Early structure formation constraints on the ultralight axion in the postinflation scenario

Vid Iršič,[1,2,3,*] Huangyu Xiao,[4] and Matthew McQuinn[1]

[1]*Department of Astronomy, University of Washington, 3910 15th Avenue NE, Seattle, Washington 98195-1580, USA*

[2]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*

[3]*Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom*

[4]*Department of Physics, University of Washington, 3910 15th Avenue NE, Seattle, Washington 98195-1580, USA*

(Received 3 December 2019; accepted 22 May 2020; published 18 June 2020)

Many works have concentrated on the observable signatures of dark matter being an ultralight axionlike particle (ALP). We concentrate on a particularly dramatic signature in the late-time cosmological matter power spectrum that occurs if the symmetry breaking that establishes the ALP happens after inflation—white-noise density fluctuations that dominate at small scales over the adiabatic fluctuations from inflation. These fluctuations alter the early history of nonlinear structure formation. We find that for symmetry-breaking scales of $f_A \sim 10^{13} - 10^{15}$ GeV, which require a high effective maximum temperature after inflation, ALP dark matter with a particle mass of $m_A \sim 10^{-13} - 10^{-20}$ eV could significantly change the number of high-redshift dwarf galaxies, the reionization history, and the Lyα forest. We consider all three observables. We find that the Lyα forest is the most constraining of current observables, excluding $f_A \gtrsim 10^{15}$ GeV ($m_A \lesssim 10^{-17}$ eV) in the simplest model for the ALP and considerably lower values in models coupled to a hidden asymptotically free strongly interacting sector ($f_A \gtrsim 10^{13}$ GeV and $m_A \lesssim 10^{-13}$ eV). Observations that constrain the extremely high-redshift tail of reionization may disfavor similar levels of isocurvature fluctuations as the forest. Future $z \sim 20$–30 21 cm observations have the potential to improve these constraints further using that the supersonic motions of the isocurvature-enhanced abundance of $\sim 10^4$ $M_\odot$ halos would shock heat the baryons, sourcing large baryon acoustic oscillation features.

DOI: 10.1103/PhysRevD.101.123518

I. INTRODUCTION

The nature of dark matter remains one of the biggest unsolved puzzles in particle physics and cosmology. We think that dark matter is a particle produced in the early Universe via one of several established mechanisms. The foremost has it thermally produced and its abundance freezing out when nonrelativistic, which can result in the observed dark matter density if it has a weak-scale mass and interaction cross section—the so-called "WIMP miracle" (see, e.g., [1]). After decades of searching for the weakly interacting massive particle (WIMP), the limits on this scenario are becoming more stringent. Perhaps our second most favored mechanism is the misalignment mechanism, discovered for the axion of quantum chromodynamics (QCD) [2–4]. At early times when the Hubble rate is greater than the axion mass—a mass that is acquired by nonperturbative effects such as instantons—the axion field is stuck outside of the minimum of its potential. However, when the Hubble rate later becomes smaller than the axion mass, the axion field begins to oscillate coherently, behaving like nonrelativistic matter with an energy density set by its initial potential energy [5–8].

The misalignment mechanism is also how the early Universe could create dark matter in the form of ultralight axionlike particles (ALPs), also known as fuzzy dark matter. The misalignment mechanism may naturally produce an ALP relic abundance of the order of the dark matter abundance if the ALP is the Goldstone boson arising from a broken grand unified theory (GUT) to Planck-scale symmetry and if it later acquires a mass of $m_A \sim 10^{-20}$ eV [9]. The nonperturbative mass generation can also naturally explain such ultralight masses, with $m_A \sim 10^{-20}$ eV motivated by the estimated size of nonperturbative effects for the GUT coupling constant [8].

Our study focuses on such ultralight ALPs in the limit where the Peccei-Quinn symmetry breaking that establishes this particle (re)occurs after inflation. For

*vi223@cam.ac.uk

string-theory-motivated models, the anticipated ranges for the symmetry-breaking scale f_A are GUT to Planck scales [8,9], although models that allow a lower scale have been devised [10]. Too low of a symmetry-breaking scale would not generate the dark matter abundance: As our constraints probe the mass range of $m_A = 10^{-16} - 10^{-20}$ eV, this requires f_A just below the GUT scale with $\sim 10^{15} - 10^{16}$ GeV to generate the relic abundance. These high values for f_A (which are far above the Hubble scale during inflation so that this symmetry must be broken during this epoch) may be strained by cosmic microwave background (CMB) B-mode observations, which limit the energy scale of inflation to $V(\phi) \lesssim 1.7 \times 10^{16}$ GeV [11]. Our mechanism requires the symmetry to be reestablished after inflation. This reestablishment can occur if the maximum postinflation thermalization temperature is greater than f_A [12] or instead during preheating, where larger effective temperatures can naturally arise from the nonthermal distribution of resonantly produced particles [14,15].

We further consider models with an asymptotically free strongly interacting sector that mimics the behavior of the QCD axion (in which the particle mass increases after the ALP behaves behaves like dark matter). Such models allow a somewhat lower f_A to match the dark matter abundance (down to $f_A \sim 10^{13}$ GeV), at the cost of introducing a sub-MeV confinement scale. The cosmological constant problem can be solved by hundreds of ALPs connected with strongly coupled sectors (as such sectors allow nondegenerate vacuum minima owing to higher instanton contributions), possibly with several hidden sectors per decade in energy [16]. (See Ref. [17] for more discussion of the strongly interacting "axiverse" scenario, as there are some challenges to this scenario in our postinflationary picture.)

