The Effect of Local Dark Matter Substructure on Constraints in Sommerfeld-Enhanced Models

Tracy R. Slatyer,[1,*] Natalia Toro,[2,1,†] and Neal Weiner[3,1,‡]

[1] *School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA*
[2] *Perimeter Institute, Waterloo ON, Canada*
[3] *Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA*

In models of dark matter with Sommerfeld-enhanced annihilation, where the annihilation rate scales as the inverse velocity, models of local substructure motivated by N-body dark matter simulations imply a local annihilation signal dominated by small, dense, cold subhalos. This contrasts with the usual assumption of a signal originating from the smooth dark matter halo, with much higher velocity dispersion. Accounting for local substructure modifies the favored parameter regions for Sommerfeld-enhanced annihilating DM explanations for the PAMELA and *Fermi* excesses. Limits from the inner galaxy and the CMB are weakened, without introducing new tension with substructure-dependent limits, such as from dwarf galaxies or isotropic gamma-ray studies. With substructure, previously excluded parameter regions with mediators of mass $m_\phi \lesssim 200$ MeV are now easily allowed. For very light mediators, subhalos in a specific range of host halo masses may be evaporated, further suppressing diffuse signals without affecting substructure in the Milky Way.

PACS numbers: 95.35.+d

I. INTRODUCTION

Interest in dark matter (DM) annihilation has been boosted in recent years as a consequence of a number of results from cosmic ray experiments. The PAMELA finding [1] of a rise in the positron fraction at high ($\sim 10 - 100$ GeV) energies coupled with harder than expected $e^+ + e^-$ spectra from *Fermi* [2, 3] and ATIC [4, 5] point to the existence of a new, primary source of high energy e^+e^-.

Attempts to explain the observations with dark matter annihilation are immediately confronted by the large rate (much larger than expected for a thermal relic), the absence of an associated antiproton signal, and the hardness of the positrons. The size of the signal, in particular, is a challenge as enhancements from dark matter substructure are typically expected (from N-body simulations) to be $\mathcal{O}(1)$, while cross sections $\mathcal{O}(100 - 1000)$ larger than thermal are required. Models of TeV-scale dark matter with light ($\lesssim$ GeV) mediators [6–8] seek to address these issues by kinematically forbidding the antiprotons, producing boosted positrons (giving rise to hard spectra), and finally raising the cross section at low velocities via the Sommerfeld enhancement[88]. Such models can explain the CR excesses while still yielding the appropriate relic abundance, and without appealing to substructure [9].

Nonetheless, it is worth reexamining the role of substructure in models with a Sommerfeld enhancement. Because bound subhalos typically have velocity dispersions much smaller than the ∼150 km/s of the smooth halo, they can be the overwhelmingly dominant contributors to local signals in such models [10–17]. As cross sections typically grow at low velocities as v^{-1} or faster (down to some saturation velocity) in these models, even $\mathcal{O}(1)$ contributions to the density-squared integral (and hence to the annihilation rate for conventional WIMPs) can change the predicted Sommerfeld-enhanced WIMP annihilation rate by an order of magnitude or more. We show in Figure 1 the relative size of the boost at saturation compared to its value in the smooth halo. For heavier mediator masses, the difference is only a factor of a few (except on resonances), but for $m_\phi \sim 100$ MeV the saturated enhancement is generically a factor of 10-100 times larger than the smooth-halo boost.

Many attempts to constrain Sommerfeld-enhanced models take the local CR excesses, and assume they are dominated by annihilation in the smooth halo, to provide the expected normalization for signals in constraining channels where no or little excess is seen. It is important to revisit these limits in the natural situation that the local signal is dominated instead by Sommerfeld-enhanced substructure. As we shall see, a wide range of constraints become dramatically weaker in this case, specifically:

*Electronic address: tslatyer@ias.edu
†Electronic address: ntoro@perimeterinstitute.ca
‡Electronic address: neal.weiner@nyu.edu

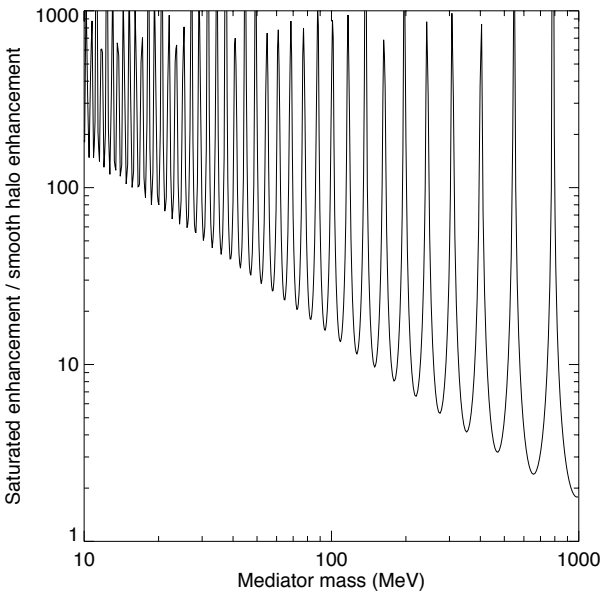

FIG. 1: The ratio of the saturated Sommerfeld enhancement to the smooth halo enhancement as a function of force carrier mass, assuming a local 1D velocity dispersion of $\sigma = 150$ km/s and a dark matter mass of 1.2 TeV.

- ICS and FSR signals from annihilation in galactic center,
- Effects on the CMB power spectrum from annihilations in the era of recombination,
- Constraints on light mediators from DM self-interactions,
- Tensions between the relic abundance and the large local signal.

In particular, the parameter space with force carrier masses below ~ 200 MeV, which is effectively ruled out in the smooth-halo-dominated case, is re-opened in the presence of an $\mathcal{O}(1)$ substructure contribution to the dark matter density-squared. This is an important point. Terrestrial fixed-target experiments and low-energy colliders can probe direct production of light hidden-sector gauge bosons [18]. Of these, collider searches [19–22] can cover the widest range of gauge boson masses, while fixed-target experiments are sensitive to the widest range of couplings, but at lower masses [21, 23, 24]. Most fixed-target results [25–32] and recent proposals [33–38] have the greatest reach at masses below ~ 200 MeV, the region where substructure effects can be most dramatic. Understanding the role of substructure is essential to clarify what regions of parameter space have astrophysical motivation.

In this note, we will explore the added parameter space that is opened at light ($\lesssim 200$ MeV) mediator masses when substructure is taken into account. In section II, we review the constraints that do not depend on substructure, specifically bounds from measurements of the CMB and limits on self-interaction of dark matter. In section III we parameterize the boost to the local density-squared integral from substructure by a single number Δ, and study how these constraints weaken when low-velocity enhancements to annihilation boost the Δ term further. In section IV, we address constraints that themselves rely upon the properties of substructure, in particular limits from diffuse γ rays, arising from substructure in the outer Milky Way and other halos, and limits from the inner Milky Way, which rely upon the amplitude of substructure locally, and the extent to which it persists in the inner galaxy.

II. A REVIEW OF CONSTRAINTS ON SOMMERFELD ENHANCED ANNIHILATION

A. Inner Galaxy limits

N-body simulations of cold dark matter structure formation predict a DM density profile with a pronounced peak in the Galactic center. Gamma-ray annihilation in this cusp can be used to set strong constraints on DM annihilation [39–47], albeit with large astrophysical uncertainties.

The gamma-ray signal from dark matter annihilation has two components: (1) photons produced in the annihilation itself, either as final state radiation (FSR) or from decays of neutral pions, and (2) starlight, infrared and CMB photons

which are inverse Compton scattered to gamma-ray energies by high-energy e^+e^-. We label these two components as "FSR" and "ICS" respectively. Both components depend on the DM density profile and the velocity profile in Sommerfeld-enhanced models; the ICS component tends to provide much stronger constraints than FSR alone for models fitting the PAMELA and *Fermi* excesses, but such limits rely on an accurate model for cosmic ray propagation.

The most recent conservative analyses [45–47] indicate that models fitting the CR data remain allowed if the final state consists of electrons and/or muons, and the DM density profile has a shallow core, similar to the "cored isothermal" or Burkert profiles which appear to be favored by observations (see e.g. [48] for a recent review). However, there is severe tension for DM density profiles possessing a central cusp, like the Einasto and NFW profiles favored by N-body simulations, and a recent analysis using less conservative assumptions finds tension even for a cored profile [49]. These analyses assume that the local and inner Galaxy signals are both dominated by the smooth halo, and the annihilation cross section does not change with Galactocentric radius.

While some authors have included models for the velocity dispersion motivated by N-body simulations (see e.g. [50, 51]), the presence of baryons is expected to significantly affect the density and velocity profiles of DM in the inner Galaxy, and even the sign of the effect is not clear (see e.g. [52–61]). For the ICS signal, which is essential to obtaining the strongest constraints on models which do not produce copious neutral pions, the magnetic field of the inner Galaxy plays an important role, and the presence of gamma-ray structures suggesting a possible large-scale high-energy outflow from the Galactic Center [62] calls into question the usual steady-state modeling of CR propagation in this region of the sky. Nonetheless, for non-cored DM density profiles the FSR signal alone is sufficient to rule out some DM explanations for the CR excesses, and so the relaxation of these constraints would be of consequence.

B. Constraints on the saturated enhancement

The Sommerfeld enhancement due to a Yukawa potential scales as $\pi\alpha_D/v$ in the regime where $v/c \gtrsim m_\phi/m_\chi$, where m_ϕ is the mass of the force carrier, m_χ is the mass of the DM, and α_D is the coupling of the DM to the force carrier. At $v/c \sim m_\phi/m_\chi$ the enhancement saturates due to the finite range of the force; a good estimate for the saturated enhancement is $12\alpha_D m_\chi/m_\phi(1-\cos\theta)$, where θ is a (real) function of $\alpha_D m_\chi/m_\phi$ that describes the resonance structure [63, 64]. Points where $\theta = 2\pi n$ correspond to values of $\alpha_D m_\chi/m_\phi$ where the potential has a zero-energy bound state; close to these resonances, the enhancement instead scales as $\sim 1/v^2$ up to saturation. Thus, the saturated low-velocity enhancement exceeds the enhancement at intermediate velocities v (as in the smooth local halo) by a ratio $\sim 4m_\chi v/m_\phi(1-\cos\theta) \geq 2m_\chi v/m_\phi$.

Some of the most stringent constraints on the Sommerfeld enhancement are obtained by placing upper bounds on the saturated enhancement from examining systems where the typical DM velocity is very small (in particular, dwarf galaxies and the early universe), and then exploiting the ratio between the enhancement in the smooth local halo and the saturated enhancement to set strong limits on the local annihilation rate (e.g. [65–67] for limits from the CMB, [68] for dwarf galaxy bounds). This technique is especially powerful for small mediator masses, due to the $1/m_\phi$ scaling of the ratio, and on resonances, as shown in Figure 1.

