dis}L}}{\xi_{\rm int}}\right). \tag{E10}$$

For large L, it's clear that the domain energy eventually becomes negative, destabilizing the ordered phase. This destruction first occurs at the emergent length scale:

$$L_* \sim \left(\frac{U}{\delta m}\frac{\xi_{\rm int}}{\xi_{\rm dis}}\right)^2 \xi_{\rm dis}. \tag{E11}$$

We conclude that when the disorder is long range, domains are expected to form once the system size is larger than L_*.

2. White noise (short-range) disorder

We now consider local, white noise disorder:

$$K\left(\frac{\boldsymbol{r}}{\xi_{\rm dis}}\right) = \xi_{\rm dis}^2\delta^2(\boldsymbol{r}). \tag{E12}$$

As in the long-range case, the parameter $\xi_{\rm dis}$ enters only as a multiplicative factor. Together with the disorder strength δm and the Fermi velocity v_F, they form a dimensionless parameter $\delta m\,\xi_{\rm dis}/\hbar v_F$ discussed in Sec. III B.

Following the arguments above, an appropriately chosen domain therefore contributes an energy

$$E_{\rm dis}(L) \sim -\delta m\frac{\xi_{\rm dis}L}{\xi_{\rm int}^2}. \tag{E13}$$

The total energy cost of the domain is

$$E_{\rm dom}(L) \sim U\frac{L}{\xi_{\rm int}} - \delta m\frac{\xi_{\rm dis}L}{\xi_{\rm int}^2} = U\frac{L}{\xi_{\rm int}}\Big(1-\alpha\Big), \tag{E14}$$

where we have defined

$$\alpha \equiv \frac{\delta m}{U}\frac{\xi_{\rm dis}}{\xi_{\rm int}}, \tag{E15}$$

as given in Eq. (18) of the main text. Notably, it is not $\delta m/U$ that controls the domain energy cost, but instead the ratio α. This feature is related to our remark that the true disorder strength is actually $g = \delta m\,\xi_{\rm dis}$. The correct energy scale is therefore obtained in units of the UV cutoff, giving $g/\xi_{\rm int} = \alpha U$, from which it follows that α is the appropriate tuning parameter, *not* $\delta m/U$. Equation (E14) simply tells us that when disorder is larger than the interaction scale, $\alpha \gtrsim 1$, there is no reason for the system to order. In this limit, the domain structure and fate of the theory is complicated and will not be relevant for us [105,106].

Conversely, for $\alpha \lesssim 1$, Eq. (E14) may appear to imply that that the system should order. However, while Eq. (E15) is sufficient for large α, the analysis above omits the effect of domain roughening. This effect should be included in general, and it completely alters our conclusions when α is small.

Roughening in the context of the RFIM was first discussed in Ref. [72], and we now summarize the reasoning made there. We begin by considering a portion of a domain wall of linear extent y, displacing it by a (small) length w, and determining the change in energy, $\delta E(w,y)$. First, the displacement increases the length of the boundary by $\delta E_{\rm int} \sim U\,w/\xi_{\rm int}$. With regard to the disorder field, we can choose to displace the boundary to either the left or the right direction, each of which has a 50% likelihood of decreasing the energy. There is therefore a 75% probability that the displacement lowers the energy by a typical amount $\delta E_{\rm dis} \sim -\delta m\xi_{\rm dis}\sqrt{wy}/\xi_{\rm int}^2$. In total, the displacement results in a typical energy change:

$$\delta E(w,y) \sim U\frac{w}{\xi_{\rm int}} - \delta m\frac{\xi_{\rm dis}}{\xi_{\rm int}^2}\sqrt{wy}. \tag{E16}$$

We now minimize δE with respect to w, to obtain

$$w_* \sim \left(\frac{\delta m}{U}\frac{\xi_{\rm dis}}{\xi_{\rm int}}\right)^2 y = \alpha^2 y,$$
$$\delta E_*(y) \equiv \delta E(w_*,y) \sim -\alpha^2 U\frac{y}{\xi_{\rm int}}. \tag{E17}$$

Next, we note that this procedure may be performed for segments of all sizes along the domain boundary. In particular, there are $N(y_\ell) = L/y_\ell$ segments of size $y_\ell = e^{-\ell}L$, each of which contributes an energy $\delta E_*(y_\ell)$. Summing over all scales returns the total energy contribution from domain-wall

roughening:

$$\delta E_{\rm tot}(L,a) = \int_0^{\log(L/a)} d\ell\, N(y_\ell)\delta E_*(y_\ell) \sim -\int_a^L \frac{dy}{y}\frac{L}{y}\alpha^2 U \frac{y}{\xi_{\rm int}} \sim -\alpha^2 U \frac{L}{\xi_{\rm int}} \log\left(\frac{L}{a}\right). \tag{E18}$$

Here, a is the smallest scale at which roughening may occur; in this context, $a \sim \xi_{\rm int}$, though we will find otherwise in the next section. (Note that this a should *not* be confused with the microscopic lattice constant of monolayer graphene). Throughout this derivation, we have assumed that a is significantly smaller than L. Finally, the total energy cost of the domain is

$$E_{\rm dom}(L) \sim \frac{L}{\xi_{\rm int}}\left[U - \alpha^2 U \log\left(\frac{L}{\xi_{\rm int}}\right)\right]. \tag{E19}$$

Solving for $E_{\rm dom}(L_*) = 0$, we find

$$L_* \sim \xi_{\rm int}\, e^{c/\alpha^2}, \tag{E20}$$

where we have introduced the nonuniversal constant $c \sim \mathcal{O}(1)$ to account for the imprecise nature of our scaling arguments. Once more, for systems larger than L_*, multiple domains should be apparent.

As we mentioned below Eq. (E18), our integration was predicated on the assumption that the domain size L was much larger than $\xi_{\rm int}$. It is clear that this is only satisfied provided the disorder is weak: $\alpha \ll 1$. When the disorder is stronger, the situation is more complicated.

3. Gaussian-correlated disorder

We now consider the situation considered in the main text, that of Gaussian correlated disorder:

$$K\left(\frac{\boldsymbol{r}}{\xi_{\rm dis}}\right) = e^{-\frac{r^2}{2\xi_{\rm dis}^2}}. \tag{E21}$$

Unlike the previous two cases, the scale $\xi_{\rm dis}$ is a true correlation length in this scenario, as is clear from the form of the disorder-induced energy reduction:

$$E_{\rm dis}(L) \sim -\delta m \frac{\xi_{\rm dis} L}{\xi_{\rm int}^2}\sqrt{1 - e^{-L^2/2\xi_{\rm dis}^2}}. \tag{E22}$$

While the domain size appeared as a ratio of the UV cutoff $\ell_{\rm UV} = \xi_{\rm int}$ in the previous two examples, here $E_{\rm dis}(L)$ is also a function of $L/\xi_{\rm dis}$.