Just like with the QCD axion in this postinflation limit, different causally disconnected patches will acquire different energy vacua depending on the random angle $\theta \in [-\pi, \pi]$ the field rolled to after symmetry breaking in a given patch, with the horizon scale setting the coherence length until $m_A \sim H$ [4,19]. At this time, the vacuum energy is then converted into nonrelativistic axions with number density $\propto \theta^2$, leading to order-unity fluctuations in the abundance of axions on the horizon scale when $m_A \sim H$ [20]. The lighter the axion, the later this occurs, and the larger the horizon-scale coherence length of the fluctuations.

These isocurvature perturbations are potentially observable. For the QCD axion [21], the mass contained in the horizon $M_{H(m_A)}$ when $m_A \sim H$—which is also the scale where there are order-unity density fluctuations—is $M_{H(m_A)} \sim 10^{-10}\ \mathrm{M}_\odot$ [20,22,23] (and axion self-interactions can lead to larger enhancements on even smaller scales [24]). This leads to the collapse of "axion miniclusters" near this mass scale at matter radiation equality, resulting in much denser dark matter structures than would be produced by the scale-invariant potential fluctuations from inflation. Still, there is no smoking gun observable for verifying whether these minute structures exist, although see Ref. [25] for a promising possibility. In contrast, for ultralight axions that are relevant for small-scale structure problems, $M_{H(m_A)}$ can approach the sizes of dwarf galaxies, and the rms fluctuations produced via these isocurvature fluctuations scale as M^{-1}, where M is the average mass contained within a spherical volume. These fluctuations are still larger than the inflationary perturbations even on mass scales of $M \gg M_{H(m_A)}$. This property has been used to place constraints on the ultralight ALPs via the CMB [26,27].

This paper shows that other observables are much more constraining than the CMB. We first focus on the Lyα forest, which is the quasilinear "large-scale" structure formation probe sensitive to the smallest scales. In addition, we show that such isocurvature perturbations could significantly affect the formation of the first stars and galaxies in the redshift of $z \sim 6$–20 Universe and discuss potential constraints. Since these isocurvature perturbations lead to the formation of dark matter halos at much higher redshifts than would occur in the standard cosmology, we also consider whether the shocks from these supersonic dense structures could ionize and heat the postrecombination

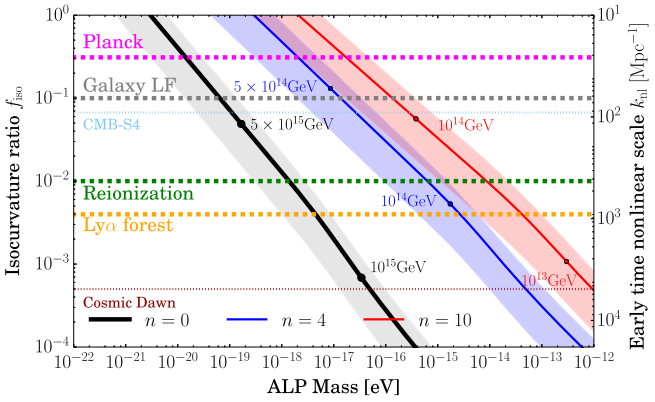

FIG. 1. The $f_{\rm iso}$—m_a constraints for the ALP dark matter in the postinflation scenario, assuming the ALP is all of the dark matter. The isocurvature to adiabatic ratio $f_{\rm iso}$ applies at the pivot scale of $k_\star = 0.05\ \mathrm{Mpc}^{-1}$, and the second y axis gives the nonlinear scale defined by $\Delta^2_{\rm iso} \equiv (k/k_{\rm nl})^3$ at early times. The solid lines show this mapping for our fiducial choice of $A_{\rm osc} = 0.1$, and the shaded bands span $0.01 \leq A_{\rm osc} \leq 0.3$. The horizontal dashed lines correspond to current upper limits on $f_{\rm iso}$ obtained using different datasets: the *Planck* 2018 CMB measurements in magenta from Ref. [27], Hubble Space Telescope (HST) galaxy luminosity function measurements in gray (Sec. IV), a combination of constraints on the reionization history in green (Sec. V), and finally Ly-α forest in orange (Sec. III). The horizontal dotted line in light blue is a forecast for CMB-S4 [27], while the horizontal dotted line in dark red is a rough forecast for where shock heating should qualitatively change the $z \sim 2021$ cm signal (Sec. VI). The labeled dots give the value of the symmetry-breaking scale f_a.

Universe. Figure 1 summarizes our constraints on the fractional amplitude of isocurvature fluctuations $f_{\rm iso}$ (defined shortly) and axion mass m_a, where dashed lines represent existing constraints and dotted represent forecasts for future efforts.

This paper is organized as follows. Section II describes the character of ALP isocurvature fluctuations. Then, we discuss the limits from several observables: the Lyα forest (Sec. III), the high-redshift galaxy luminosity function (Sec. IV), measurements that constrain early Universe star formation from the electron scattering optical depth through reionization (Sec. V), and finally from future 21 cm observations and the potential shock heating of cosmic gas (Sec. VI). While some of these observables are inherently very astrophysical and, hence, the constraints dependent on modeling, we show that isocurvature fluctuations can result in qualitatively different trends. Our numerical calculations take $\Omega_m = 0.308$, $\Omega_\Lambda = 0.692$, $\Omega_b = 0.0484$, $h = 0.678$, $\sigma_8 = 0.815$, and $n_s = 0.968$, consistent with the results of Ref. [28]. When convenient, our calculations will use natural units where $c = \hbar = k_b = 1$. Cosmological distances and wave numbers are given in comoving units. All mass function calculations use the mass function of Sheth and Tormen [29]. Even though we are considering nonstandard cosmologies, the well-tested universality of the mass function means that Sheth and Tormen [29] still holds at the 10% fractional level [30], and some of us have also have been involved in running simulations testing this.