In the absence of substructure, the strongest constraints in this category arise from measurements of the cosmic microwave background (CMB)[89]. Dark matter annihilation during the epoch of recombination injects ionizing electrons and photons which broaden the last scattering surface and give rise to increased damping of temperature anisotropies, combined with enhanced polarization anisotropies [65]. The annihilation cross section can therefore be constrained by high-precision measurements of the CMB. The typical velocity of WIMPs at $z \sim 1000$ is of order $v \sim 10^{-8}c$ [66], so we expect the Sommerfeld enhancement to be saturated; the bound from *WMAP* 5 can be summarized as $\langle\sigma v\rangle_{\rm sat} \lesssim (120/f)\,(m_\chi/1{\rm TeV})\,3\times10^{-26}{\rm cm}^3/{\rm s}$, where $f \sim 0.2-0.7$ is an efficiency factor depending on the annihilation final state (see [67] for details)[90]. This bound is typically only a factor of $\sim 2-3$ higher than the local annihilation rate required to fit the CR excesses; consequently, if the smooth halo dominates the local signal, mediator masses lighter than ~ 200 MeV can be ruled out, at least for simple Sommerfeld models [9].

C. Relic density limits

There has recently been some debate over whether models with Sommerfeld-enhanced annihilation can provide a large enough cross section to fit the local CR data at all, while remaining consistent with the measured DM relic density. Assuming the signal originates entirely from the smooth halo, [69] found that the maximal local enhancement was too low by a factor of ~ 15 to explain the CR signals, assuming a 2.35 TeV DM candidate annihilating through a light force carrier solely into muons (providing a good spectral fit to the data), with a local halo density of 0.3 GeV/cm^3. Using a more up-to-date local density estimate of 0.4 GeV/cm^3, and specific models where the light force carrier was a vector and decayed to SM states according to their charge (via kinetic mixing with the photon), [9]

found that the discrepancy was a factor of ~ 3 or less (depending on the DM mass and final state) if the states in the dark-charged DM multiplet were taken to be degenerate, and that there was no discrepancy if a small splitting ($\sim 0.1 - 1$ MeV) between the states was permitted.

D. Self-interaction limits

The longevity and morphology of various astrophysical systems constrain the self-interaction cross-sections of dark matter. Several authors have noted that because dark matter self-interaction is also subject to Sommerfeld enhancement at low velocities, these self-interaction bounds are relevant to theories with Sommerfeld-enhanced annihilation. Feng, Kaplinghat, and Yu independently examined the constraints from elliptic galaxies [70]. Buckley and Fox ([71] and references therein) identify seven classes of constraints, from observations of the Bullet Cluster, evaporation of galaxies and dwarf galaxies, the stability of elliptical cores in galaxy clusters, the growth rate of supermassive black holes, thermodynamics of galaxies, and the structure of dwarf galaxies.

Each of these constraints can be formulated in terms of a velocity-averaged *transfer cross-section* at an appropriate characteristic velocity, which for a particle of mass m_χ in a Yukawa potential controlled by the mediator mass m_ϕ and coupling α_D is well approximated by [72],

$$\sigma_T \approx \begin{cases} \frac{4\pi}{m_\phi^2}\beta^2 \ln(1+\beta^{-1}), & \beta < 0.1, \\ \frac{8\pi}{m_\phi^2}\beta^2/(1+1.5\beta^{1.65}), & 0.1 \leq \beta \leq 1000, \\ \frac{\pi}{m_\phi^2}\left(\ln\beta + 1 - \frac{1}{2}\ln^{-1}\beta\right)^2, & \beta > 1000, \end{cases} \tag{1}$$

where $\beta = 2\alpha_D m_\phi/(m_\chi v_{\rm rel}^2)$. Limits on σ_T in systems of different velocity dispersions can be formulated as upper limits on α_D, as a function of m_ϕ and m_χ, and compared to one another.

The shape of dwarf galaxies presents a particularly strong potential constraint on Sommerfeld-enhanced models, because it depends on self-interaction at velocities as low as 10 km/s, where the Sommerfeld enhancement for light mediators is significant. Transfer cross-sections $\sigma/m_\chi \gtrsim 0.1$ cm^2/ g at velocity dispersions $v_0 \approx 10$ km/s are expected to cause significant departures in halo structure from cold DM models (see e.g. [73]).

Transfer cross-sections above this "self-interaction threshold" may not be ruled out, but would at least have significant effects on the structure of dwarf galaxies. Indeed, it has even been argued that such a velocity-dependent force can explain the origin of cores in dwarf galaxies [74][91]. This threshold cross-section is reached even for tiny couplings at low mediator masses, but rapidly becomes irrelevant at high mediator masses. For example, for $m_\chi = 1$ TeV and mediators masses below 7 MeV, $\sigma/m_\chi \lesssim 0.1$ cm^2/ g requires $\alpha_D \lesssim 10^{-4}$, but for larger mediator masses, it requires

$$\alpha_D \lesssim 0.023 \times \left(\frac{20\ \text{MeV}}{m_\phi}\right)^{4.7} \qquad (7 \lesssim m_\phi \lesssim 20\ \text{MeV}). \tag{2}$$

Thus, self-interaction effects are typically not significant for Sommerfeld-enhanced explanations of the PAMELA and *Fermi* excesses with $m_\phi > 20$ MeV. While this "bound" is much weaker than that arising from the cosmic microwave background, we shall show that it becomes the leading constraint on light-mediator models with significant substructure.

A weaker, but more robust constraint arises from the *evaporation* of dwarf galaxies, which however depends on the velocity dispersion of the host galaxy. The presence of dwarf galaxies in the Milky Way implies a bound $\sigma_T/m_\chi \lesssim 0.1$ cm^2/ g at velocities $v_0 \approx 100$ km/s. This bound permits α_D approximately 100 times larger than the self-interaction threshold ($\alpha_D \lesssim 0.01$ for $m_\phi < 20$ MeV). Even in the presence of significant substructure, this constraint is weaker than the one from the CMB.

III. INDIRECT DM SIGNALS AND CONSTRAINTS IN THE PRESENCE OF SUBSTRUCTURE

The presence of substructure can have a dramatic impact on the preferred ranges of parameters needed to fit PAMELA and *Fermi*. This will immediately affect a number of constraints that have confronted models previously, namely, achieving the appropriate relic abundance, as well as limits from the CMB and self-interaction of dark matter.

An appropriate relic abundance can be readily achieved in these models in conjunction with a fit to the CR excesses, although this constrains the value of α_D when other parameters are fixed. On the other hand, the CMB and self-interaction constraints can be quite restrictive. While both are independent of the presence of substructure, they are nonetheless sensitive to the same low-velocity physics that is relevant for Sommerfeld enhancement in subhalos.

We must be certain that invoking a large low-velocity boost for substructure does not place us in conflict with these observations.

We thus proceed to consider the circumstances under which large low-velocity Sommerfeld enhancement in substructure can give an important contribution to the observed positron signals, but does not imply too-large effects on the CMB or measures of self-interaction.

A. Smooth halo vs substructure-dominated scenarios

The effect of local substructure on a wide variety of constraints can be understood through a simple parametrization, treating the amount of local substructure as a free parameter. We write $1 + \Delta(r) = \langle\rho^2(r)\rangle/\langle\rho(r)\rangle^2$, where $r =$ Galactocentric radius, and defer discussion of the expected value of Δ to the next section. When substructure boosts are non-negligible, most of the substructure signal comes from small dense subhalos with saturated Sommerfeld enhancements. The enhancement factor from substructure and Sommerfeld enhancement can then be written as

$$\mathrm{S_{eff}} \approx S_{v(r)} + S_{v\to 0}\Delta(r), \tag{3}$$

where S_v is the Sommerfeld enhancement factor at velocity v. Generally, one or the other term will dominate, in which case we are either "smooth halo" dominated (former term), or substructure dominated (latter term). For cosmic-ray signals, propagation of the CRs from the point of annihilation means that strictly the relevant substructure boost is given by $\Delta(r)$ averaged over some volume. However, for the energies relevant for the PAMELA and *Fermi* signals, most of the observed positrons come from within 1 kpc [75], and it is reasonable to approximate this average by the local value $\Delta(8.5\mathrm{kpc})$.

The term "boost factor" is commonly used to described any number of enhancements to the cross section, but generally refers to either the boost from substructure compared to a smooth halo, or the boost by the Sommerfeld enhancement relative to an uncorrected s-wave cross section. We will attempt to clearly distinguish between the various "boosts" (in particular because both of these contributions will vary from place to place). We define the general "boost factor" (BF) as the enhanced annihilation rate $\langle\sigma v\rho^2\rangle$ divided by the canonical $(3\times10^{-26}\ \mathrm{cm^3/s})$ $\times\langle\rho\rangle^2$, from all combined effects. Then we obtain the relation,

$$\mathrm{BF} = \mathrm{BF_{smooth}}\left(1 + \frac{S_{v\to 0}}{S_{v(r)}}\Delta(r)\right) = \frac{\langle\sigma v\rangle_{v\sim 150\mathrm{km/s}}}{3\times10^{-26}\mathrm{cm^3/s}}\left(1 + \frac{S_{v\to 0}}{S_{v(r)}}\Delta(r)\right). \tag{4}$$

If the second term (proportional to $\Delta(r)$) is dominant locally, then one needs to understand its scaling with r – not just that of the smooth $\rho(r)^2$ – to understand how limits are affected.

Consider constraints from the inner Galaxy (or another system where the substructure has been entirely disrupted and the characteristic velocity is quite high); let us assume that $\Delta(r) = 0$ there, i.e. all the substructure has been disrupted. Then the ratio,

$$\frac{\mathrm{BF_{GC}}}{\mathrm{BF_{local}}} = \frac{\mathrm{BF_{GC,smooth}}}{\mathrm{BF_{local,smooth}}}\left(1 + \frac{S_{v\to 0}}{S_{v\sim 150\mathrm{km/s}}}\Delta(8.5\mathrm{kpc})\right)^{-1} = \frac{S_{v(r=0)}}{S_{v\sim 150\mathrm{km/s}} + S_{v\to 0}\Delta(8.5\mathrm{kpc})} \to \frac{1}{\Delta(8.5\mathrm{kpc})}\frac{S_{v(r=0)}}{S_{v\to 0}}, \tag{5}$$

is in general not equal to one as most studies of the inner Galaxy limits have assumed. In the substructure-dominated case, the final expression of (5) approximates the ratio of boosts, which can easily weaken constraints from the inner Galaxy by up to three orders of magnitude (see Figure 1) even for moderate $\Delta(8.5\mathrm{kpc}) \sim 0.1 - 1$.