There are two natural limits to consider. In the first, we take the domain size to be small enough relative to $\xi_{\rm dis}$ that the smoothness of the disorder is still important, i.e., we cannot simply ignore the exponential in Eq. (E23). As an extreme example, when $L \ll \xi_{\rm dis}$,

$$E_{\rm dis}(L) \sim -\delta m \frac{L^2}{\xi_{\rm int}^2}. \tag{E23}$$

That is, the change in energy is proportional to the *volume* of the domain. This observation makes sense given that $\mathcal{B}(\boldsymbol{r})$ should be essentially constant for two points within a distance $\xi_{\rm dis}$ of one another. In fact, it seems clear that an energetically favorable domain should be at least $\xi_{\rm dis}$ in extent: $L_* \gtrsim \xi_{\rm dis}$. We therefore examine the threshold scenario given by $L = \xi_{\rm dis}$. We conclude that domain formation is favorable when

$$E_{\rm dom}(\xi_{\rm dis}) \sim U\frac{\xi_{\rm dis}}{\xi_{\rm int}}\left(1 - \frac{\delta m}{U}\frac{\xi_{\rm dis}}{\xi_{\rm int}}\right) = U\frac{\xi_{\rm dis}}{\xi_{\rm int}}(1-\alpha) \lesssim 1. \tag{E24}$$

The parameter α that appeared in the white noise case, Eq. (E15), has showed up again. When it is greater than unity, $\alpha \gtrsim 1$, the disorder destroys long-range order, resulting in domains of typical size $\xi_{\rm dom} \sim \xi_{\rm dis}$.

When $\alpha \lesssim 1$, the interaction energy cost associated with the boundary of a domain of linear extent $\xi_{\rm dis}$ is greater than the gain associated with aligning with the random field. For domains larger than $\xi_{\rm dis}$, the random field within the domain is only weakly correlated. The exponential under the square root may therefore be neglected, resulting in an expression identical to our original estimate for the domain energy with white noise disorder in Eq. (E14). As we discussed there, this expression was not complete: the roughening of the domain walls must also be taken into account, resulting in the contribution given in Eq. (E18). The arguments made in Sec. E 2 follow through for weak, Gaussian-correlated disorder in all respects save for one minor caveat. Unlike the white noise disorder case, the roughening cutoff for Gaussian-correlated disorder is not necessarily $\xi_{\rm int}$. Instead, only scales down to at most $\xi_{\rm dis}$ should be included, since this is where our omission of the exponential ceases to be valid, i.e., $a = \max(\xi_{\rm int}, \xi_{\rm dis})$. Setting the domain energy to zero, we find

$$\xi_{\rm dom} \lesssim \max(\xi_{\rm int}, \xi_{\rm dis})\, e^{c/\alpha^2}, \tag{E25}$$

where $c \sim \mathcal{O}(1)$ is again a nonuniversal constant. In Fig. 9(b), we show $\xi_{\rm dom}$ for Gaussian-correlated disorder for both regimes, $\alpha \lesssim 1$ and $\alpha \gtrsim 1$.

APPENDIX F: COMPETING ORDERS

We now address the possibility considered in Sec. V C that the QVH state is not the ground state of the clean theory at charge neutrality—either one of the other three $\mathcal{C}_2\mathcal{T}$-breaking insulators (QSH, QH, or QSVH) or a completely different order minimizes the energy of the homogeneous theory.

We are interested in studying the conditions under which the QVH phase is realized. To simplify the analysis, we assume that there is a single competing phase whose order parameter does not couple to disorder, but whose ground-state energy density, $\mathcal{E}_{\rm comp}$, is lower than the energy density of the QVH phase, $\mathcal{E}_{\rm QVH}$, by a small amount. We measure this distinction in terms of the energy difference $\delta\epsilon$ within a region of area $\ell_{\rm UV}^2 = \xi_{\rm int}^2$:

$$\frac{\delta\epsilon}{\xi_{\rm int}^2} = \mathcal{E}_{\rm QVH} - \mathcal{E}_{\rm C} \geqslant 0. \tag{F1}$$

Throughout this section, we assume that $\delta\epsilon \ll U$. While this ground-state energy difference implies that the competing phase is realized in a perfectly clean sample, disorder exclusively favors the local realization of the QVH phase. We therefore expect the majority of the sample to be in the QVH phase when $\delta\epsilon$ is sufficiently small. Using the Ising notation

of Sec. V A and Appendix E, we quantify this expectation as

$$\left[\frac{1}{\text{vol}}\int_{\boldsymbol{r}}\langle\phi^2(\boldsymbol{r})\rangle\right]^{1/2} \gtrsim \frac{1}{2}, \tag{F2}$$

where vol denotes the sample volume.

We approach the problem in two complementary fashions. The question of an Ising order parameter competing with another phase may bring to mind dilute Ising physics, where here vacancies represent regions where the Ising ϕ field is not ordered. In Appendix F 1, we describe a mean-field solution of a classical 2D lattice model formulated to tackle this type of question.

While useful, because of the low-dimensionality of the problem, mean-field theory is not particularly reliable. In particular, we are free to take the limit $\delta\epsilon \to -\infty$, effectively removing the competing phase from the problem. In this limit, our results should agree with those of Sec. V A and Appendix E. There, we found that any disorder was sufficient to destroy long-range order. In contrast, the mean-field calculation falsely finds long-range order in this limit. We therefore devise an Ising formulation of the problem in Appendix F 2, which allows us to make Imry-Ma arguments similar to those of Appendix E.

1. Blume-Capel description

In keeping with the Ising description of the QVH insulator, we view the ordering of the competing phase as the presence of an annealed vacancy. At finite temperature, this physics is known to give rise to the tricritical Ising fixed point, though this observation is not relevant for our discussion. While continuum descriptions do exist, for our purposes, it is most convenient to employ a lattice model. We therefore consider the Blume-Capel model [74,75] on an (unspecified) lattice of coordination number z with quenched random-field disorder,

$$H_{\text{BC}} = -\frac{J}{z}\sum_{\langle \boldsymbol{r},\boldsymbol{r}'\rangle} s_{\boldsymbol{r}}s_{\boldsymbol{r}'} + \mu\sum_{\boldsymbol{r}} s_{\boldsymbol{r}}^2 + \sum_{\boldsymbol{r}} h_{\boldsymbol{r}}s_{\boldsymbol{r}}, \tag{F3}$$

where the classical spins may take three values: $s_{\boldsymbol{r}} \in \{+1, -1, 0\}$. As above, the quenched disorder is represented through a random magnetic field $h_{\boldsymbol{r}}$. For simplicity, we assume that $h_{\boldsymbol{r}}$ satisfies Gaussian white noise disorder. The corresponding probability distribution reads

$$\mathcal{P}(h_{\boldsymbol{r}}) = \frac{e^{-\frac{h_{\boldsymbol{r}}^2}{2h_0^2}}}{\sqrt{2\pi h_0^2}}. \tag{F4}$$

The use of this distribution is equivalent to our previous definitions of the disorder distribution, entirely in terms of moments:

$$\overline{h_{\boldsymbol{r}}} = 0, \qquad \overline{h_{\boldsymbol{r}}h_{\boldsymbol{r}'}} = h_0^2\delta_{\boldsymbol{r},\boldsymbol{r}'}. \tag{F5}$$

We associate $s_{\boldsymbol{r}} = \pm 1$ with the realization of the QVH phase, i.e., $\langle\phi\rangle \sim \pm 1$, and vacancies $s_{\boldsymbol{r}} = 0$ with competing phase. The exchange energy J corresponds to the Coulomb interaction strength, $J \sim U$, while the random field strength h_0, should be mapped to the disorder strength $h_0 \sim \delta m\,\xi_{\text{dis}}/\xi_{\text{int}}$ in units of the UV cutoff ξ_{int} [see the discussion below Eq. (E15)]. Finally, the so-called crystal field, μ, can be related to the energy splitting $\delta\epsilon$ by establishing when the competing phase (all $s_{\boldsymbol{r}} = 0$) and QVH phase (all $s_{\boldsymbol{r}} = +1$ or -1) are degenerate, indicating that $\mu = \delta\epsilon + J/2 \sim \delta\epsilon + U/2$.