II. ISOCURVATURE POWER FROM POSTINFLATION AXIONS

After perturbative effects break the degeneracy between different θ vacua, the vacuum misalignment of the ALP translates into a component that behaves like nonrelativistic matter with local density [4,31,32]:

$$\rho_a(T,\theta_{\rm ini}) = \frac{1}{2} f_a^2 m_a(T) m_a(T_{\rm osc}) \theta_{\rm ini}^2 \left(\frac{a(T_{\rm osc})}{a(T)}\right)^3, \quad (1)$$

where $\theta_{\rm ini}$ is the initial vacuum misalignment angle after symmetry breaking, $a(T)$ is the scale factor, and $m_a(T)$ is the axion mass. This formula holds after the axion starts oscillating at an oscillation temperature that we define as $m_a = 3H(T_{\rm osc})$. Equation (1) allows for the possibility that the axion temperature is also evolving at $T_{\rm osc}$ as could occur in strongly interacting sectors (as discussed later). We average $\rho_a(T,\theta_{\rm ini})$ over space, noting that we use the simple relation for the spatial average $\langle\theta_{\rm ini}^2\rangle = \pi^2/3$, to calculate the average dark matter abundance. The axion decay constant f_A (which we also refer to as the "symmetry-breaking scale") will be adjusted to match the observed dark matter abundance.

Because different causal horizons have different $\theta_{\rm ini}$, this translates into a white spectrum of isocurvature fluctuations in the matter overdensity at times after the field behaves like nonrelativistic matter but well into the radiation era with a growing mode dimensionless power spectrum of (e.g., [27])

$$\Delta_{\mathcal{S}}^2(k) \equiv \frac{k^3}{2\pi^2} P_{\mathcal{S}}(k) = A_{\rm osc}\left(\frac{k}{k_{\rm osc}}\right)^3 \quad \text{at } k < k_{\rm osc}, \quad (2)$$

where $P_{\mathcal{S}}(k) \equiv V^{-1}|\tilde{\delta}_{\mathbf{k}}|^2$, V is the volume, $\tilde{\delta}_k$ is the Fourier transform of the configuration-space dark matter matter overdensity $\delta(\mathbf{x})$ [which we assume to be entirely composed of ALPs such that $\delta(\mathbf{x}) = \rho_a(\mathbf{x})/\langle\rho_a\rangle - 1$], $k_{\rm osc} = aH|_{T_{\rm osc}}$ is the size of the horizon when the ALP starts to oscillate in its potential [4], and $A_{\rm osc}$ sets the normalization for which the order-unity fluctuations on the oscillations scale mean $A_{\rm osc} \sim 1$. While irrelevant for this study, at scales $k \gtrsim k_{\rm osc}$ a sharp cutoff is expected, as the vacuum misalignment fluctuations have been smoothed out by the Kibble mechanism [19]. Typical values of $k_{\rm osc}$ are between 100 and 1000 Mpc^{-1} for the ALP masses of 10^{-19} and 10^{-17} eV, respectively. The signatures we study are sourced by structures that are coming from an order of magnitude smaller wave numbers.

Simulations of the QCD axion find that values of the isocurvature variance at initial conditions are $A_{\rm osc} \sim 0.01$–0.3 [27,33], somewhat smaller than unity because some of the misalignment power is not in the zero mode and because this signal is diluted by relativistic axions radiated by axionic strings. However, for our ALP we expect the details that shape $A_{\rm osc}$ to depend on the specific model. When we connect our results to the axion mass m_a, we take as a fiducial value $A_{\rm osc} = 0.1$, but our results are easily rescaled to other values.

We use the standard growth and transfer function parameterization to model the subsequent evolution of the isocurvature fluctuations (as well as the standard inflationary adiabatic fluctuations). We parameterize the isocurvature fluctuations as

$$\Delta_{\rm iso}^2(k,z) = D_{\rm iso}^2(z) T_{\rm iso}^2(k,z) A_{\rm iso}\left(\frac{k}{k_\star}\right)^3, \quad (3)$$

where $D_{\rm iso}(z)$ is the growth function that tends to a constant deep in the radiation era and $T_{\rm iso}^2(k,z)$ is the transfer function that is normalized to unity at high k [34]. This transfer function is approximately constant for modes that enter the horizon during radiation domination. We take $k_\star = 0.05$ Mpc^{-1} for the pivot scale. Similarly, for the adiabatic fluctuations from inflation

$$\Delta_{\rm ad}^2(k,z) = D_{\rm ad}^2(z) T_{\rm ad}^2(k,z) A_s\left(\frac{k}{k_\star}\right)^{n_s-1}, \quad (4)$$

with analogous definitions as for $\Delta_{\rm iso}^2$ except that the adiabatic transfer function is normalized to unity at

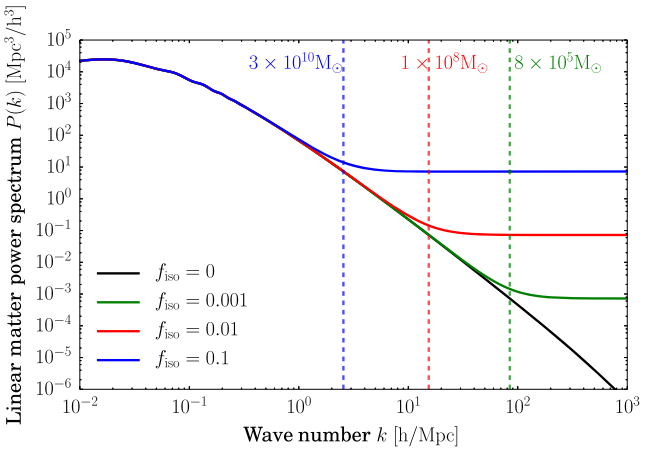

FIG. 2. The linear matter power spectrum at $z = 0$ for different values of $f_{\rm iso}$. The scales where isocurvature contribution become important, relative to the adiabatic power spectrum, are marked with vertical dashed lines and labeled by their $M = 4\pi/3\rho_m(z=0)k_c^{-3}$, where k_c is the wave number where adiabatic and isocurvature fluctuations are equal.