Constraints from systems where the Sommerfeld enhancement is already saturated must also be modified to account for local substructure, and behave quite differently depending on whether the smooth halo or substructure dominates the local signal. Specifically,

$$\frac{\mathrm{BF_{sat}}}{\mathrm{BF_{local}}} = \frac{S_{v\to 0}}{S_{v\sim 150\mathrm{km/s}} + S_{v\to 0}\Delta(8.5\mathrm{kpc})} \to \begin{cases} S_{v\to 0}/S_{v\sim 150\mathrm{km/s}} & \text{smooth-halo-dominated} \\ 1/\Delta(8.5\mathrm{kpc}) & \text{substructure-dominated} \end{cases} \tag{6}$$

The ratio applicable when the smooth halo dominates can be very small, particularly for mediator masses below 100 MeV, making bounds from the CMB particularly constraining of these models. In the substructure-dominated case, by contrast, Sommerfeld-enhanced models behave like models with a large but velocity-independent annihilation cross section, with a local density-squared rescaled by Δ. The resulting constraints are potentially orders of magnitude weaker than one would have thought by considering only the smooth component of the local halo.

B. Consistent scenarios for thermal freeze-out

In general, in the presence of a dominant substructure contribution, the value of α_D that produces the CR excesses is too *small* to generate the observed relic density by thermal freezeout, at least in the simplest models (in contrast to the smooth-halo case with larger DM mass and different decay modes studied in [69]). The benchmarks given in [9], which achieve the correct relic density and local boost factor assuming the entire signal originates from the smooth halo, are generally not appropriate for the substructure-dominated case because they overproduce the CR signal.

Instead, one is led to consider models where additional annihilation channels are important during freeze-out, which may or may not be relevant for the CR excesses today. Inclusion of such processes significantly affects which models are allowed, by breaking the linkage between the Sommerfeld enhancement and the annihilation cross section, and hence allowing extra depletion of the relic density while still producing the desired CR signal. To illustrate the interplay between CMB/self-interaction constraints and present substructure enhancements, we consider two limiting cases: a case with only new "irrelevant" annihilations, and a case with only new "relevant" annihilations, in the sense of being (ir)relevant to indirect detection.

The first possibility is that there may be extra annihilation channels that are important for freezeout, but irrelevant to the CR excesses today. These processes might include p-wave suppressed annihilation, channels that experience a repulsive Sommerfeld effect, involvement of excited states that are present in the early universe but have decayed by the present day, very soft annihilation channels (which contribute to e^+e^- signals where the backgrounds are large), or annihilations into invisible channels, such as neutrinos or dark-neutralinos. In this case (new "irrelevant" channels), we assume that the annihilation rate relevant to indirect detection in the Galactic halo is calculable from the parameters α_D, m_ϕ, and m_χ alone, and take it to be given by the minimal t-channel annihilation cross section into dark gauge bosons, $\langle\sigma v\rangle = \pi\alpha_D^2/m_\chi^2\times$ the Sommerfeld enhancement. We further assume that the Sommerfeld enhancement is controlled by the same force carriers into which the dark matter annihilates.

Where the calculated annihilation rate is smaller than required to generate the correct relic density, we assume the difference is made up by these extra annihilation channels which – for whatever reason – do not contribute signal to present-day indirect detection experiments. Where this rate is too large to generate the correct relic density, we say that this point is ruled out by the relic density constraint. In other words, we take the relic density to provide an upper bound on the early universe cross section, rather than fixing its value.

Alternatively, there could be additional annihilation channels that contribute to both freezeout and the local CR signal, but which do not correspond to force carriers mediating additional Sommerfeld enhancement. This could arise for instance by annihilations into an additional force carrier that is not as effective for Sommerfeld enhancement (e.g. due to a more massive force carrier), but has a larger coupling. In this class of scenarios (new "relevant" channels), we again take α_D, m_ϕ, and m_χ, and assume that these parameters determine the Sommerfeld enhancement, but that the underlying cross section relevant for both freezeout and indirect detection experiments $\langle\sigma v\rangle$, which is Sommerfeld-enhanced at low velocities, can be larger than the naive $\pi\alpha_D^2/m_\chi^2$. When even the "bare" rate of $\langle\sigma v\rangle = \pi\alpha_D^2/m_\chi^2$ alone is larger than allowed by the relic density, we again say that this point is ruled out by the relic density constraint (again, taking the relic density as an upper bound on the cross section).

In both cases, we must take into account the fact that the presence of Sommerfeld enhancement reduces the "bare" annihilation cross section that gives the correct relic density [9, 69, 76, 77]. We follow [68] and use a Taylor expansion of the Sommerfeld enhancement during freezeout to estimate the magnitude of this effect.

C. Consequences of substructure for CMB and self-interaction constraints

Figure 2 shows how constraints from the CMB and self-interaction limits bound the force carrier mass as a function of the substructure contribution, when the parameters are tuned to achieve a desired boost factor locally, for the Sommerfeld enhancement induced by a Yukawa potential. More complicated models for the enhancement are certainly possible – one example is discussed briefly in Appendix A – but this simple case adequately demonstrates the effect of non-zero Δ. We use a DM mass of 1.2 TeV and a local boost factor of 100 as a benchmark, and assume the dark force carrier decays only into electrons: this benchmark provides an adequate fit to the PAMELA and *Fermi* measurements. For each point in $\Delta - m_\phi$ parameter space, we solve for the largest value of the dark sector coupling α_D that gives rise to the desired local boost factor: smaller values of α_D may achieve the same local boost factor, but only if they lie near the tips of narrow resonance peaks (a requirement which becomes increasingly finely tuned as α_D decreases, since the enhancement must be increasingly close to perfectly resonant to cancel out the usual reduction in the enhancement from lowered α_D). Consequently, choosing the largest possible α_D yields the most "natural" parameter set giving the desired boost factor, in the sense that small changes to the parameters will not change the boost factor very much. The resulting self-interaction cross section and signal in the CMB can then be compared to the limits. There is a third constraint from relic density, as described above for the cases of extra relevant and

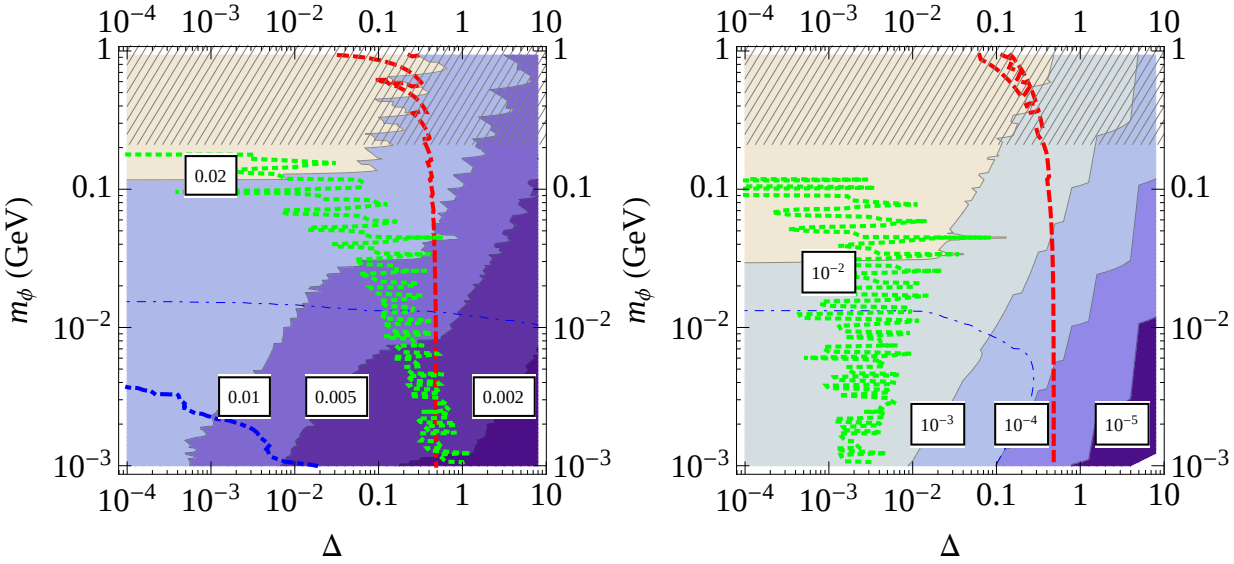

FIG. 2: Contours of constant dark sector coupling α_D as a function of mediator mass m_ϕ and substructure contribution Δ, for a fixed dark matter mass of 1.2 TeV and local boost factor (BF) of 100, in the scenarios with new "irrelevant channels" (*left panel*) and "relevant channels" (*right panel*). The BF includes contributions from Sommerfeld enhancement and substructure and is defined by BF $= \frac{\langle\sigma v\rangle_{v\sim 150\mathrm{km/s}}}{3\times10^{-26}\mathrm{cm}^3/\mathrm{s}}\left(1+\frac{S(v\to 0)}{S(v\sim 150\mathrm{km/s})}\Delta(r\sim 8.5\mathrm{kpc})\right)$. Regions to the left of and/or below the red dashed (blue dot-dashed) lines are ruled out by constraints from the CMB (self-interaction bounds). The thin blue dot-dashed line denotes the threshold at which self-interaction effects in dwarf galaxies become significant, but may be allowed. The region to the left and below the green dotted line is where WIMPonium formation becomes relevant; at this DM mass and for this class of models, WIMPonium formation is almost always ruled out by the CMB constraints, although it can become relevant at higher DM masses. The dark gauge boson is assumed to decay into electrons only, in which case this boost factor and DM mass provide a good fit to the PAMELA and *Fermi* data. When the gauge boson mass exceeds twice the muon mass, the true final state may become more complicated, so this region is indicated by cross-hatching. All points on this plot have been checked and do not over-deplete the thermal relic density.

irrelevant channels, but for a required boost factor of 100 at $m_\chi = 1.2$ TeV, the constraint curve does not appear on these plots: even for zero local substructure and zero mediator mass, a boost factor of 100 is attainable via the Coulomb-like $\pi\alpha_D/v$ Sommerfeld enhancement while maintaining consistency with the relic density bound (as can be seen from e.g. [69]).

The major difference between the two scenarios we consider is the rate at which the preferred value of α_D falls with increasing Δ, and hence the strength of the self-interaction bound at low masses. In the first case, with extra irrelevant annihilation channels, the saturated annihilation rate scales as α_D^3, whereas in the case with additional relevant channels, the bare annihilation rate is largely fixed by the relic density alone, and so the saturated annihilation rate scales roughly as α_D. Consequently, in the "irrelevant channels" case relatively small changes to α_D are sufficient to greatly reduce the signal, compensating for the increased boost factor from saturated enhancement in subhalos; α_D changes by only a single order of magnitude over the parameter space we consider. If the self-interaction threshold discussed earlier is treated as a limit, it remains quite stringent at mediator masses below $m_\phi \sim 10$ MeV. In the "relevant channels" scenario, on the other hand, a large reduction in the saturated annihilation rate requires a large reduction in α_D: the very low values of α_D at low mediator mass and $\mathcal{O}(1)$ Δ also greatly relax any self-interaction bounds.