As discussed, we analyze this model in mean-field theory [76,77]. Letting $m \equiv \langle s_{\boldsymbol{r}}\rangle$ be the average magnetization, the mean-field free energy is

$$\begin{aligned} f_{\text{BC}}(m) &= \frac{1}{2}Jm^2 - \overline{\log(1 + e^{-\beta\mu}2\cosh[\beta(Jm + h)])} \\ &= \frac{1}{2}Jm^2 - \int \frac{dh}{\sqrt{2\pi}h_0}e^{-h^2/2h_0^2} \\ &\quad \times \log(1 + e^{-\beta\mu}2\cosh[\beta(Jm + h)]), \end{aligned} \tag{F6}$$

where β is the inverse temperature and we explicitly average over the Gaussian distribution of Eq. (F4) in the second line. Taking the zero temperature limit, $\beta \to \infty$, the integral can be evaluated exactly, giving

$$\begin{aligned} f_{\text{BC}}(m) &= \frac{1}{2}Jm^2 + \frac{1}{2}\Bigg[(\mu - Jm)\text{Erfc}\left(\frac{\mu - Jm}{\sqrt{2}h_0}\right) \\ &\quad + (\mu + Jm)\text{Erfc}\left(\frac{\mu + Jm}{\sqrt{2}h_0}\right)\Bigg] \\ &\quad - \frac{h_0}{\sqrt{2\pi}}\left(e^{-(\mu - Jm)^2/2h_0^2} + e^{-(\mu + Jm)^2/2h_0^2}\right), \end{aligned} \tag{F7}$$

where $\text{Erfc}(x)$ is the complementary error function. The magnetization is determined by extremizing f_{BC}, resulting in the self-consistency equation

$$m = \frac{1}{2}\left[\text{Erfc}\left(\frac{\mu - Jm}{\sqrt{2}h_0}\right) - \text{Erfc}\left(\frac{\mu + Jm}{\sqrt{2}h_0}\right)\right]. \tag{F8}$$

The expectation value of the spin squared, $q \equiv \sqrt{\langle s_{\boldsymbol{r}}^2\rangle}$, is directly analogous to the expression on the right-hand side of Eq. (F2), i.e., when $q \gtrsim 1/2$, QVH order prevails. It is calculated by taking the derivative of f_{BC} with respect to μ:

$$q^2 = \frac{\partial}{\partial\mu}f_{\text{BC}} = \frac{1}{2}\left[\text{Erfc}\left(\frac{\mu - Jm}{\sqrt{2}h_0}\right) + \text{Erfc}\left(\frac{\mu + Jm}{\sqrt{2}h_0}\right)\right]. \tag{F9}$$

In Figs. 10(a) and 10(c), we plot m and q as functions of $\delta\epsilon/U$ and α, respectively. To make contact with the phase diagram in the main text, Fig. 5, we also plot m and q with the y axis given by $\gamma\,\delta\epsilon/\delta m$, where $\gamma = \xi_{\text{int}}/\xi_{\text{dis}}$, in Figs. 10(b) and 10(d).

Figures 10(a) and 10(b) indicate that m orders for $\delta\epsilon \lesssim 0$ when disorder is sufficiently small. While these calculations agree with our expectations when $\delta m = 0$, we showed in Appendix E that any nonzero disorder destroys long-range order. The presence of regions with $m \neq 0$ is therefore an artifact of the mean-field theory; given the low dimension, the failure of mean-field theory in this regard is not surprising. Nevertheless, we take it as a good sign that m approaches zero close to $\alpha \sim 0.8 \sim 1$ for $\delta\epsilon < 0$ since this condition defines the crossover regime identified in Appendix E. We therefore optimistically associate mean-field-ordered regions with those that in reality possess exponentially large domains.

The density plots in Figs. 10(c) and 10(d) display q. Obviously, when our mean-field prescription indicates that m is ordered, q is nonzero as well, as a quick comparison with