low k. For our chosen value of σ_8, $A_s = 2.054 \times 10^{-9}$. The total matter power at redshift z is the sum of the isocurvature and adiabatic contributions, $\Delta^2_{\rm iso} + \Delta^2_{\rm ad}$. The transfer functions at late times were calculated using the CAMB Boltzmann code solver [35]. Following convention, we define $f_{\rm iso}$ to be the ratio of isocurvature to adiabatic fluctuations at $k_\star = 0.05$ Mpc^{-1}:

$$f^2_{\rm iso} = \frac{A_{\rm iso}}{A_s} = \frac{A_{\rm osc}}{A_s}\left(\frac{k_\star}{k_{\rm osc}}\right)^3, \tag{5}$$

where the second equation uses that deep into the radiation-dominated universe $A_{\rm iso}/{k_\star}^3 = A_s/k^3_{\rm osc}$, since $D_{\rm iso}T_{\rm iso} \to 1$. In the late-time matter power, the ratio of isocurvature-sourced to adiabatic-sourced fluctuations is highly scale dependent, scaling approximately as k^3 at high wave numbers. This is illustrated in Fig. 2, where different colors represent different values of $f_{\rm iso}$, with the highest value of $f_{\rm iso}$ resulting in the highest small-scale power. Dashed vertical lines show the mass scale at which the adiabatic and isocurvature contributions to the power spectrum are equal. The contribution of isocurvature fluctuations becomes important at different mass scales, following the approximate scaling of $f_{\rm iso}$ with mass as $M^{1/2}$. This is a direct consequence of the definition of $f_{\rm iso}$, which is fixed on large scales ($k_\star$), and leads to a natural expectation that observables probing smaller mass scales will result in tighter constraints on $f_{\rm iso}$.

We also specify the level of isocurvature by its early-time nonlinear scale $k_{\rm nl}$, where $k_{\rm nl} \equiv k_{\rm osc}A_{\rm osc}^{-1/3}$ such that deep into the radiation era $\Delta^2_{\rm iso} = (k/k_{\rm nl})^3$. The nonlinear scale represents a more straightforward quantification of the white-noise power, because it does not convolve in the well-understood amplitude of adiabatic fluctuations and because it does not single out a specific $k_\star$.

One likely scenario is that the m_a does not exhibit strong temperature dependence in the early Universe. This limit applies to ALPs whose mass is acquired by nonperturbative effects associated with the perturbative gauge couplings in GUT theories [9]. In this case, the nonperturbative mass is exponentially suppressed relative to the symmetry-breaking scale, and the ALP field obtains its zero-temperature mass at $T \gg T_{\rm osc}$. We also consider a QCD-like case of a asymptotically free strongly interacting sector where the nonperturbative effects increase with decreasing temperature until the temperature reaches the confinement scale Λ; evolution of the mass occurs even after the ALP behaves like nonrelativistic matter if $\Lambda < T_{\rm osc}$, with the final mass equal to $m_A = \Lambda^2/f_a$. The ALP mass evolution can be characterized at $T \lesssim T_{\rm osc}$ by

$$m_a(T) = m_a\left(\frac{\Lambda}{T}\right)^n \quad \text{for } T > \Lambda, \tag{6}$$

$$m_a(T) = m_a \quad \text{otherwise}, \tag{7}$$

where we use the notation that m_a without an argument is the zero-temperature mass and where n parameterizes the temperature dependence of the instanton effects. The case $n = 4$ mimics the scaling found for the QCD axion, but the details of this scaling will depend on the strong sector. For the $n = 0$ perturbative case, we note that this parameterization still holds (trivially).

With this parameterization,

$$T_{\rm osc} = 3\left(\frac{10}{\pi^2 g_{\rm eff}}\right)^{1/4}[m_A(T_{\rm osc})M_P]^{1/2} \tag{8}$$

$$\propto \langle\theta^2_{\rm ini}\rangle^{-[n/(8+3n)]} m_a^{(4+n)/(8+3n)}, \tag{9}$$

$$k_{\rm osc} = a_{\rm osc}H(T_{\rm osc}) = \frac{T_{\rm cmb,0}}{T_{\rm osc}}\frac{m_A(T_{\rm osc})}{3} \propto T_{\rm osc}, \tag{10}$$

$$f_A \propto \langle\theta^2_{\rm ini}\rangle^{-2/(8+3n)} m_a^{-(2+n)/(8+3n)}, \tag{11}$$

where $M_P = 1/\sqrt{8\pi G}$ is the reduced Planck mass and $T_{\rm osc}$ evaluates to 1–100 keV for m_A of interest, indicating $g_{\rm eff} \approx 3.4$. For the proportionality relations, we have eliminated the Λ dependence in favor of m_a and f_A. We note that at fixed $m_A(T_{\rm osc})$ the amplitude of isocurvature fluctuations does not depend on n, and our constraints in Fig. 1 translate to $m_A(T_{\rm osc}) = 10^{-20} - 10^{-17}$ eV. For our $n = 4$ (10) models in Fig. 1, the particle mass increases by 3 (4) orders of magnitude to reach m_a at $T = \Lambda$.