If the force carrier mass is sufficiently light, $m_\phi < \alpha_D^2 m_\chi/4$, then it is possible for two DM particles to radiatively capture into a bound state at low velocities, referred to as WIMPonium. The capture cross section scales in the same way as Sommerfeld-enhanced annihilation in the low-velocity limit, but is larger by a factor of ~ 6 in the limit where $m_\phi \ll \alpha_D^2 m_\chi/4$. For this DM mass, WIMPonium formation primarily affects regions of parameter space that are already ruled out by the CMB bounds, but it can be marginally relevant for $\Delta \sim 0.5-1$ and few-MeV force carriers, and is included in the plots.

Figure 2 is useful for showing how the different constraints compare, but relies on picking a specific target boost factor. A question of perhaps more general interest is how the maximal boost factor *consistent with all constraints* varies as a function of Δ and m_ϕ. We again proceed by sampling the $\Delta - m_\phi$ parameter space holding m_χ fixed at 1.2 TeV, and at each point scan over α_D to obtain the maximum boost factor consistent with the CMB, self-interaction

and relic density bounds. We include radiative capture to WIMPonium in the boost factor, for values of α_D where it is kinematically allowed.

As mentioned previously, at resonance peaks relatively low values of α_D can give rise to very large saturated enhancements. Relying on such resonance peaks is problematic for several reasons:

- The usual treatment of the Sommerfeld enhancement neglects higher-order corrections that regulate the resonances; the saturated enhancement in our current approximate treatment diverges at the exact centers of the resonances, and this is not physical.
- On the resonances, the enhancement saturates at lower velocities than in the non-resonant case; close to the centers of the resonances, it is not clear that we can assume the enhancement is saturated in the smallest subhalos.
- Our perturbative treatment of Sommerfeld corrections to thermal freezeout fails in the case of highly resonant enhancement, since in this case annihilations can recouple after kinetic decoupling.
- From an aesthetic perspective, demanding that the Sommerfeld enhancement be highly resonant implies fine-tuning of the parameters.

Consequently, we impose the further condition that the saturated Sommerfeld enhancement must not exceed the expected non-resonant value, $12\alpha_D m_\chi/m_\phi$, by more than a factor of 10. Increasing this factor to 100 has a negligible effect on the results.

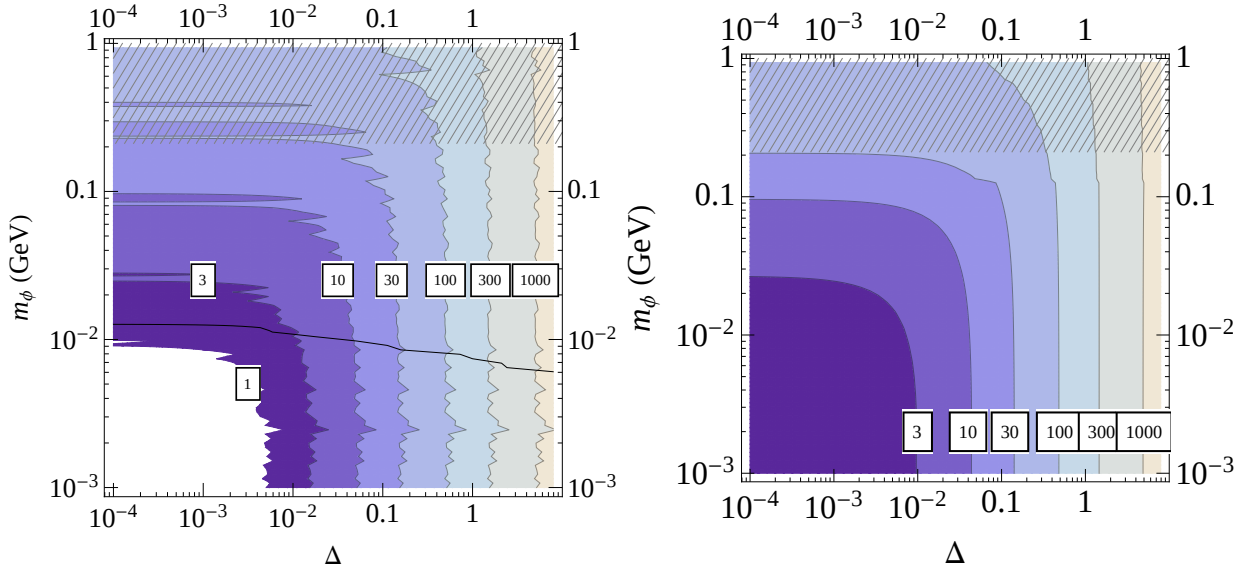

FIG. 3: The maximum local boost factor for 1.2 TeV dark matter consistent with constraints from the thermal relic density, the CMB, self-interaction bounds, and naturalness (in the sense of not relying on the resonance peaks), in the scenarios with new "irrelevant channels" (*left panel*) and "relevant channels" (*right panel*). The dark gauge boson is assumed to decay into electrons only; when the gauge boson mass exceeds twice the muon mass, the true final state may become more complicated, so this region is indicated by cross-hatching. In the left panel, parameter points that maximize the boost for $m_\phi < 13$ MeV have transfer cross-sections at 10 km/ s above the "self-interaction threshold" of 0.1 cm^2/ g. Treating this threshold as a hard constraint would extend the white region ($\mathrm{BF}_{\mathrm{local,max}} \leq 1$) out to the black curve, while having no effect at all on the contours for $m_\phi > 13$ MeV. In contrast, the self-interaction bounds are never constraining for scenarios with new relevant channels.

The results of this analysis for scenarios with either "irrelevant" or "relevant" annihilation channels are shown in Figure 3. The limiting constraint in most of the parameter space is that from the CMB, with dwarf evaporation becoming a significant constraint only at low Δ and m_ϕ, in the case of "irrelevant channels" (which require larger α_D to produce a given boost). The more stringent "self-interaction threshold" of $\sigma_T = 0.1$ cm^2/ g at $v_0 = 10$ km/s, above which DM scattering within the dwarf halo would change its shape significantly from non-interacting CDM simulations, is not included as a constraint, but the parameter points shown in the left plot (irrelevant channels) cross this threshold at $m_\phi = 13$ MeV. If we were to treat this threshold as a constraint, it would leave the region $m_\phi > 13$ MeV completely unaffected, but sharply change the maximum boosts in the region $m_\phi < 13$ MeV, with the

white region $BF_{\rm local} \leq 1$ pushed out to the near-horizontal black line. No self-interaction effects are important in the "relevant channels" scenario because of the much slower scaling of the annihilation rate with α_D.

In Figure 3 we see that a maximal boost factor of ~ 100 is first achieved at $\Delta \sim 0.4 - 0.5$ for small mediator masses; in this region of parameter space, the CMB provides the strongest constraint. This is consistent with Figure 2, where we see that the target boost factor of 100 is first consistent with the CMB limits at $\Delta \sim 0.4 - 0.5$. We have checked the effect of running the analysis with and without including WIMPonium formation and found that it makes essentially no difference to our results (although the plots we show do include it): the parameters for which WIMPonium formation is kinematically allowed are in general excluded by the constraints, or at most marginally allowed, and so contribute little to the maximum boost factor.

In the $\Delta = 0$ limit, the maximum boost factor is strongly dependent on m_ϕ, but for Δ of $\mathcal{O}(1)$ this dependence is almost completely removed. This is to be expected: in this regime the CMB provides the strongest limits, the local enhancement is substructure-dominated, and the ratio of the local enhancement to the saturated enhancement depends only on Δ. (We remind the reader, however, that models saturating the CMB limit with very low $m_\phi \lesssim 10$ MeV have transfer cross-sections of order $0.1-1$ cm^2/ g at 10 km/s, whose effect on the substructure of dwarf galaxies is significant.) Figure 4 shows the increase in maximum boost factor (consistent with all constraints) in the presence of substructure, as a function of Δ and m_ϕ: even for Δ of $\mathcal{O}(1)$, the factor can be two or more orders of magnitude.

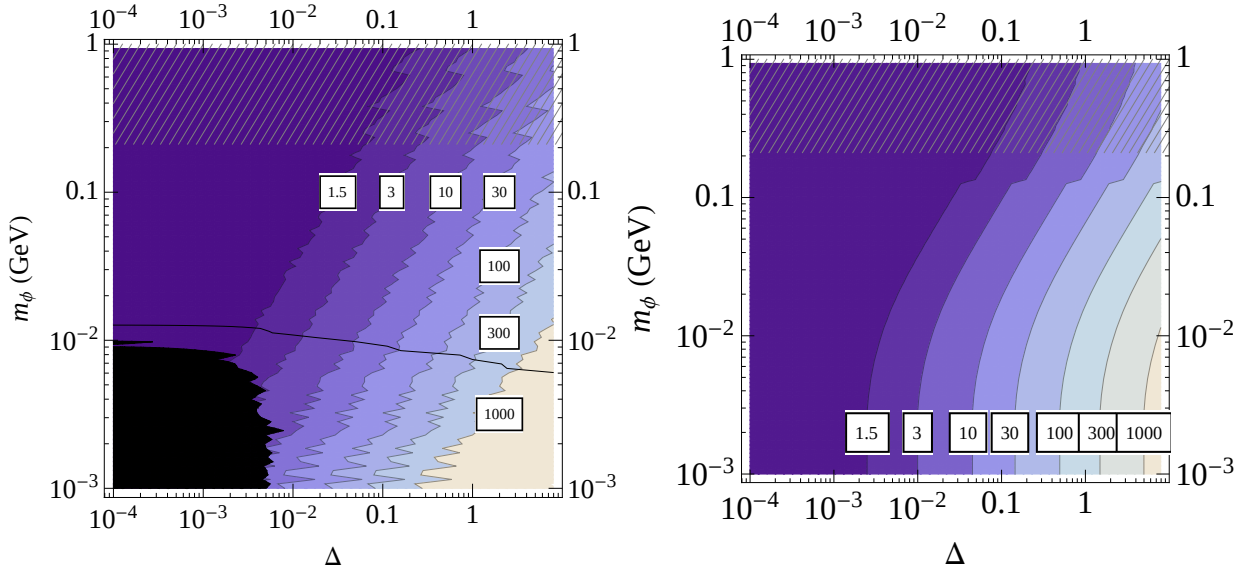

FIG. 4: The ratio of the maximum local boost factor for 1.2 TeV dark matter to the maximum local boost factor with $\Delta = 0$, where in both cases the parameters achieving the maximum local boost are required to respect constraints from the thermal relic density, the CMB, self-interaction bounds, and naturalness (in the sense of not relying on the resonance peaks), in the "irrelevant channels" (*left panel*) and "relevant channels" (*right panel*) scenarios. The dark gauge boson is assumed to decay into electrons only; when the gauge boson mass exceeds twice the muon mass, the true final state may become more complicated, so this region is indicated by cross-hatching. In the "irrelevant channels" scenario, the region where the substructure-enhanced local boost is less than 1 is blacked out, since while this region may have a larger boost factor with substructure than without, the boost factor is still very small and so its details are not very interesting. If the "self-interaction threshold" from self-interaction of DM within dwarf galaxies were treated as a hard constraint, this black region would extend out to the black curve.