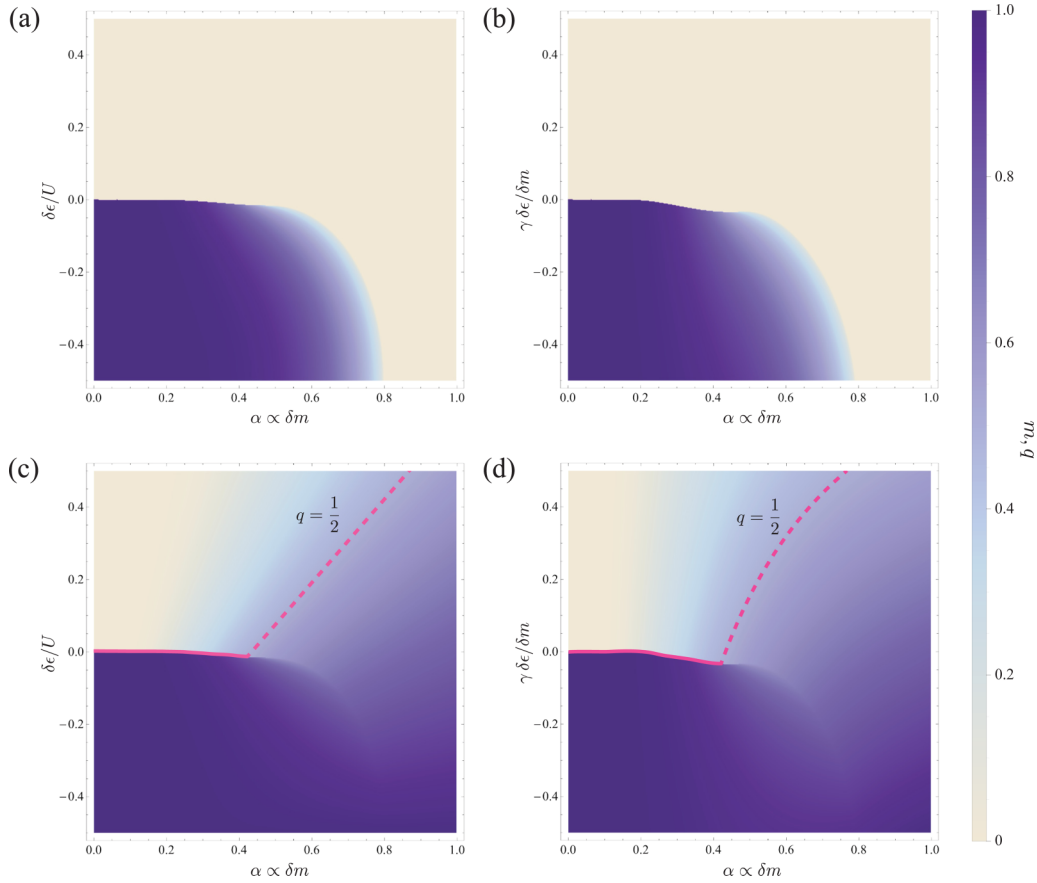

FIG. 10. (a), (b) Density plot of the (absolute value of the) magnetization, obtained by minimizing $f_{\mathrm{BC}}(m)$ in Eq. (F7). (c), (d) Density plots of q, as given in Eq. (F9). The color scheme for all plots, (a)–(d), is shown on the right, and, in (b) and (d), $\gamma = \xi_{\mathrm{int}}/\xi_{\mathrm{dis}}$. The solid pink line in (c) and (d) indicates the first-order phase transition between regions with q small and regions with $q \sim 1$ (as follows from having $m \sim \pm 1$ in that region). The dashed pink line, on the other hand, is the contour along which $q = 1/2$ and $m = 0$; we view it as demarcating a crossover between regions where the competing phase percolates and regions where the QVH insulator percolates. It follows that for both (c) and (d), the network scenario we propose should be valid in the regions below and to the right of the pink lines.

Figs. 10(a) and 10(b) clearly shows. Outside of these regions, however, we find that q only vanishes exactly when $\delta m \to 0$ (equivalently, $h_0 \to 0$) as well. From Eq. (F9), we verify that when $m = 0$,

$$q(m=0) = \sqrt{\mathrm{Erfc}\left(\frac{\mu}{\sqrt{2}h_0}\right)}, \tag{F10}$$

implying that contours of constant q are represented by straight lines extending from the $\mu = 0$ origin (not to be confused with $\delta\epsilon = 0$ origin), as shown in Fig. 10(c). More precisely, we can numerically solve for the line along which $q = 1/2$,

$$\frac{1}{2} = \sqrt{\mathrm{Erfc}\left(\frac{\eta_{1/2}}{\sqrt{2}}\right)}, \tag{F11}$$

to obtain $\eta_{1/2} \cong 1.15$. Then, provided $\delta\epsilon/U$ and α are such that $m = 0$, we find that $q = 1/2$ along the line

$$\frac{\delta\epsilon}{U} = \eta_{1/2}\alpha - \frac{1}{2}. \tag{F12}$$

We plot this contour with a pink dashed line in Fig. 10(c). It follows that the system is primarily in the QVH phase when either

$$\delta\epsilon \lesssim U\left(\eta_{1/2}\alpha - \frac{1}{2}\right) \quad \text{or} \quad \delta\epsilon \lesssim 0. \tag{F13}$$

The modification needed to obtain the crossover lines drawn in Fig. 10(d) is straightforward:

$$\delta\epsilon \lesssim \delta m \frac{\xi_{\mathrm{dis}}}{\xi_{\mathrm{int}}}\left(\eta_{1/2} - \frac{1}{2\alpha}\right) \quad \text{or} \quad \delta\epsilon \lesssim 0. \tag{F14}$$

2. Competing Ising field description

The mean-field theory discussed above had the advantage of simplicity but did not correctly capture the absence of long-range order. We therefore employ an Imry-Ma description, similar to the analysis of Appendix E. The ordering of both phases is now modeled by two distinct Ising fields. As above, we associate ϕ with the QVH insulator (i.e., $\mathcal{C}_2$ symmetry breaking) and Φ with the competing phase. The total energy is given by $H_{\mathrm{Ising}} + H'_{\mathrm{Ising}} + H_{\phi\Phi} + H_{\phi,\mathrm{dis}}$

where

$$H_{\text{Ising}} = \int d^2\boldsymbol{r}\left[\mathcal{K}(\nabla\phi)^2 - \frac{|r|}{2}\phi^2 + \frac{u}{4!}\phi^4\right],$$
$$H'_{\text{Ising}} = \int d^2\boldsymbol{r}\left[\mathcal{K}'(\nabla\Phi)^2 - \frac{|r'|}{2}\Phi^2 + \frac{u'}{4!}\Phi^4\right],$$
$$H_{\phi\Phi} = \int d^2\boldsymbol{r}\,\lambda\,\phi^2\Phi^2,$$
$$H_{\text{dis}} = \int d^2\boldsymbol{r}\,\mathcal{B}(\boldsymbol{r})\phi(\boldsymbol{r}). \tag{F15}$$

Since both ϕ and Φ are dimensionless, $\mathcal{K}$, $\mathcal{K}'$ have dimensions of energy. We assume that the interaction scales of the QVH and competing phases are similar, prompting us to set both to $\sim U$. Similarly, the remaining parameters describing H_{Ising} and H'_{Ising}, r, r', u, and u', have units of energy over length squared. Their natural scale is therefore U/ℓ_{UV}^2, where ℓ_{UV} is the UV cutoff, which should in turn be approximately given by $\xi_{\text{int}} = \hbar v_F/\Delta_{\text{CNP}}$, as discussed in Sec. V A. However, this assignment of energy scales cannot be the entire story since the difference in ground-state energies, Eq. (F1), has not yet been included. Because $\delta\epsilon$ is assumed to be much smaller than U, and we ignore coefficients of $\mathcal{O}(1)$, the exact implementation is unimportant. Nevertheless, to be concrete, we note that if one wishes to ensure that Eq. (F1) holds while also requiring the magnitudes of ϕ and Φ to be identical in their respective ordered phases, the following choice is sufficient:

$$|r'| \sim |r| + \frac{2|r|}{3u}\frac{\delta\epsilon}{\xi_{\text{int}}^2}, \qquad u' \sim u + \frac{2}{3}\frac{\delta\epsilon}{\xi_{\text{int}}^2}. \tag{F16}$$

The parameter λ in $H_{\phi\Phi}$ is assumed to be larger than the other scales of the theory to guarantee that $\langle\phi\rangle \neq 0$ and $\langle\Phi\rangle \neq 0$ do not occur within the same region. Finally, the last term, H_{dis}, describes the behavior of disorder. We will consider both white noise and Gaussian-correlated, as defined in Eqs. (E12) and (E21), respectively.