Figure 1 foreshadows the constraints we find in the following sections in the $m_a - f_{\rm iso}$ plane. The different horizontal limits show the upper limit on $f_{\rm iso}$, bounding the viable parameter space to be below the curves. The $n = 0$

corresponds to the most likely case where the mass is established well before the particle commences oscillations, and the QCD axion yields a scaling with $n = 4$. The dots on the lines correspond to the values of the decay constant f_A for those models (color coded to match the lines), while the shaded regions around the lines correspond to the uncertainty in the value of $A_{\rm osc}$. The sold lines themselves were evaluated at the value of $A_{\rm osc} = 0.1$.

III. LYMAN-α FOREST

The Lyman-α forest is used to infer the initial conditions using significantly smaller comoving scales than other large-scale structure observables, to 3D wave numbers of $k \approx 10$–100 Mpc^{-1} [36,37]. The Lyman-α forest circumvents many of the difficulties of modeling structure formation at these nonlinear scales by being sensitive exclusively to low densities ($\Delta \sim 1$ as the absorption of higher densities is saturated [38]), where our nonlinear models for the cosmic web appear to be under control [39–41] and where astrophysical processes appear to be less of a contaminant (e.g., [37]). Indeed, the forest has been used to place the tightest constraints on the small-scale cutoff in the spectrum of primordial matter fluctuations, which may owe to the free streaming of warm dark matter and the de Broglie wavelength of fuzzy dark matter [42–44]. In the context of ALPs, combining the Lyα constraints with the limits on the isocurvature fluctuations from the CMB can lead to interesting bounds on the tensor-to-scalar ratio [45].

A typical Lyα forest analysis is sensitive to 1D wave numbers between 0.1 and 10 Mpc/h, which would naively lead to a typical mass of 10^8 M$_\odot$ (see Fig. 2). However, the nonlinear mapping from the 3D density field to the 1D flux field in the quasar spectra makes the Lyα forest sensitive to even smaller wave numbers (see, e.g., [46]). Additionally, the nonlinearity of the gravitational evolution does not dominate over the clustering signal at high redshifts, which helps to better constrain cosmology at a given scale.

The forest is also sensitive to an enhancement in power as would occur from the white isocurvature fluctuations from axions in the postinflation scenario. Indeed, the allowed level of enhancement has been constrained in the context of primordial black holes, which also may have a white spectrum [46,47]. Conveniently, the adiabatic plus white-noise simulations run for the primordial black holes in Ref. [46] are the same as would be run in the context of ALP isocurvature perturbations; the difference comes in the interpretation of the isocurvature amplitude and how it is linked to the actual physical model. In particular, Murgia and co-workers [46] find that the isocurvature fraction of $f_{\rm iso} = \sqrt{A_{\rm iso}/A_s}$ at the pivot scale of $k = 0.05$ Mpc^{-1} should be lower than 0.004 at 2σ confidence level when adopting conservative priors on the thermal history. This constraint can be remapped to our models by solving Eqs. (5) and (11) for a given ALP mass evolution model. The relation between f_A and m_a is fixed by assuming that all of dark matter is composed of the axionlike particle. This gives a lower bound on the mass of the ALP of $m_A > 2 \times 10^{-17}$ eV for the most natural case of no mass evolution after the axion starts oscillating ($n = 0$). This constraint further shows that the forest is effectively able to probe structure in the dark matter to mass scales as small as $\sim 3 \times 10^7$ M$_\odot$ (using Fig. 2), a number that is helpful for putting the forest in context with the other constraints we discuss.

Figure 1 shows the constraints from the forest. A primary result of this paper is that we find the Lyα forest is more constraining than other probes, although future observations of the high-redshift universe using redshift 21 cm radiation may ultimately be more constraining.

IV. GALAXY LUMINOSITY FUNCTION

Small galaxies are a second observable that has been used to constrain the primordial fluctuations on small scales, with observations both probing them as satellite galaxies to the Milky Way [48] and at high redshifts when they are forming the bulk of their stars [49,50]. Since the white-noise isocurvature fluctuations in our ultralight axion models dramatically increase fluctuations on small scales, such scenarios may predict a large increase in the number of low-luminosity galaxies. Foreshadowing the result of this section, for galaxies that are directly observable in the future, we find that this enhancement is small for the $f_{\rm iso}$ allowed by the forest, although in Sec. V we show that for smaller galaxies (whose effects can only be indirectly probed via their ionization and enrichment) the enhancement can be more substantial.

To model the enhanced number of small galaxies, we use a simple but successful model for star formation, where the predicted number density of galaxies n_g per UV luminosity between L and $L + dL$ is related to the halo mass function dn_h/dM_h by

$$\phi(L) \equiv \frac{dn_g}{dL} = \frac{dn_h}{dM_h}\frac{dM_h}{dL}. \tag{12}$$

This model assumes the common one-to-one mapping between halo mass and observed UV luminosity described by dM_h/dL. As this function has significant astrophysical uncertainty, we will use qualitatively different shapes for the galaxy luminosity function dn_g/dL, as a signature that a given axion cosmology is excluded.

To calculate the terms in Eq. (12), we use the Sheth-Tormen mass function [29] to model dn_h/dM_h [51]. The "universality" of the halo mass function makes it likely that the same mass function should be a good approximation to cases with isocurvature fluctuations (e.g., [30,53]). Additionally, we adopt a common assumption that a galaxy's star formation rate is proportional to its gas accretion rate $\dot{M}_b$, with proportionality

constant $f_\star(M_h, z)$ called stellar efficiency. Note that the star formation rate directly maps to the UV luminosity of the galaxy. We follow Furlanetto *et al.* [54] to calculate $f_\star$, who calculate it from an analytic model that considers an energy-regulated stellar feedback process plus virial shocking. In this model, the stellar efficiency of the baryons peaks at around $M_h = 10^{11.5}$ $M_\odot$, where it reaches values of just below 0.05. This efficiency has a steep tail toward smaller masses, reaching 10^{-3} by $M_h = 10^8$ $M_\odot$. One worry, which we will address, is that this efficiency depends on uncertain astrophysics and so any differences we find may not be distinguishable.