IV. THE DISTRIBUTION OF SUBSTRUCTURE AND ITS IMPLICATIONS

The amount of local substructure is highly uncertain. N-body simulations cannot resolve the small subhalos that are expected to contribute the bulk of the signal, so some extrapolation procedure must be employed (over 12 or more orders of magnitude). The mass of the smallest subhalos scales as the cube of the temperature of kinetic decoupling of the DM from the SM, which can easily range from $1 - 100$ MeV. Furthermore, DM self-interactions and baryonic physics, neither of which are included in N-body simulations, may deplete substructure and/or flatten the density profiles of subhalos at later times. Consequently, it is important to explore the consequences of a broad range of substructure scenarios. In particular, we consider directly the formulations of [16, 17, 78], all of which attempt to understand the implications of N-body simulations on the expected local boost.

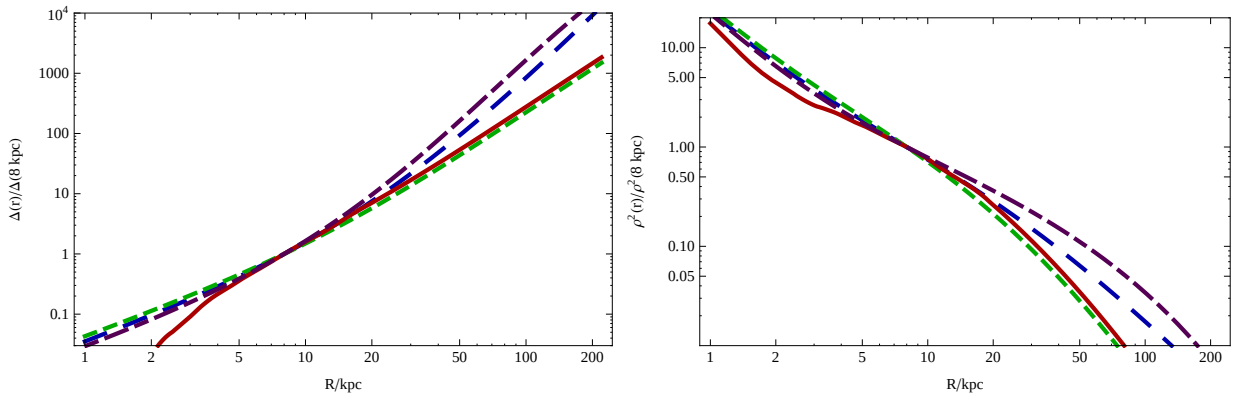

FIG. 5: *left*: the profile $\Delta(r)/\Delta(8.5\text{kpc})$ for four different approaches to extrapolating N-body results into the inner halo. *right*: $\rho^2(r)/\rho^2(8.5\text{kpc})$ (relevant for line-of-sight calculations) with $\Delta(8.5\text{kpc}) \times \sigma v_{sat}/\sigma v_{smooth} = 10$ for the same extrapolations. Lines are: the approach of [17] to the Via Lactea II simulation (*blue, long-dashed*), the approach of [16] to the Aquarius simulations (*purple, long-short dashed*), the approach of [78] to the VLII simulation, with tidal disruption (*red, solid*), and without tidal disruption (*green, dashed*).

A. Parameterization of unresolved substructure

There are many different approaches to the question of substructure boosts, often relating to whether the study focuses on the local boost, the boost to emission from a distant source (such as a dwarf galaxy), or the boost to the extragalactic, all-halo and all-redshift, isotropic signal. However, in general, there are three parameters which influence the relevant substructure boost: $1 - f_{sm}$, the fraction of the DM in bound substructures (f_{sm} is the fraction in the "smooth" halo), α_{sub}, the exponent of the power law that controls the distribution of subhalos of different masses[92], and finally α_h, which is the parameter which determines the distribution of unresolved *main* halos, relevant for studies of isotropic diffuse photons.

The local boost is most dependent on a) the total amount of substructure as well as the mass distribution of subhalos controlled by α_{sub}, and b) the amount of substructure present in the solar neighborhood $1 - f_{sm}$. The boost in the outer part of the MW is controlled by essentially the same two parameters. Thus, the ratio between the two is dominantly controlled by the evolution of $1 - f_{sm}$ as a function of radius (in addition to the change in concentration as a function of radius from tidal stripping [78]).

B. Inner Galaxy gamma-ray flux

While various approaches differ on the rate at which the substructure signal is disrupted towards the GC, there is general agreement that it *should* be suppressed as one moves to smaller radii. Thus, in the limit that the local signal is dominated by substructure, one can estimate the maximum amount by which any inner galaxy signals would be suppressed (compared to the no substructure case) by looking at the evolution of $\Delta(r)$, which we show in figure 5. Even in the case with the slowest evolution of substructure of the four cases we consider (from [78] without tidal disruption), the signal is suppressed in the inner 1 kpc by a factor of $\sim$20. Thus, in the limit that local substructure dominates the PAMELA signal *even the most stringent ICS or FSR constraints will be irrelevant.*

C. Outer halo gamma-ray flux

At the same time, one might be concerned that by boosting the local substructure signal, one is also boosting *other* signals that are already dominated by substructure, namely, gamma rays from the outer galaxy and unresolved extragalactic sources (the latter contributing to the isotropic diffuse gamma-ray flux). The limits from gamma-ray emission from dwarf galaxies, briefly mentioned previously, will also become more stringent if substructure is taken into account.

For the outer galaxy, there are already studies [16, 79] that consider the role of substructure in ICS signals from the outer halo. [16] employs the Aquarius simulation, normalizing the local signal assuming that the entire local PAMELA signal arises from substructure (i.e., $S_{v\to 0}\Delta \gg S_{v\sim 150\text{km/s}}$), while [79] uses the approach of [17] to the VLII

simulation, with a local value $\Delta \sim 0.5$, and implicitly take $S_{v\to 0} = S_{v\sim 150\mathrm{km/s}}$. Both papers calculate the ICS signals from the outer halo and use this to set constraints.

If we assume that the local signal is substructure dominated, then the outer halo is substructure dominated as well. In this limit, only the product $\Delta(r)S(v \to 0)$ is relevant. As an example, [16] assumes substructure domination, thus their limits are directly applicable here. On the other hand, since [79] takes $\Delta(8.5\text{ kpc}) \sim 0.5$, their limits should be strengthened by a factor of ~ 2 in the substructure-dominated case. Importantly, both analyses assume that the e^+e^- produce their ICS signals at the point of annihilation. Since the energy loss time for these particles is $\mathcal{O}$(Myr), the particles would at least partially diffuse away, and suppress these limits by up to $\mathcal{O}$(few).

Even without accounting for this correction, [16] finds these signals are only borderline, not excluded. Moreover, observing figure 5, we see that, of the three approaches we have considered, in this approach the substructure is depleted most rapidly as one moves in from the outer halo. Other formulations of substructure, with weaker dependence on Galactocentric radius, will yield weaker constraints from the outer MW halo, relative to a fixed local signal. In light of this uncertainty and the effects of e^+e^- diffusion, we conclude the the ICS signals of the outer MW halo do not strongly constrain the substructure dominated scenario.

D. Gamma-rays from dwarfs

The non-observation of a gamma ray excess from the dwarf galaxy Segue 1 constrains the DM annihilation cross section to be no more than ~ 100 times larger than required to fit the CR excesses with $4e$ annihilation (neglecting substructure both locally and in the dwarf) [68]. The presence of substructure in both systems would enhance the annihilation signal from Segue 1 more than the local CR signal, because the full substructure boost to Segue 1 is dominated by its substructure-rich outer halo. However, this ratio is not likely to provide the two-order-of-magnitude relative enhancement to the Segue 1 signal that would be needed to derive a strong constraint from the current measurement. We may expect the overall boost factor for Segue 1 to be at most comparable to that of a Milky Way-like galaxy, and plausibly smaller because the dwarf possesses substructure over fewer decades of mass. As discussed below, we find that the total boost of the Milky Way seen from far away exceeds the local boost Δ by a factor of $\sim 5 - 30$, following the approaches of [78] and [17]. A comparable boost to the Segue 1 signal would not be sufficient to derive a strong constraint.

E. Diffuse extragalactic gamma-rays

A more stringent constraint involving substructure is the limit from isotropic diffuse gamma rays [47, 80, 81]. The distribution of main halos, and the distribution of substructure (by mass), are critical for this constraint, but the limit is essentially divorced from the question of how the substructure evolves with Galactocentric radius in the inner galaxy, which is key for local signals. In a recent analysis, [81] employed the approach previously developed by [82], where two independent quantities are modeled as power laws with parameters fitted from N-body simulations: the boost from main halos below the resolution limit of the simulation, and the boost associated with each main halo due to its substructure, as a function of the main halo mass.

For the extragalactic diffuse gamma-ray signal, we take as our comparison point the integrated flux from all smooth main halos with masses between $6.89 \times 10^8 h^{-1} M_\odot$ and $10^{15} M_\odot$. The lower limit corresponds to the resolution of the Millennium-II simulation, and the power-law behavior described in [82] may no longer be accurate for halo masses $\gtrsim 10^{15} M_\odot$. All boosts for the extragalactic diffuse signal are "scaled isotropic boosts," defined with respect to this quantity.

The scaled boost factor to the extragalactic diffuse signal from the combined substructure and unresolved main halo contributions then ranges from $\sim 20 - 2500$ when varying the parameters $A_{\rm sub}$ and $\alpha_{\rm sub}$, which respectively determine the normalization and slope of the power law governing the substructure mass function, across the ranges $10^{-0.5} \leq A_{\rm sub} \leq 10^{0.1}$, $-1.15 < \alpha_{\rm sub} < -0.95$. This range of scaled isotropic boosts assumes the parameters of the power law governing the unresolved main halos are held fixed at their best-fit values, with a fixed minimum subhalo mass of $10^{-6}\ M_\odot$.

In the case with the *smallest* amount of unresolved substructure within the included parameter space, corresponding to a scaled boost to the isotropic signal of ~ 20 by our definition, [81] have shown that the maximum *saturated* annihilation cross section is characteristically very close that which is required to explain PAMELA in the absence of substructure, even before any subtraction of astrophysical backgrounds (the uncertainties on the backgrounds are sufficiently large that their contribution to the signal could very well be subdominant). In other words, in substructure-dominated scenarios we would simultaneously require the local boost $\Delta(8.5\text{kpc}) \gtrsim 1$ when the total scaled boost to the extragalactic emission is ~ 20. More generally, any scenario where $\Delta(8.5\text{kpc})/(\text{scaled isotropic boost}) \gtrsim 1/20$

could simultaneously fit PAMELA and evade this constraint. A question we wish to confront is: is such a ratio of substructure boosts reasonable?