We examine this system in several steps. Using Imry-Ma type arguments similar to those of Appendix E, we begin by studying the formation of a ϕ-ordered domain within a uniformly Φ-ordered system for both white noise and Gaussian-correlated disorder. As we did in Appendix E, coefficients of $\mathcal{O}(1)$ are ignored. Next, we argue that if the physical parameters favor the formation of a single ϕ-ordered domain, a macroscopically large fraction of the system should also be ϕ-ordered. Our final result is a function of the ratio α [see Eq. (E15)], $\delta\epsilon_c(\alpha)$, that parametrizes a crossover between the two regimes of interest: when $\delta\epsilon \lesssim \delta\epsilon_c(\alpha)$, the system is primarily ϕ-ordered, whereas when $\delta\epsilon \gtrsim \delta\epsilon_c(\alpha)$, the system is primarily Φ-ordered. Figure 5 shows the resulting phase diagram.

a. Single ϕ-domain formation: White noise disorder

To make contact with the mean-field theory of Appendix F 1, we begin by considering white noise disorder. We assume that the competing phase is realized, $\langle\Phi\rangle \neq 0$, and examine the energy cost associated with the formation of a ϕ-ordered domain. As in Appendix E, there are energy contributions from interactions along the domain boundary and from the random field $\mathcal{B}(\boldsymbol{r})$. Since we assume that $\mathcal{K} \sim \mathcal{K}' \sim U$, the interaction energy cost E_{int} is identical to the expression given in Eq. (E2).[6] Similarly, the contribution from disorder, E_{dis}, follows from the expression in Eq. (E6), giving the same result as in Eq. (E13). Unlike Appendix E, there is an important additional cost associated with the difference in ground-state energy. On general grounds, the cost must increase with the domain *area*:

$$E_{\text{comp}}(L) \sim \delta\epsilon\frac{L^2}{\xi_{\text{int}}^2}. \tag{F17}$$

We could also have obtained this result from the Hamiltonian defined in Eq. (F15) with the coefficients defined in Eq. (F16). The total energy cost of a ϕ-ordered domain is given by the sum of this expression with E_{int} and E_{dis}:

$$E_{\phi\text{-dom}}(L) \sim \delta\epsilon\frac{L^2}{\xi_{\text{int}}^2} + U\frac{L}{\xi_{\text{int}}} - \delta m\frac{\xi_{\text{dis}}L}{\xi_{\text{int}}^2}$$
$$= \delta m\frac{\xi_{\text{dis}}L}{\xi_{\text{int}}^2}\left(\frac{\delta\epsilon}{\delta m}\frac{L}{\xi_{\text{dis}}} + \frac{1}{\alpha} - 1\right). \tag{F18}$$

This result is the analog of Eq. (E14). There, we concluded that when $\alpha \gtrsim 1$, disorder was "large" and the system would not order. While this expression also indicates that $\alpha \gtrsim 1$ is necessary to destroy the local order (here, Φ-order instead a different type of ϕ-order), the energy cost of the ϕ domain is also dependent on its size, L: the smaller the domain size, the more favorable it is. A threshold value of $\delta\epsilon$ can therefore be defined by the condition $E_{\phi\text{-dom}}(a) < 0$, where a is the smallest possible domain size. (Again, a should not be confused with the microscopic lattice constant of monolayer graphene here or below). For the current situation, clearly $a \sim \xi_{\text{int}}$; nevertheless, with an eye to the subsequent section, it is convenient to leave a unspecified. That is, $E_{\phi\text{-dom}}(a) < 0$ provided

$$\delta\epsilon \lesssim \delta\epsilon_c(\alpha), \quad \delta\epsilon_c(\alpha) \equiv \delta m\frac{\xi_{\text{dis}}}{a}\left(1 - \frac{1}{\alpha}\right), \quad \text{when} \quad \alpha \gtrsim 1. \tag{F19}$$

Here, we have defined the critical energy difference $\delta\epsilon_c(\alpha)$ in the region where $\alpha \gtrsim 1$ for white noise disorder with a minimal domain size $a = \xi_{\text{int}}$. We generalize this definition to smaller values of α below.

We note that up to coefficients of $\mathcal{O}(1)$, this inequality has the same dependence on α as our mean-field result in Eq. (F14)! At least in the simple regime, the Blume-Capel and Imry-Ma descriptions are in agreement.

As we saw in Appendix E 2, once $\alpha \lesssim 1$, the effects of domain-wall roughening become important and must be included. Because roughening does not change the domain area significantly, the roughening contribution Eq. (E18) remains valid.[7] We note that this situation is similar to what occurs in the absence of a competing order when a small, uniform

[6] One might argue that it is more honest to define $\mathcal{K}' \sim \mathcal{K} + \delta\epsilon \sim U + \delta\epsilon$ in analogy with the definitions of Eq. (F16). However, since $\delta\epsilon \ll U$ by assumption, this difference is negligible.

[7] Alternatively, we can argue that since the displacement is equally likely to increase or decrease the domain area, Eq. (E16) remains valid on average.

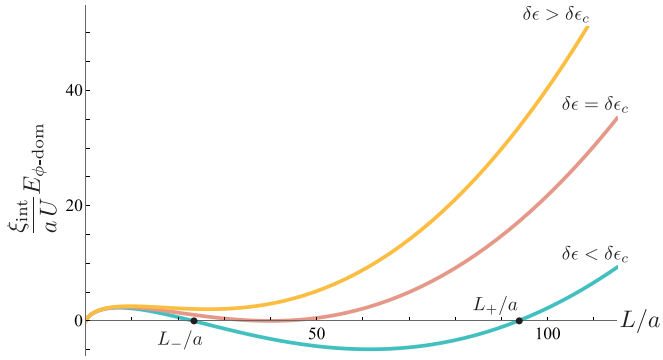

FIG. 11. Plot of the energy cost associated with adding a ϕ-ordered domain to a uniformly Φ-ordered system when $\delta\epsilon > \delta\epsilon_c$ (orange), $\delta\epsilon = \delta\epsilon_c$ (pink), and $\delta\epsilon < \delta\epsilon_c$ (turquoise). For $\delta\epsilon < \delta\epsilon_c$, we see that domain formation is energetically favorable, $E_{\phi\text{-dom}} < 0$, for domains with linear extent L satisfying $L_- < L < L_+$. Here, we have set $\alpha \sim 0.6$, for which $\delta\epsilon_c \sim 0.015\,\delta m\,\xi_{\rm dis}/a$.

magnetic field is applied [72,106]. The resulting cost of a ϕ domain is

$$E_{\phi\text{-dom}}(L) \sim \delta\epsilon\left(\frac{L}{\xi_{\rm int}}\right)^2 + U\frac{L}{\xi_{\rm int}} - U\left(\frac{\delta m}{U}\frac{\xi_{\rm dis}}{\xi_{\rm int}}\right)^2 \frac{L}{\xi_{\rm int}}\log\left(\frac{L}{a}\right) = U\frac{L}{\xi_{\rm int}}\left[\frac{\delta\epsilon}{U}\frac{L}{\xi_{\rm int}} + 1 - \alpha^2\log\left(\frac{L}{a}\right)\right]. \tag{F20}$$

Again, a is the minimal domain size, which is equivalent to $\xi_{\rm int}$ in this case. We can now define a critical energy difference in the small α regime. We find that there exists a solution $E_{\rm dom}(L) = 0$ provided $\delta\epsilon$ satisfies

$$\delta\epsilon \lesssim \delta\epsilon_c(\alpha),\ \ \delta\epsilon_c(\alpha) \equiv \frac{\xi_{\rm dis}}{a}\delta m\,\alpha\, e^{-c\left(\frac{1}{\alpha^2}+1\right)},\ \text{ when }\ \alpha \lesssim 1. \tag{F21}$$