To model the gas accretion rate $\dot{M}_b$, numerical results are typically obtained from cosmological simulations (e.g., [55]), but for the isocurvature case, $\dot{M}_b$ has not been determined using simulations. However, the time evolution of the halo accretion rate is driven largely by the time evolution of the mass variance $\sigma(M)$ (see, e.g., [56]). We set

$$\frac{d\ln M_b}{dt} = \left|\frac{d\ln\sigma}{d\ln M_h}\right|^{-1}\frac{d\ln D}{dt}, \tag{13}$$

and D is the growth function. This allows us to build a consistent approach to calculating the gas accretion for any $f_{\rm iso}$. Our results on the gas accretion are in good agreement [56] in the limit they consider of $f_{\rm iso} = 0$.

Figure 3 shows the resulting comparison of the galaxy luminosity function. Our model is compared to the measurements of Refs. [57–60] but also include the $z = 6$ lensed galaxy sample of Ref. [61] that extends the measurement to fainter immensities. We use the standard convention of writing the UV luminosity in terms of absolute AB magnitude, where $M_{\rm AB} = -2.5\log_{10}(L_{\rm UV}) + M_{\rm ref}$, where $M_{\rm ref}$ is a constant. We have not performed any dust correction at this stage, as the typical corrections (e.g., [62]) are significant only for the higher-mass systems and lead to a shallower relation between the halo mass and the UV magnitude [63].

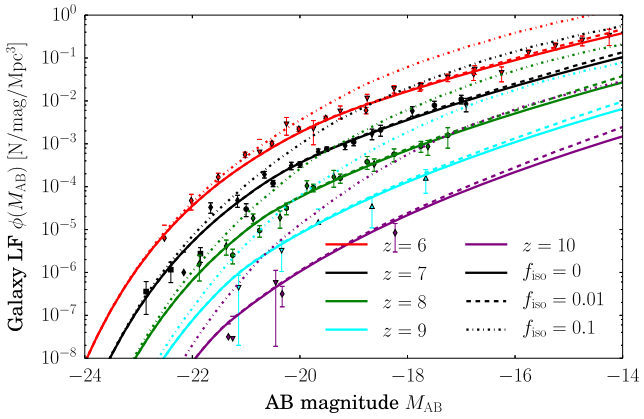

FIG. 3. The effect of white-noise isocurvature fluctuations on the galaxy luminosity function. The various line styles show different levels of isocurvature fluctuations, with color indicating redshift. Overplotted is a compilation of observational data ranging over typical redshifts probed by the future surveys. We consider $f_{\rm iso} = 0.1$ to be ruled out by these observations, as by this value the luminosity function has a qualitatively different behavior, especially at the highest redshifts probed.

However, including isocurvature fluctuations, even at the level already excluded by Lyα forest of $f_{\rm iso} = 0.01$, results only in a small signal at a lower end of the luminosity function. This is mainly due to the fact that even the observed high-redshift galaxies behind cluster lenses reside in $>10^9$ $M_\odot$ halos in our models. In contrast, the Lyα forest is sensitive to scales of $M \sim 10^8$ $M_\odot$, as illustrated in Fig. 2. We find that current observations of the high-redshift luminosity function rule out $f_{\rm iso} > 0.1$, as this leads to a large qualitative change that likely cannot be mimicked by the large astrophysical uncertainty in our star formation efficiency model. One can already start to see this large effect for the $f_{\rm iso} = 0.05$ model in Fig. 2. These limits translate into a lower bound on the ALP mass to be $m_a > 10^{-19}$ eV.

Future observations at higher redshifts would help in discriminating between different isocurvature models and could potentially provide constraints comparable to the ones derived from the small-scale structure of the Lyα forest. Namely, the James Webb Space Telescope (JWST) is able to go a few magnitudes deeper at $z \approx 6$ and, more importantly, has the infrared sensitivity that allows better constraints at higher redshifts. With lensed galaxy samples, JWST should be able to place similar constraints to HST at $z = 6$ (reaching to absolute magnitudes of $M_{\rm AB} = -14$) but all the way to $z = 10$, constraining $f_{\rm iso} \sim 0.01$. Unfortunately, astrophysical uncertainties require a qualitative change in behavior, making it difficult to probe beyond $f_{\rm iso} = 0.01$. Thus, the Lyα forest is likely to always provide a more sensitive probe than direct measurements of galaxy luminosity functions.

V. HIGH-REDSHIFT STAR FORMATION RATE AND REIONIZATION

Though we find that the galaxy luminosity function is not competitive with the Lyα forest, the collapsed fraction of halos that can form stars can be orders of magnitude larger than the $f_{\rm iso} = 0$ prediction at $z = 10$, and this difference is even larger at higher redshifts, if we take $f_{\rm iso} = 0.01$—comparable to the constraint coming from Lyα. This is illustrated in Fig. 4, noting that stars can form in halos with $M_h \gtrsim 10^{7-8}$ $M_\odot$ only if the gas condenses by cooling via atomic transitions and $M_h \gtrsim 10^{5-6}$ $M_\odot$ halos if instead by molecular ones. Unfortunately, the direct luminosity function measurements with HST (and in the future with JWST) are not sufficiently sensitive to detect the stars or galaxies that likely lie in these diminutive halos.