It is difficult to address this question directly within the framework of [81], because the simulations used there cannot resolve structures below $\sim 10^8 M_\odot$, and do not address local boosts deep inside the Galactic halo. On the other hand, the other methods in the literature are not readily expressed in the parameterization of [82]. However, we note that for the models we have tested, the total integrated boost from a distant halo is essentially independent of behavior in the inner Galaxy (being dominated entirely by the outer halo); furthermore, if we follow [78] and use the Roche criterion to parameterize tidal disruption of substructure, the resulting reduction in the inner-Galaxy boost factor is nearly independent of the parameters $\alpha_{\rm sub}$, $A_{\rm sub}$. Thus, it makes sense to treat the radial profile of substructure in the inner Galaxy and the overall normalization of the boost factor as unrelated, and simply use the results from [17, 78] to obtain the ratio of the local ($R \sim 8.5$ kpc) boost to the overall boost of the halo. Thus, to determine the ratio of the local boost to the isotropic boost and whether it is larger than 1/20, we take

$$\frac{\Delta(8.5\text{ kpc})}{\text{BF}_{\rm isotropic}} \simeq \left(\frac{\Delta(8.5\text{ kpc})}{\text{BF}_{\rm MW}}\right)_{\rm out\to in} \left(\frac{\text{BF}_{\rm MW}}{\text{BF}_{\rm isotropic}}\right)_{\rm ZSB}. \tag{7}$$

Here the first factor is calculated using the approaches [78] or [17] to connect the local inner boost to the "total boost" of the MW as seen from far away (dominated by outer halo structure), while the second factor is calculated using the approach of [82] to connect the MW total boost to the scaled isotropic boost.

Using the approach of [78] for the first factor, we find that the total boost of the MW, integrating out to 200 kpc, is approximately 6× the local value of Δ, while for [17] the corresponding factor is 18. Thus the extragalactic gamma ray bounds can be evaded if the scaled diffuse gamma-ray boost, integrated over all halos (up to $10^{15} M_\odot$) and all substructure, is $\lesssim 1-3\times$ greater than the total boost of the Milky Way.

Working in the formalism of [82], we can now determine that this condition naturally holds for $\alpha_{\rm sub} \lesssim -1.00$ for the largest values of $A_{\rm sub}$ and $\alpha_{\rm sub} \lesssim -1.05$ for the smallest $A_{\rm sub}$. Note that this is *not* the low-substructure limit: as an example, taking the central values $\alpha_{\rm sub} = -1.05$, $A_{\rm sub} = 10^{-0.2}$, the boost factor for a 10^{12} solar mass halo is ~ 35 (corresponding to a local $\Delta \sim 6$ under the formalism of [78]), and for the extragalactic diffuse signal is ~ 90. Thus, for these cases, we estimate $\Delta(8.5\text{ kpc})/\text{BF}_{\rm isotropic} \sim 1/15$ and 1/50.

So far, we have assumed the parameters of the unresolved *main* halos and the cutoff mass are perfectly known, but because the signal is generically dominated by the small halos and subhalos, even a small uncertainty in these parameters can have a substantial impact on the boost. Changing the power law index α_h from -1.05 to -1.0, in the previous example, leaves the Milky Way boost factor unaffected, but reduces the integrated boost factor for the diffuse gamma-rays to ~ 30, in which case $\Delta(8.5\text{ kpc})/\text{BF}_{\rm isotropic} \sim 1/5$ and 1/15.

Raising the cutoff mass above $10^{-6} M_\odot$ also improves the consistency with the constraints (at least for the default value of $\alpha_h = -1.05$), simply because most of the extragalactic signal comes from small, dense subhalos which are almost entirely destroyed if the cutoff is sufficiently raised. As an example, again taking $\alpha_{\rm sub} = -1.05$, $A_{\rm sub} = 10^{-0.2}$ but raising the cutoff mass to $1 M_\odot$ reduces the boosts for the Milky Way and the diffuse emission to ~ 15 and ~ 25 respectively, yielding $\Delta(8.5\text{ kpc})/\text{BF}_{\rm isotropic} \sim 1/10$ and 1/30. Raising the cutoff mass thus opens additional allowed parameter space at larger values of $\alpha_{\rm sub}$.

In summary, the diffuse extragalactic gamma-ray background is probably (together with the CMB) one of the most sensitive probes of the substructure-dominated scenario for the PAMELA excess. The expected isotropic signals in such a case are typically of the same order as the current limits, but they are extremely sensitive to even small changes in the parametrization of substructure. As such, the substructure-dominated scenario for the PAMELA excess does not appear to be in clear conflict with [81], although it may still require a relatively small contribution from star forming galaxies and blazars to the gamma ray background. It is interesting to note that the naively least-constrained scenario, where the substructure is minimized, may not actually be least constrained as an explanation for PAMELA when local substructure is self-consistently taken into account.

F. Depletion of substructure in light mediator models

The usually assumed low-mass cutoff of $\sim 10^{-6} M_\odot$ presumes a kinetic decoupling temperature of $\mathcal{O}(100)$ MeV, which is appropriate for a standard WIMP. For models with a light mediator kinetically mixed with the photon, however, the cross section for scattering of dark matter on charged Standard Model particles can be much larger. Direct detection experiments constrain the mixing if the scattering is elastic, but small ($\mathcal{O}(100)$ keV) mass splittings δ between the states in the dark matter multiplet can remove these limits.

For plausible parameters, the DM may remain coupled to the SM via DM-electron scattering at temperatures down

to m_e: specifically, the kinetic decoupling temperature is [69],

$$T^e_{\rm kd} \sim \max\left\{ m_e, \delta, 0.82{\rm MeV}\left[\frac{10^{-3}}{\epsilon}\right]^{1/2} \times \left[\frac{m_\phi}{30{\rm MeV}}\right]\left[\frac{0.021}{\alpha_D}\right]^{1/4}\left[\frac{m_\chi}{\rm TeV}\right]^{1/4}\right\}. \quad (8)$$

For these relatively low kinetic decoupling temperatures, the mass cutoff scale is given by [83],

$$M_{\rm cutoff} = 3.4\times 10^{-6}\left(\frac{T_{\rm kd} g_{\rm eff}^{1/4}}{50{\rm MeV}}\right)^{-3}. \quad (9)$$

For $T_{\rm kd} \sim m_e$, we find $M_{\rm cutoff} \sim 1 M_\odot$. It is certainly not *necessary* that the cutoff mass be this small, since small ϵ and α_D would lower the cutoff scale, but it is plausible for scenarios with very small mediator mass.

After structure formation, subhalos may also be evaporated in the presence of self-interaction, by collisions with more energetic particles in the host halos. This mechanism has already been invoked to set constraints on the self-interaction by demanding that dwarf galaxies within the Milky Way and galaxies within clusters have not yet evaporated, following [71], but what of subhalos in smaller, denser hosts? Would they evaporate early, for models saturating the self-interaction limits we have imposed?

Note that the properties of the *subhalo* do not matter for this question, since the particles of the host halo are definitionally not bound to substructure and have enough energy to remove particles from any subhalo. Only the characteristic velocity and density of the host halo are relevant. Consequently, this effect cannot strongly affect local Δ (since the Milky Way has dwarf galaxy subhalos), and we need only ask if it can be relevant in smaller host halos, thus affecting the diffuse gamma-ray limits.

The timescale for evaporation scales as $(n\sigma(v)v)^{-1}$, where v and n are the characteristic velocity and number density of the host halo. If we use the parameterization of [17] and take $v \propto n^{-1.75}$, the evaporation timescale varies as $(v^{0.43}\sigma(v))^{-1}$. For small v, σ scales as $v^{-0.7}$ before leveling off to log dependence on v (Equation 1). Thus, while the dependence on v is always quite weak, it seems possible in principle for the saturation of the self-interaction to pick out a particular range of halo masses in which evaporation is faster than the age of the universe, with halos both above and below this characteristic mass range not evaporating.

In general, for couplings smaller than the self-interaction threshold (where self-scattering is too weak to change halo properties), evaporation is never fast enough to efficiently destroy subhalos, at least using this simple estimate. For lighter force carriers, for which the dark matter departs from the collisionless limit, evaporation can naturally occur over some range of host halo masses. This evaporation could further weaken diffuse isotropic gamma-ray limits, which under the usual assumptions receive large contributions from the substructure of low-mass halos. Moreover, even in the nominally "collisionless" region of parameter space, the gap between the timescale for evaporation and the age of the halo can be less than an order of magnitude, and a more careful analysis is justified.

V. CONCLUSIONS

The e^+ excess observed by PAMELA and now *Fermi* points to a new primary source of cosmic ray electrons and positrons. One of the most exciting, if speculative, explanations of the excess is that it arises from Sommerfeld-enhanced dark matter annihilation. Such models naturally provide annihilation rates much larger than what would be expected from a thermal WIMP with substructure enhancement alone.

Nonetheless, in models with Sommerfeld enhancement, in the presence of $\mathcal{O}(1)$ substructure, the substructure is often the *dominant* source of the signal, because of the low velocity dispersion of the bound subhalos. Most constraints on these models have been calculated assuming the signal arises from the smooth halo, and the limits become dramatically weaker if substructure dominates the CR signal.

In particular, the constraints on parameter space from the CMB are removed for $\Delta \gtrsim 0.4$, since the local signal as well as the early universe signal are both determined by the saturated cross section, in contrast to the smooth halo piece that is generally unsaturated. Because substructure is depleted in the inner regions of galaxies, constraints from FSR and ICS signals in the inner Milky Way are strongly suppressed. As Δ increases, lower couplings between the DM and the force carrier are required to fit the CR excesses, and this in turn relaxes limits on the force carrier mass from bounds on DM self-interaction; such constraints are subsumed by the CMB bounds, but depending on the model, the more stringent requirement that self-interaction have negligible effect on dwarf galaxy structure may imply $m_\phi \gtrsim 10$ MeV. This has profound implications for terrestrial searches. Considering substructure not only opens the lighter ($m_\phi \lesssim 200$ MeV) ranges of parameter space, but possibly makes them the preferred range, once additional constraints are considered. Given the sensitivity that many experiments have in this region, these searches become even more motivated.

Limits on the DM annihilation rate from measurements of the diffuse extragalactic gamma-ray background, or from gamma-rays from the outer halo of the Milky Way, generally become stronger as the amount of substructure is increased. However, the key quantity is the ratio of the signal in these constraining channels to the local substructure-enhanced annihilation rate, and this can easily be substantially *smaller* at higher Δ than for $\Delta = 0$, relaxing the limits on DM explanations for the CR excesses.

There is no established consensus on what the local boost from substructure should be, due to the large uncertainties in extrapolating the results from N-body simulations below their mass resolutions, but it is not thought to be extremely large. Nonetheless, in the presence of Sommerfeld enhancement the substructure contribution can still easily dominate and open up new regions of parameter space, especially for sub-200 MeV force carriers. These regions, too, will be tested by further observations, with the exciting possibility that we might already be detecting the cosmic ray signals of dark matter substructure.