In Fig. 11, we plot $E_{\phi\text{-dom}}(L)$ as a function L for several values of $\delta\epsilon$. As indicated in the figure, when $\delta\epsilon < \delta\epsilon_c$, there is an entire region where $E_{\phi\text{-dom}} < 0$ for $L_- < L < L_+$. Naturally, as $\delta\epsilon \to 0$, $L_- \to L_*$ [as defined in Eq. (E20)] while $L_+ \to \infty$.

b. Single ϕ-domain formation: Gaussian correlated disorder

We now repeat the exercise above for Gaussian-correlated disorder. The energy cost of inserting a ϕ-ordered domain into a uniformly Φ-ordered system is on average

$$E_{\phi\text{-dom}}(L) \sim \delta\epsilon\left(\frac{L}{\xi_{\rm int}}\right)^2 + U\frac{L}{\xi_{\rm int}} - \delta m\frac{\xi_{\rm dis}L}{\xi_{\rm int}^2}\sqrt{1 - e^{-L^2/2\xi_{\rm dis}^2}}. \tag{F22}$$

We first study the regime where the smoothness of the disorder is important, i.e., the exponential under the square root is important. In this case, we expect the ϕ domains to track the disorder potential and therefore be of the same size as the disorder correlation length $\xi_{\rm dis}$. For this to be energetically favorable, we must have

$$0 > E_{\phi\text{-dom}}(\xi_{\rm dis}) \sim \delta\epsilon\frac{\xi_{\rm dis}^2}{\xi_{\rm int}^2} + U\frac{\xi_{\rm dis}}{\xi_{\rm int}} - \delta m\frac{\xi_{\rm dis}^2}{\xi_{\rm int}^2} = U\left(\frac{\xi_{\rm dis}}{\xi_{\rm int}}\right)^2\left[\frac{\delta\epsilon}{\delta m} + \frac{1}{\alpha} - 1\right], \tag{F23}$$

It follows that ϕ-ordered domains of linear extent $\xi_{\rm dis}$ should form once

$$\delta\epsilon \lesssim \delta\epsilon_c(\alpha),\quad \delta\epsilon_c(\alpha) \equiv \delta m\left(1 - \frac{1}{\alpha}\right),\quad \text{when}\quad \alpha \gtrsim 1. \tag{F24}$$

This critical energy difference is nearly identical to the analogous expression obtained for white noise disorder in Eqs. (F14) and (F19). The most notable difference between the two inequalities is the prefactor $\xi_{\rm dis}/\xi_{\rm int}$ multiplying the right-hand side. Going back to the previous section, we see that this coefficient originates from setting the minimal domain size to $\xi_{\rm int}$. In contrast, for Gaussian-correlated disorder, the smallest allowed domains are expected to be $\xi_{\rm dis}$, and so $\delta\epsilon_c(\alpha)$ contains no such prefactor.

As we saw in Appendix E, once $\alpha \lesssim 1$, Gaussian-correlated disorder can be treated as local white-noise disorder, which necessitates a treatment that includes the effects of domain-wall roughening. The relevant expression for $E_{\phi\text{-dom}}(L)$ is therefore identical to the one given in Eq. (F20), save that the smallest domain size is given by $a = \max(\xi_{\rm int}, \xi_{\rm dis})$. The inequality describing the favorability of domain formation is now

$$\delta\epsilon \lesssim \delta\epsilon_c(\alpha),\quad \delta\epsilon_c(\alpha) \equiv \frac{\xi_{\rm dis}}{a}\delta m\,\alpha\, e^{-c\left(\frac{1}{\alpha^2}+1\right)},\quad \text{when}\quad \alpha \lesssim 1. \tag{F25}$$

c. Multiple ϕ-domains

The formation of a single domain does not necessarily imply the network model we propose as a description for mTBG at charge neutrality. Instead, we want the ϕ-ordered regions to percolate throughout the sample, as implied by the condition given in Eq. (F2). We argue that ϕ order should start dominating at a crossover set by the scale $\delta\epsilon_c(\alpha)$. As discussed in Appendix F 2 a, within our approximation, domain boundaries between different ϕ orientations have the same cost as domains between Φ- and ϕ-ordered regions. As a result, we can imagine 'tiling' the ϕ-ordered regions into domains of some size ξ_*. For instance, when $\alpha \lesssim 1$, the energy difference between a uniformly Φ-ordered system and a (nonuniformly) ϕ-ordered system is

$$\Delta E \sim \delta\epsilon\left(\frac{L}{\xi_{\rm int}}\right)^2 + \left(\frac{L}{\xi_*}\right)^2\left[U\frac{\xi_*}{\xi_{\rm int}} - U\left(\frac{\delta m}{U}\frac{\xi_{\rm dis}}{\xi_{\rm int}}\right)^2\frac{\xi_*}{\xi_{\rm int}}\log\left(\frac{\xi_*}{a}\right)\right] = U\frac{L^2}{\xi_{\rm int}\xi_*}\left[\frac{\delta\epsilon}{U}\frac{\xi_*}{\xi_{\rm int}} + 1 - \alpha^2\log\left(\frac{\xi_*}{a}\right)\right] = \frac{L^2}{\xi_*^2}E_{\phi\text{-dom}}(\xi_*), \tag{F26}$$

where this expression is the same for both white noise and Gaussian-correlated disorder provided we recall that $a = \xi_{\text{int}}$ in the former case while $a = \max(\xi_{\text{dis}}, \xi_{\text{int}})$ in the latter. It follows that when the typical domain size ξ_* is such that $E_{\phi\text{-dom}}(\xi_*) < 0$ (i.e., $L_- < \xi_* < L_+$), a wholly (but nonuniformly) ϕ-ordered sample may be considered energetically favorable. An identical argument holds for $\alpha \gtrsim 1$ with $\xi_* = \xi_{\text{dis}}$. If we now imagine fixing α and increasing $\delta\epsilon$, we expect Eq. (F2) to hold up to some value, $\delta\tilde{\epsilon}_c(\alpha)$, of the same order as $\delta\epsilon_c(\alpha)$. Given the general lack of precision throughout this Appendix, we assume that $\delta\tilde{\epsilon}_c(\alpha) \sim \delta\epsilon_c(\alpha)$. This identity sets the dashed line in Fig. 5.

[1] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, Nature (London) **556**, 43 (2018).

[2] Y. Cao, V. Fatemi, A. Demir, S. Fang, S. L. Tomarken, J. Y. Luo, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras, R. C. Ashoori, and P. Jarillo-Herrero, Nature (London) **556**, 80 (2018).

[3] J. M. B. Lopes dos Santos, N. M. R. Peres, and A. H. Castro Neto, Phys. Rev. Lett. **99**, 256802 (2007).

[4] R. Bistritzer and A. H. MacDonald, Proc. Natl. Acad. Sci. **108**, 12233 (2011).

[5] Y. Kim, B. J. Wieder, C. L. Kane, and A. M. Rappe, Phys. Rev. Lett. **115**, 036806 (2015).

[6] H. C. Po, L. Zou, A. Vishwanath, and T. Senthil, Phys. Rev. X **8**, 031089 (2018).

[7] R. de Gail, M. O. Goerbig, F. Guinea, G. Montambaux, and A. H. Castro Neto, Phys. Rev. B **84**, 045436 (2011).