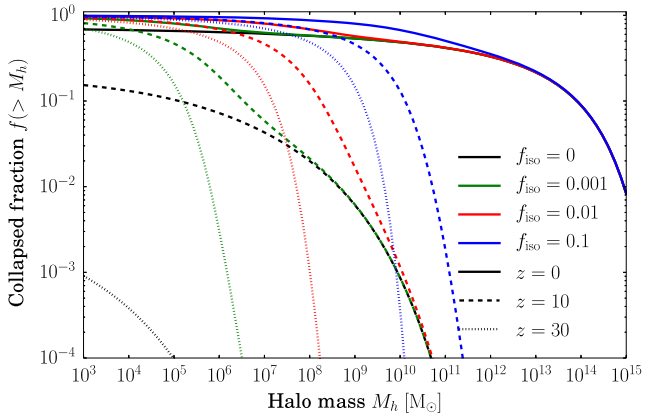

FIG. 4. The collapsed fraction in halos above a mass of M_h. Different colors show the collapsed fraction for isocurvature fractions of $f_{\rm iso} = 0.1$ (blue), $f_{\rm iso} = 0.01$ (red), $f_{\rm iso} = 0.001$ (green), and $f_{\rm iso} = 0$ (black). The line styles differentiate redshifts.

However, the enhanced extremely high-redshift star formation from an increased abundance of these small halos could also heat and ionize the cosmic gas (and their UV photons can pump the 21 cm line) in a manner that may allow constraints on $f_{\rm iso}$. There is also some indirect evidence that the smallest galaxies contribute disproportionately to the ionizing photons that escape into and, hence, ionize the intergalactic medium (IGM) (e.g., [64]), which would make our mass-independent escape in what follows conservative.

To illustrate just how much isocurvature fluctuations could change the mass in halos that are massive enough to host stars, we calculate the fraction of mass that is collapsed in halos with masses above M_h using the extended Press-Schechter theory [65,66]. This yields $f_{\rm coll}(>M_h) = {\rm erfc}(\nu(M_h)/\sqrt{2})$, where ${\rm erfc}(x) \equiv \pi^{-1/2}\int_x^\infty dx\exp[-x^2]$ and $\nu \equiv \delta_c/\sigma(M,z)$ and $\sigma(M,z)$ is the standard deviation of the density in a spherical top-hat Lagrangian volume with mass M. The virial temperature of the halo (the characteristic temperature the gas can shock heat) is the property of a halo that sets whether its gas can cool and form stars rather than the halo mass. The two are related by $M_h \propto [aT_{\rm vir}]^{3/2}$—at higher redshifts, the same virial $T_{\rm vir}$ halo has smaller M_h. For the isocurvature fluctuations with a constant power spectrum on small scales during the matter-dominated epoch, this leads to $\sigma^2(M) \propto a^2M^{-1}$, whereas for $f_{\rm iso} = 0$ we have $\sigma^2(M) \propto a^2\log[M]$. The result is that the redshift evolution of the collapsed fraction at a fixed virial radius is *much* flatter for masses where isocurvature fluctuations dominate, with the difference given by

$$f_{\rm coll}(>T_{\rm vir}) = {\rm erfc}[1.7\ Z_{10}] \quad \text{for adiabatic;}$$
$$f_{\rm coll}(>T_{\rm vir}) = {\rm erfc}[1.0\, f_{i,-2}^{-1} Z_{10}^{1/4} T_{\rm vir,4}^{3/4}] \quad \text{isocurvature,}$$

where $f_{\rm i,-2} \equiv f_{\rm iso}/10^{-2}$, $Z_{10} \equiv (1+z)/10$, and $T_{\rm vir,4} \equiv T_{\rm vir}/10^4$ K [67].

The former function falls off exponentially with increasing redshift for rare (large ν) objects noting asymptotic form $\sqrt{\pi}{\rm erf}(x) = \exp[-x^2](x^{-1} + \mathcal{O}(x^{-3}))$, whereas the latter (while still exponentially sensitive once the argument becomes greater than unity) is much flatter, allowing halos that can cool at much higher redshifts.

An enhancement in the number of star-forming halos in the manner of our white isocurvature fluctuations should lead to an enhanced number of hydrogen ionizing photons, causing the reionization of the Universe to start earlier and be a much more prolonged process. Such a reionization history would be constrained by direct estimates of the ionized fraction using quasar spectra and Lyman-α emitters. The ionized state of the intergalactic gas can be measured through the time evolution of the volume-averaged ionized fraction, that depends on the balance between recombination and ionization due to photoionization [68]:

$$\frac{dx_i}{dt} = \frac{d(\zeta f_{\rm coll})}{dt} - \bar{n}_H(t)\alpha_{\rm re}(T_e)C_{\rm HII}x_i, \qquad (14)$$

where $\zeta = A_{\rm He}f_\star f_{\rm esc}N_\gamma$ is the ionizing efficiency: a product of the correction factor for singly ionized helium, $A_{\rm He} \approx 1.22$; the star formation efficiency $f_\star$; the escape fraction of ionizing photons, $f_{\rm esc}$; and the average number of ionizing photons produced per stellar baryon, N_γ. In the recombination term, the number density of hydrogen, $\bar{n}_H$, is time dependent as $\bar{n}_H = \bar{n}_H(z=0)(1+z)^3$ at redshift z; the recombination rate $\alpha_{\rm re}$ is temperature dependent such that $\alpha_{\rm re}(T_e) = 2.6\times10^{-13}(T_e/10^4\ {\rm K})^{0.76}\ {\rm cm^3\,s^{-1}}$, at the electron temperature T_e; and the volume-averaged clumping factor is defined to be $C_{\rm HII} \equiv \langle n_e^2\rangle/\langle n_e\rangle^2$.