Acknowledgments

We thank Nima Arkani-Hamed, Rouven Essig, Mike Kuhlen, Julien Lavalle, Annika Peter, Jennifer Siegal-Gaskins, and Jesús Zavala for comments and conversations. NW is supported by NSF grant #0947827, as well as support from the Ambrose Monell Foundation. TS is supported by NSF grant AST-0807444 and DOE grant DE-FG02-90ER40542. Research at the Perimeter Institute is supported in part by the Government of Canada through NSERC and by the Province of Ontario through MEDT. NW and TS acknowledge the hospitality of the Aspen Center for Theoretical Physics during the early stages of this work. This research was supported in part by the National Science Foundation under Grant No. NSF PHY05-51164.

Appendix A: An inelastic example

The observant reader may have noticed that in Figure 2, the zero-Δ case appears to be ruled out for masses up to 1 GeV by the CMB bounds, and in Figure 3, the maximum local boost for $\Delta = 0$ is only ~ 30 at $m_\phi = 200$ MeV. At first glance, this seems to contradict the claim that the PAMELA and *Fermi* results can be explained by Sommerfeld-enhanced models in the absence of substructure, for sub-GeV mediators.

However, this is just the factor of $\sim 2-3$ discrepancy noted in [9], for the case where the states in the dark-charged DM multiplet were taken to be degenerate. As briefly mentioned in §II C, the benchmark models presented in [9] possessed two non-degenerate dark matter states, and a larger smooth-halo Sommerfeld enhancement as a result. The same analysis described above can be applied to a model of this type with a fixed mass splitting, with the modification that now we also need to take into account annihilation channels involving the excited states, at least for the freezeout calculation.

We again consider two scenarios: in our new scenario (1), for our "baseline" annihilation cross section we include all the annihilation channels described in [9] (which are all "irrelevant" by the definition given in §III B); a point in parameter space is ruled out by the relic density constraint if this baseline cross section over-depletes the relic density. We assume that additional "irrelevant" (to indirect detection) annihilation channels exist to deplete the relic density: this is easily achieved in the models of [9] by e.g. introducing new states charged under the dark-sector gauge interaction, since s-channel annihilation through an off-shell dark gauge boson provides a coannihilation channel which is large at freezeout but suppressed by the abundance of the excited state in the present day.

In the inelastic scenario (2), rather than use the specific models of [9], we assume a single universal s-wave annihilation cross section for co-annihilation and self-annihilation of dark matter particles in the ground or excited states, choose this cross section to obtain the correct relic density, and assume this annihilation rate is multiplied by the Sommerfeld enhancement. A point in parameter space is "ruled out by the relic density constraint" if the required cross section is smaller than $\sigma v = \pi\alpha_D^2/m_\chi^2$. This approach removes differences between the inelastic and elastic cases due to differences in the self-annihilation vs co-annihilation rates and model-dependent extra annihilation channels, allowing us to clearly see the effect of the modified Sommerfeld enhancement in inelastic models.

We employ the approximation for the inelastic Sommerfeld enhancement derived by [64]. This approximation is expected to break down for mass splittings $\delta \gtrsim \alpha_D m_\phi$: consequently, we cannot study the small-mediator-mass region in detail in the inelastic case, especially in scenario (2) where at small mediator masses and/or large Δ very small values of α_D are favored.

Our results are shown in Figures 6-7: in Figure 6, we take a target boost factor of 65, a mass splitting of 700 keV, and a DM mass of 1 TeV, motivated by the lowest-mediator-mass benchmark model in [9]. The qualitative features are similar to the elastic case, although we cannot study the self-interaction bounds since they only apply to small

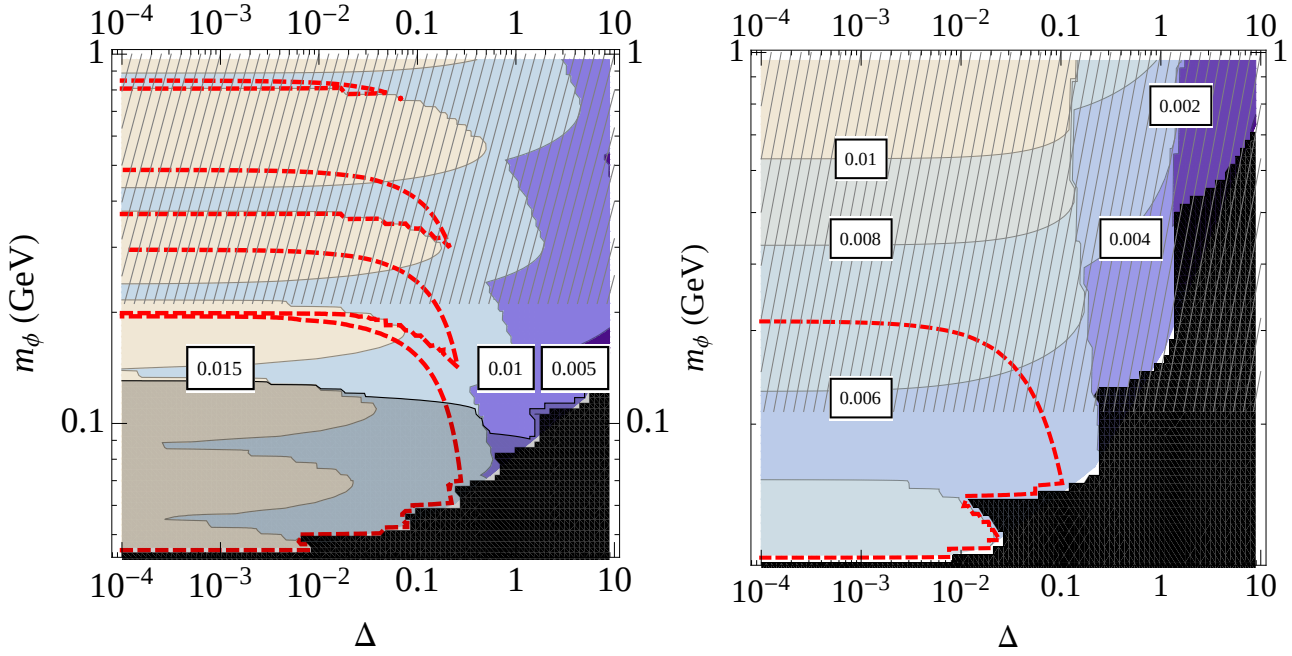

FIG. 6: The dark sector coupling α_D as a function of mediator mass m_ϕ and substructure contribution Δ, for a fixed dark matter mass of 1 TeV, mass splitting of 700 keV, and local aggregate boost factor of 65, in scenarios 1 (*left panel*) and 2 (*right panel*); see text for descriptions of the two scenarios. Regions to the left of and/or below the red dashed line are ruled out by constraints from the CMB; the self-interaction bounds lie at mediator masses below the range of this plot. In the regions overlaid in solid black, the approximation we have used for the multi-state Sommerfeld enhancement is expected to break down; in the grayed-out regions, the model is unphysical as the required cross section to obtain the correct relic density is smaller than the minimal contribution from t-channel annihilation into dark gauge bosons. The dark gauge boson is assumed to decay into electrons only, in which case this boost factor and DM mass provide a good fit to the PAMELA and *Fermi* data. When the gauge boson mass exceeds twice the muon mass, the true final state may become more complicated, so this region is indicated by cross-hatching.

mediator masses where our approximations are expected to break down. We note that as expected, a local boost of 65 is permitted for a 200 MeV mediator and 1 TeV DM in the zero-Δ limit.

[1] O. Adriani et al. (PAMELA), Nature **458**, 607 (2009), 0810.4995.
[2] A. A. Abdo et al. (The Fermi LAT), Phys. Rev. Lett. **102**, 181101 (2009), 0905.0025.
[3] M. Ackermann et al. (Fermi LAT), Phys. Rev. **D82**, 092004 (2010), 1008.3999.
[4] J. Chang et al., Nature **456**, 362 (2008).
[5] A. D. Panov et al., Astrophys. Space Sci. Trans. **7**, 119 (2011), 1104.3452.
[6] D. P. Finkbeiner and N. Weiner, Phys. Rev. **D76**, 083519 (2007), astro-ph/0702587.
[7] N. Arkani-Hamed, D. P. Finkbeiner, T. R. Slatyer, and N. Weiner, Phys. Rev. **D79**, 015014 (2009), 0810.0713.
[8] M. Pospelov and A. Ritz, Phys. Lett. **B671**, 391 (2009), 0810.1502.
[9] D. P. Finkbeiner, L. Goodenough, T. R. Slatyer, M. Vogelsberger, and N. Weiner (2010), 1011.3082.
[10] M. Lattanzi and J. I. Silk, Phys. Rev. **D79**, 083523 (2009), 0812.0360.
[11] M. Kuhlen and D. Malyshev, Phys. Rev. **D79**, 123517 (2009), 0904.3378.
[12] M. Kuhlen, P. Madau, and J. Silk, Science **325**, 970 (2009), 0907.0005.
[13] J. Bovy, Phys. Rev. **D79**, 083539 (2009), 0903.0413.
[14] Q. Yuan et al., JCAP **0912**, 011 (2009), 0905.2736.
[15] A. C. Vincent, W. Xue, and J. M. Cline (2010), 1009.5383.
[16] M. D. Kistler and J. M. Siegal-Gaskins, Phys. Rev. **D81**, 103521 (2010), 0909.0519.
[17] M. Kamionkowski, S. M. Koushiappas, and M. Kuhlen, Phys. Rev. **D81**, 043532 (2010), 1001.3144.
[18] M. Pospelov, Phys.Rev. **D80**, 095002 (2009), 0811.1030.
[19] B. Batell, M. Pospelov, and A. Ritz, Phys. Rev. **D79**, 115008 (2009), 0903.0363.
[20] R. Essig, P. Schuster, and N. Toro, Phys. Rev. **D80**, 015003 (2009), 0903.3941.
[21] M. Reece and L.-T. Wang, JHEP **07**, 051 (2009).