[8] W.-Y. He, Z.-D. Chu, and L. He, Phys. Rev. Lett. **111**, 066803 (2013).

[9] M. Yankowitz, S. Chen, H. Polshyn, Y. Zhang, K. Watanabe, T. Taniguchi, D. Graf, A. F. Young, and C. R. Dean, Science **363**, 1059 (2019).

[10] A. L. Sharpe, E. J. Fox, A. W. Barnard, J. Finney, K. Watanabe, T. Taniguchi, M. A. Kastner, and D. Goldhaber-Gordon, Science **365**, 605 (2019).

[11] M. Serlin, C. L. Tschirhart, H. Polshyn, Y. Zhang, J. Zhu, K. Watanabe, T. Taniguchi, L. Balents, and A. F. Young, Science **367**, 900 (2020).

[12] N. Bultinck, S. Chatterjee, and M. P. Zaletel, Phys. Rev. Lett. **124**, 166601 (2020).

[13] Y.-H. Zhang, D. Mao, and T. Senthil, Phys. Rev. Res. **1**, 033126 (2019).

[14] X. Lu, P. Stepanov, W. Yang, M. Xie, M. A. Aamir, I. Das, C. Urgell, K. Watanabe, T. Taniguchi, G. Zhang, A. Bachtold, A. H. MacDonald, and D. K. Efetov, Nature **574**, 653 (2019).

[15] A. Kerelsky, L. J. McGilly, D. M. Kennes, L. Xian, M. Yankowitz, S. Chen, K. Watanabe, T. Taniguchi, J. Hone, C. Dean, A. Rubio, and A. N. Pasupathy, Nature **572**, 95 (2019).

[16] Y. Choi, J. Kemmer, Y. Peng, A. Thomson, H. Arora, R. Polski, Y. Zhang, H. Ren, J. Alicea, G. Refael, F. von Oppen, K. Watanabe, T. Taniguchi, and S. Nadj-Perge, Nat. Phys. **15**, 1174 (2019).

[17] Y. Jiang, X. Lai, K. Watanabe, T. Taniguchi, K. Haule, J. Mao, and E. Y. Andrei, Nature **573**, 91 (2019).

[18] Y. Xie, B. Lian, B. Jäck, X. Liu, C.-L. Chiu, K. Watanabe, T. Taniguchi, B. A. Bernevig, and A. Yazdani, Nature **572**, 101 (2019).

[19] A. Uri, S. Grover, Y. Cao, J. A. Crosse, K. Bagani, D. Rodan-Legrain, Y. Myasoedov, K. Watanabe, T. Taniguchi, P. Moon, M. Koshino, P. Jarillo-Herrero, and E. Zeldov, Nature **581**, 47 (2020).

[20] J. H. Wilson, Y. Fu, S. Das Sarma, and J. H. Pixley, Phys. Rev. Res. **2**, 023325 (2020).

[21] S. Liu, E. Khalaf, J. Y. Lee, and A. Vishwanath, Phys. Rev. Res. **3**, 013033 (2021).

[22] M. Xie and A. H. MacDonald, Phys. Rev. Lett. **124**, 097601 (2020).

[23] F. Zhang, J. Jung, G. A. Fiete, Q. Niu, and A. H. MacDonald, Phys. Rev. Lett. **106**, 156801 (2011).

[24] I. Martin, Y. M. Blanter, and A. F. Morpurgo, Phys. Rev. Lett. **100**, 036804 (2008).

[25] Z. Qiao, J. Jung, Q. Niu, and A. H. MacDonald, Nano Lett. **11**, 3453 (2011).

[26] R. V. Gorbachev, J. C. W. Song, G. L. Yu, A. V. Kretinin, F. Withers, Y. Cao, A. Mishchenko, I. V. Grigorieva, K. S. Novoselov, L. S. Levitov, and A. K. Geim, Science **346**, 448 (2014).

[27] M. Sui, G. Chen, L. Ma, W.-Y. Shan, D. Tian, K. Watanabe, T. Taniguchi, X. Jin, W. Yao, D. Xiao, and Y. Zhang, Nat. Phys. **11**, 1027 (2015).

[28] Y. Shimazaki, M. Yamamoto, I. V. Borzenets, K. Watanabe, T. Taniguchi, and S. Tarucha, Nat. Phys. **11**, 1032 (2015).

[29] L. Ju, Z. Shi, N. Nair, Y. Lv, C. Jin, J. Velasco, C. Ojeda-Aristizabal, H. A. Bechtel, M. C. Martin, A. Zettl, J. Analytis, and F. Wang, Nature (London) **520**, 650 (2015).

[30] J. Li, K. Wang, K. J. McFaul, Z. Zern, Y. Ren, K. Watanabe, T. Taniguchi, Z. Qiao, and J. Zhu, Nat. Nanotechnol. **11**, 1060 (2016).

[31] L.-J. Yin, H. Jiang, J.-B. Qiao, and L. He, Nat. Commun. **7**, 11760 (2016).

[32] D. A. Abanin, S. A. Parameswaran, S. A. Kivelson, and S. L. Sondhi, Phys. Rev. B **82**, 035428 (2010).

[33] S. A. Parameswaran and B. E. Feldman, J. Phys.: Condens. Matter **31**, 273001 (2019).

[34] D. F. Mross, Y. Oreg, A. Stern, G. Margalit, and M. Heiblum, Phys. Rev. Lett. **121**, 026801 (2018).

[35] C. Wang, A. Vishwanath, and B. I. Halperin, Phys. Rev. B **98**, 045112 (2018).

[36] B. Lian and J. Wang, Phys. Rev. B **97**, 165124 (2018).

[37] J. T. Chalker and P. D. Coddington, J. Phys. C **21**, 2665 (1988).

[38] D. K. K. Lee and J. T. Chalker, Phys. Rev. Lett. **72**, 1510 (1994).

[39] D. K. K. Lee, J. T. Chalker, and D. Y. K. Ko, Phys. Rev. B **50**, 5272 (1994).

[40] C. M. Ho and J. T. Chalker, Phys. Rev. B **54**, 8708 (1996).

[41] D.-H. Lee, Phys. Rev. B **50**, 10788 (1994).

[42] N. N. T. Nam and M. Koshino, Phys. Rev. B **96**, 075311 (2017).