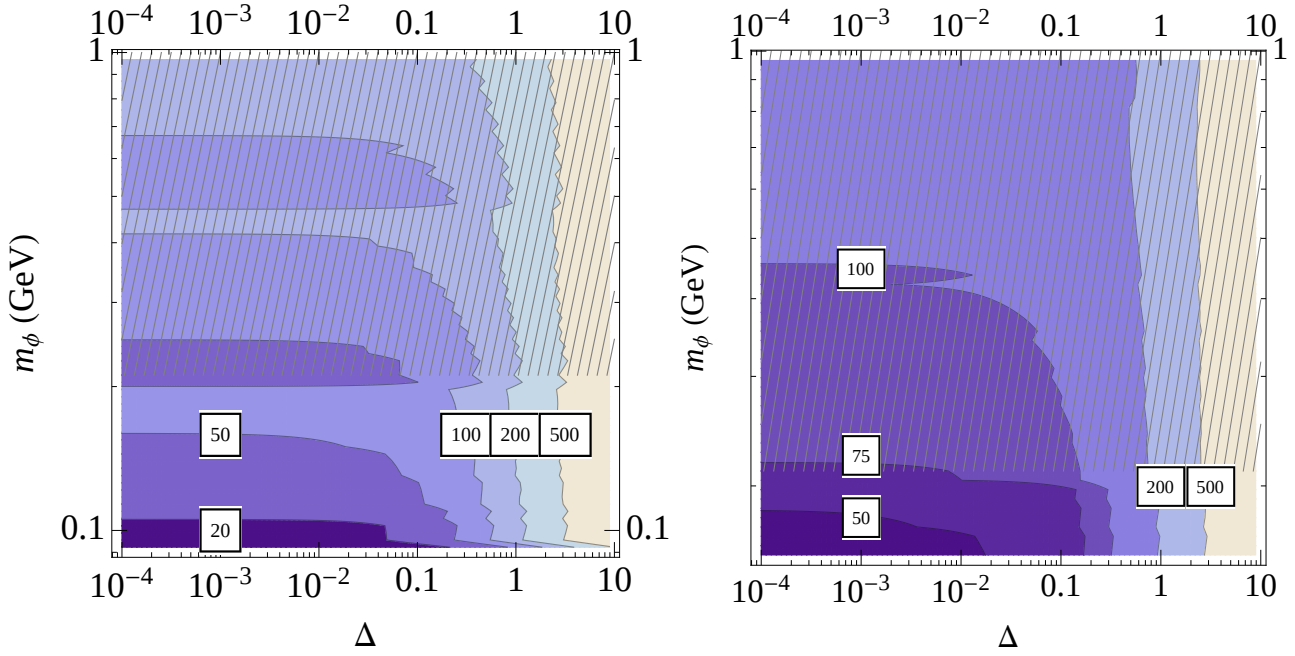

FIG. 7: The maximum local boost factor for 1 TeV dark matter with a 700 keV mass splitting, consistent with constraints from the thermal relic density, the CMB, self-interaction bounds, and naturalness (in the sense of not relying on the resonance peaks), in scenarios 1 (*left panel*) and 2 (*right panel*). The dark gauge boson is assumed to decay into electrons only; when the gauge boson mass exceeds twice the muon mass, the true final state may become more complicated, so this region is indicated by cross-hatching. The regions overlaid in solid black indicate where our approximation for the multi-state Sommerfeld enhancement is expected to break down.

[22] B. Aubert et al. (BABAR Collaboration) (2009), 0908.2821.
[23] J. D. Bjorken, R. Essig, P. Schuster, and N. Toro, Phys. Rev. **D80**, 075018 (2009), 0906.0580.
[24] B. Batell, M. Pospelov, and A. Ritz, Phys. Rev. **D80**, 095024 (2009), 0906.5614.
[25] F. Bergsma et al. (CHARM), Phys. Lett. **B157**, 458 (1985).
[26] E. M. Riordan et al., Phys. Rev. Lett. **59**, 755 (1987).
[27] J. D. Bjorken et al., Phys. Rev. **D38**, 3375 (1988).
[28] A. Bross et al., Phys. Rev. Lett. **67**, 2942 (1991).
[29] S. Andreas and A. Ringwald (2010), arXiv:1008.4519.
[30] F. Archilli et al. (2011), arXiv:1107.2531.
[31] H. Merkel et al. (A1), Phys. Rev. Lett. **106**, 251802 (2011).
[32] S. Abrahamyan et al. (2011), 1108.2750.
[33] R. Essig, P. Schuster, N. Toro, B. Wojtsekhowski, et al., JLab Experiment E12-10-009 (2009).
[34] R. Essig, P. Schuster, N. Toro, and B. Wojtsekhowski, JHEP **02**, 009 (2011).
[35] Jefferson Lab PAC37 Proposal PR-11-006. http://www.jlab.org/exp-prog/proposals/11prop.html.
[36] B. Wojtsekhowski, AIP Conf. Proc. **1160**, 149 (2009).
[37] M. Freytsis, G. Ovanesyan, and J. Thaler, JHEP **01**, 111 (2010).
[38] e.g. http://www.desy.de/˜ringwald/axions/talks/bim_dark_follow.pdf.
[39] N. F. Bell and T. D. Jacques (2008), 0811.0821.
[40] G. Bertone, M. Cirelli, A. Strumia, and M. Taoso (2008), 0811.3744.
[41] L. Bergstrom, G. Bertone, T. Bringmann, J. Edsjo, and M. Taoso (2008), 0812.3895.
[42] M. Cirelli and P. Panci, Nucl. Phys. **B821**, 399 (2009), 0904.3830.
[43] M. Pato, L. Pieri, and G. Bertone, Phys. Rev. **D80**, 103510 (2009), 0905.0372.
[44] P. Meade, M. Papucci, A. Strumia, and T. Volansky, Nucl. Phys. **B831**, 178 (2010), 0905.0480.
[45] M. Cirelli, P. Panci, and P. D. Serpico, Nucl. Phys. **B840**, 284 (2010), 0912.0663.
[46] M. Papucci and A. Strumia, JCAP **1003**, 014 (2010), 0912.0742.
[47] G. Hutsi, A. Hektor, and M. Raidal, JCAP **1007**, 008 (2010), 1004.2036.
[48] W. J. G. de Blok, Advances in Astronomy **2010** (2010), 0910.3538.
[49] G. Zaharijas, A. Cuoco, Z. Yang, and J. Conrad (for the Fermi-LAT) (2010), 1012.0588.
[50] M. Cirelli and J. M. Cline (2010), 1005.1779.
[51] J. F. Navarro et al. (2008), 0810.1522.
[52] G. R. Blumenthal, S. M. Faber, R. Flores, and J. R. Primack, Astrophys. J. **301**, 27 (1986).

[53] O. Y. Gnedin, A. V. Kravtsov, A. A. Klypin, and D. Nagai, Astrophys. J. **616**, 16 (2004), arXiv:astro-ph/0406247.
[54] A. El-Zant, I. Shlosman, and Y. Hoffman (2001), astro-ph/0103386.
[55] A. A. El-Zant, Y. Hoffman, J. Primack, F. Combes, and I. Shlosman, Astrophys. J. **607**, L75 (2004), astro-ph/0309412.
[56] E. Romano-Diaz, I. Shlosman, Y. Hoffman, and C. Heller (2008), 0808.0195.
[57] F. Governato et al. (2009), 0911.2237.
[58] E. Romano-Diaz, I. Shlosman, C. Heller, and Y. Hoffman, Astrophys. J. **702**, 1250 (2009), 0901.1317.
[59] M. G. Abadi, J. F. Navarro, M. Fardal, A. Babul, and M. Steinmetz (2009), 0902.2477.
[60] S. E. Pedrosa, P. B. Tissera, and C. Scannapieco (2009), 0910.4380.
[61] P. B. Tissera, S. D. M. White, S. Pedrosa, and C. Scannapieco (2009), 0911.2316.
[62] M. Su, T. R. Slatyer, and D. P. Finkbeiner (2010), 1005.5480.
[63] S. Cassel (2009), 0903.5307.
[64] T. R. Slatyer, JCAP **1002**, 028 (2010), 0910.5713.
[65] N. Padmanabhan and D. P. Finkbeiner, Phys. Rev. **D72**, 023508 (2005), astro-ph/0503486.
[66] S. Galli, F. Iocco, G. Bertone, and A. Melchiorri, Phys. Rev. **D80**, 023505 (2009), 0905.0003.
[67] T. R. Slatyer, N. Padmanabhan, and D. P. Finkbeiner, Phys. Rev. **D80**, 043526 (2009), 0906.1197.
[68] R. Essig, N. Sehgal, L. E. Strigari, M. Geha, and J. D. Simon (2010), 1007.4199.
[69] J. L. Feng, M. Kaplinghat, and H.-B. Yu (2010), 1005.4678.
[70] J. L. Feng, M. Kaplinghat, and H.-B. Yu (2009), 0911.0422.
[71] M. R. Buckley and P. J. Fox, Phys. Rev. **D81**, 083522 (2010), 0911.3898.
[72] S. A. Khrapak, A. V. Ivlev, G. E. Morfill, and S. K. Zhdanov, Phys. Rev. Lett. **90**, 225002 (2003).
[73] S. Hannestad (2000), astro-ph/0008422.
[74] A. Loeb and N. Weiner (2010), 1011.6374.
[75] T. Delahaye et al., Astron. Astrophys. **501**, 821 (2009), 0809.5268.
[76] J. B. Dent, S. Dutta, and R. J. Scherrer, Phys. Lett. **B687**, 275 (2010), 0909.4128.
[77] J. Zavala, M. Vogelsberger, and S. D. M. White, Phys. Rev. **D81**, 083502 (2010), 0910.5221.
[78] L. Pieri, J. Lavalle, G. Bertone, and E. Branchini, Phys. Rev. **D83**, 023518 (2011), 0908.0195.
[79] K. N. Abazajian, S. Blanchet, and J. P. Harding (2010), 1011.5090.
[80] A. A. Abdo et al. (Fermi-LAT), JCAP **1004**, 014 (2010), 1002.4415.
[81] J. Zavala, M. Vogelsberger, T. R. Slatyer, A. Loeb, and V. Springel (2011), 1103.0776.
[82] J. Zavala, V. Springel, and M. Boylan-Kolchin, Mon. Not. Roy. Astron. Soc. **405**, 593 (2010), 0908.2428.
[83] T. Bringmann, New J. Phys. **11**, 105027 (2009), 0903.0189.
[84] J. Hisano, S. Matsumoto, and M. M. Nojiri, Phys. Rev. Lett. **92**, 031303 (2004), hep-ph/0307216.
[85] J. Hisano, S. Matsumoto, M. M. Nojiri, and O. Saito, Phys. Rev. **D71**, 063528 (2005), hep-ph/0412403.
[86] S. Galli, F. Iocco, G. Bertone, and A. Melchiorri (2011), 1106.1528.
[87] G. Hutsi, J. Chluba, A. Hektor, and M. Raidal (2011), 1103.2766.
[88] The Sommerfeld enhancement was first discussed in the context of dark matter by [84, 85].
[89] However, if the dwarf galaxies contain significant substructure, the bounds on the low-velocity annihilation cross section from dwarfs will be strengthened; we consider this bound in substructure-dominated scenarios in §IV D.
[90] More recent analyses using *WMAP* 7 data have found limits lower by a factor of $\sim 2/3$ [86] and larger by a factor of $\sim 19/14$ [87]: the reason for this factor-of-2 discrepancy is not well understood, and for simplicity we use the *WMAP* 5 limits for illustration in this work.
[91] We note that this threshold is treated as a limit in some literature, but is not clearly excluded. A more conservative upper limit is $\sigma/m_\chi \lesssim 5.6\ \mathrm{cm}^2/\ \mathrm{g}$, the threshold for collapse of the dwarf halo's core in less than a Hubble time [73] (we thank M. Kuhlen for bringing this point to our attention). In Sommerfeld-enhanced models, the latter limit is never reached in parameter regions permitted by the dwarf evaporation limit, discussed below.
[92] In different formulations, this can describe the distribution of subhalos in main halos [78, 81] or the probability of finding a particle in a region of density ρ (related to the number of small subhalos of that characteristic density) [17].