[43] M. Koshino, N. F. Q. Yuan, T. Koretsune, M. Ochi, K. Kuroki, and L. Fu, Phys. Rev. X **8**, 031087 (2018).
[44] H. C. Po, L. Zou, T. Senthil, and A. Vishwanath, Phys. Rev. B **99**, 195455 (2019).
[45] J. Kang and O. Vafek, Phys. Rev. X **8**, 031088 (2018).
[46] L. Zou, H. C. Po, A. Vishwanath, and T. Senthil, Phys. Rev. B **98**, 085435 (2018).
[47] M. R. Zirnbauer, J. Math. Phys. **37**, 4986 (1996).
[48] A. Altland and M. R. Zirnbauer, Phys. Rev. B **55**, 1142 (1997).
[49] A. W. W. Ludwig, M. P. A. Fisher, R. Shankar, and G. Grinstein, Phys. Rev. B **50**, 7526 (1994).
[50] P. M. Ostrovsky, I. V. Gornyi, and A. D. Mirlin, Phys. Rev. Lett. **98**, 256801 (2007).
[51] A. F. Morpurgo and F. Guinea, Phys. Rev. Lett. **97**, 196804 (2006).
[52] A. Pruisken, Nucl. Phys. B **235**, 277 (1984).
[53] A. M. M. Pruisken, Field theory, scaling and the localization problem, in *The Quantum Hall Effect*, edited by R. E. Prange and S. M. Girvin (Springer New York, 1990), pp. 117–173.
[54] I. L. Aleiner and K. B. Efetov, Phys. Rev. Lett. **97**, 236801 (2006).
[55] A. Altland, Phys. Rev. Lett. **97**, 236802 (2006).
[56] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).
[57] P. A. Lee and T. V. Ramakrishnan, Rev. Mod. Phys. **57**, 287 (1985).
[58] E. Fradkin, Phys. Rev. B **33**, 3257 (1986).
[59] E. Fradkin, Phys. Rev. B **33**, 3263 (1986).
[60] S. Carr, S. Fang, Z. Zhu, and E. Kaxiras, Phys. Rev. Res. **1**, 013001 (2019).
[61] F. D. M. Haldane, Phys. Rev. Lett. **61**, 2015 (1988).
[62] C. Weeks and M. Franz, Phys. Rev. B **81**, 085105 (2010).
[63] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 226801 (2005).
[64] G. Tarnopolsky, A. J. Kruchkov, and A. Vishwanath, Phys. Rev. Lett. **122**, 106405 (2019).
[65] Y.-H. Zhang, D. Mao, Y. Cao, P. Jarillo-Herrero, and T. Senthil, Phys. Rev. B **99**, 075127 (2019).
[66] J. Y. Lee, E. Khalaf, S. Liu, X. Liu, Z. Hao, P. Kim, and A. Vishwanath, Nat. Commun. **10**, 5333 (2019).
[67] N. Bultinck, E. Khalaf, S. Liu, S. Chatterjee, A. Vishwanath, and M. P. Zaletel, Phys. Rev. X **10**, 031034 (2020).
[68] V. S. Dotsenko and V. S. Dotsenko, Adv. Phys. **32**, 129 (1983).
[69] Y. Imry and S.-k. Ma, Phys. Rev. Lett. **35**, 1399 (1975).
[70] T. Nattermann and J. Villain, Phase Transit. **11**, 5 (1988).
[71] T. Nattermann, *Spin Glasses and Random Fields*, Series on Directions in Condensed Matter Physics, Vol. 12 (World Scientific, Singapore, 1977), p. 277.
[72] K. Binder, Z. Phys. B **50**, 343 (1983).
[73] M. Aizenman and J. Wehr, Phys. Rev. Lett. **62**, 2503 (1989).
[74] M. Blume, Phys. Rev. **141**, 517 (1966).
[75] H. Capel, Physica **32**, 966 (1966).
[76] M. Kaufman and M. Kanner, Phys. Rev. B **42**, 2378 (1990).
[77] R. Vasseur and T. Lookman, Phys. Rev. B **81**, 094107 (2010).
[78] Y.-Z. Chou, R. M. Nandkishore, and L. Radzihovsky, Phys. Rev. B **99**, 165108 (2019).
[79] F. Evers and A. D. Mirlin, Rev. Mod. Phys. **80**, 1355 (2008).
[80] D. K. Efimkin and A. H. MacDonald, Phys. Rev. B **98**, 035404 (2018).
[81] X.-C. Wu, C.-M. Jian, and C. Xu, Phys. Rev. B **99**, 161405(R) (2019).
[82] Y.-Z. Chou, Y.-P. Lin, S. Das Sarma, and R. M. Nandkishore, Phys. Rev. B **100**, 115128 (2019).
[83] C. Chen, A. H. Castro Neto, and V. M. Pereira, Phys. Rev. B **101**, 165431 (2020).
[84] C. R. Woods, L. Britnell, A. Eckmann, R. S. Ma, J. C. Lu, H. M. Guo, X. Lin, G. L. Yu, Y. Cao, R. V. Gorbachev, A. V. Kretinin, J. Park, L. A. Ponomarenko, M. I. Katsnelson, Y. N. Gornostyrev, K. Watanabe, T. Taniguchi, C. Casiraghi, H. J. Gao, A. K. Geim, and K. S. Novoselov, Nat. Phys. **10**, 451 (2014).
[85] M. M. van Wijk, A. Schuring, M. I. Katsnelson, and A. Fasolino, 2D Mater. **2**, 034010 (2015).
[86] S. Dai, Y. Xiang, and D. J. Srolovitz, Nano Lett. **16**, 5923 (2016).
[87] N. Y. Kim, H. Y. Jeong, J. H. Kim, G. Kim, H. S. Shin, and Z. Lee, ACS Nano **11**, 7084 (2017).
[88] F. Gargiulo and O. V. Yazyev, 2D Mater. **5**, 015019 (2018).
[89] K. Zhang and E. B. Tadmor, J. Mech. Phys. Solids **112**, 225 (2018).
[90] Cory Dean (private communication).
[91] D. A. Abanin, A. V. Shytov, L. S. Levitov, and B. I. Halperin, Phys. Rev. B **79**, 035304 (2009).
[92] D. R. Hofstadter, Phys. Rev. B **14**, 2239 (1976).
[93] J. Tworzydło, B. Trauzettel, M. Titov, A. Rycerz, and C. W. J. Beenakker, Phys. Rev. Lett. **96**, 246802 (2006).
[94] N. M. R. Peres, A. H. Castro Neto, and F. Guinea, Phys. Rev. B **73**, 195411 (2006).
[95] S. Das Sarma, S. Adam, E. H. Hwang, and E. Rossi, Rev. Mod. Phys. **83**, 407 (2011).
[96] T. Ando, J. Phys. Soc. Jpn. **75**, 074716 (2006).
[97] V. V. Cheianov and V. I. Fal'Ko, Phys. Rev. Lett. **97**, 226801 (2006).
[98] K. Nomura and A. H. MacDonald, Phys. Rev. Lett. **96**, 256602 (2006).
[99] E. H. Hwang, S. Adam, and S. Das Sarma, Phys. Rev. Lett. **98**, 186806 (2007).
[100] K. Nomura and A. H. MacDonald, Phys. Rev. Lett. **98**, 076602 (2007).
[101] M. Trushin and J. Schliemann, Europhys. Lett. **83**, 17001 (2008).
[102] M. I. Katsnelson, F. Guinea, and A. K. Geim, Phys. Rev. B **79**, 195426 (2009).
[103] M. I. Katsnelson and A. K. Geim, Philos. Trans. R. Soc. London Ser. A **366**, 195 (2008).
[104] T. Nattermann, J. Phys. C **16**, 6407 (1983).
[105] E. T. Seppälä, V. Petäjä, and M. J. Alava, Phys. Rev. E **58**, R5217 (1998).
[106] E. T. Seppälä and M. J. Alava, Phys. Rev. E **63**, 066109 (2001).