A rough approximation during HI reionization [68,69] is to fix $C_{\rm HII} = 3$, and $T_e = 10^4$ K. It would be natural to expect a redshift evolution of the clumping factor (see, e.g., [64]), which might change the reionization history. In our simple scenario, changing the value of the clumping factor to 5 (1) leads to a largely redshift-independent change in the ionized fraction in our calculations by a factor of 0.8 (1.4) (at least at high redshifts). The value of the mean number of ionizing photons produced, N_γ, depends on the initial mass function and metallicity of the stellar population. We use $N_\gamma = 4,000$ for population II (pop-II) stars, assuming Salpeter initial mass function (IMF) and 5% of the solar metallicity (although the results are weakly sensitive to these choices at least assuming empirically motivated IMFs). Pop-II stars are the second generation of stars that are born in metal-enriched gas and likely have properties similar to stars observed at low redshifts. Unless otherwise stated, we use the escape fraction of 20% for the pop-II stars. In the fiducial pop-II model, we assume all halos above $M_{\rm min}$ form stars, and at each redshift the value of

$M_{\rm min}$ is fixed to the mass at the virial temperature of $T_{\rm vir} = 10^4$ K. The basic photoionization rate can be evaluated using the halo mass accretion rates discussed in Sec. IV,

$$\frac{d(\zeta f_{\rm coll})}{dt} = A_{He} N_\gamma f_{\rm esc} \int_{M_{\rm min}}^{\infty} \frac{dM_h}{\bar{\rho}_m} n(M_h) f_\star \dot{M}_h, \quad (15)$$

where $f_\star$ is the mass-dependent stellar efficiency and $n(M_h)$ is the halo mass function.

In the context of the early star formation, a population III (pop-III) stellar contribution is often discussed, which is the first generation of stars which are born metal-free and expected to be more massive. Since this contribution is at present largely unconstrained [70], we adopt a toy model to characterize their effect on the progression of the reionization. In this case, an additional photoionization term is added, mimicking the structure of $d(\zeta_{\rm III} f_{\rm coll})/dt$, but with the ionizing efficiency characteristic of the pop-III models. Namely, following Eq. (15), we write down the pop-III photoionization rate as

$$\frac{d(\zeta_{\rm III} f_{\rm coll})}{dt} = A_{He} N_\gamma^{\rm III} \int_{M_{\rm min}^{\rm III}}^{M_{\rm min}} \frac{dM_h}{\bar{\rho}_m} n(M_h) f_\star^{\rm III} \dot{M}_h. \quad (16)$$

The integration is only over halos where molecular cooling is efficient and atomic is not (as atomic leads to our normal mode of star formation), i.e., between $T_{\rm vir} = 500$ K ($M_{\rm min}^{\rm III}$), warm enough to excite rotational transitions of molecular hydrogen, and the mass at the virial temperature of 10^4 K ($M_{\rm min}$). We use $N_\gamma^{\rm III} = 40,000$ as anticipated for the hotter photospheres of these metal-free stars [71] and assume that all ionizing photons escape as anticipated for star formation in these diminutive halos. We also take a stellar efficiency of $f_\star^{\rm III} = 5 \times 10^{-4}$, although the escape of ionizing photons can be pulled into this parameter. This efficiency is on the lower end of what is typically used in the literature [72,73], with most commonly used values being $10^{-3} - 10^{-2}$. However, in our simplified model, our fiducial value of $f_\star^{\rm III}$ leads to the star formation rate density of pop-III stars comparable to that of Ref. [70] (see Ref. [74]).

Once enough stars form in the Universe, the ∼11 eV Lyman-Werner radiation they produce dissociates molecular hydrogen, turning off cooling in molecular cooling halos and preventing the formation of further pop-III stars [75,76]. To model this, we follow Refs. [70,77,78], where we modify the lower integration limit ($M_{\rm min}^{\rm III}$) in Eq. (16) to also include self-regulations due to Lyman-Werner background. The numerical calculations of Refs. [79,80] found that the gas is able to cool in halos with mass

$$M_{\rm min}^{\rm III} = M_h(T_{\rm vir} = 500\ {\rm K})[1 + 6.69 F_{\rm LW,21}^{0.47}], \quad (17)$$

where $F_{\rm LW,21}$ is the Lyman-Werner intensity integrated over a solid angle in units of 10^{-21} erg s^{-1} Hz^{-1} cm^{-2}. To estimate the Lyman-Werner intensity given a star formation rate ($\dot{\rho}_{\rm SFR}$), we use the relations of Refs. [70,78]:

$$F_{\rm LW,21} = 7.22 \frac{(1+z)^3}{H(z)} e^{-\tau_{\rm LW}} (N_{\rm LW}^{\rm II} \dot{\rho}_{\rm SFR}^{\rm II} + N_{\rm LW}^{\rm III} \dot{\rho}_{\rm SFR}^{\rm III}), \quad (18)$$

where $H(z)$ is the Hubble rate of expansion and $\tau_{\rm LW}$ is the intergalactic opacity for the Lyman-Werner photons, which can be 1–2 in the absence of dissociations [81] and can be larger once the first HII regions have formed [82]. We use $\exp(-\tau_{\rm LW}) = 0.5$; however, we note that in the isocurvature model the value of $\tau_{\rm LW}$ might increase due to more small-scale structure obscuring the Lyman-Werner background.

The number of Lyman-Werner photons produced per baryon in stars is taken to be $N_{\rm LW}^{\rm II} = 9690$ for pop-II stars and $N_{\rm LW}^{\rm III} = 100,000$ for pop-III stars [78]. The value of $\dot{\rho}_{\rm SFR}$ is modeled through Eqs. (15) and (16), such that $\dot{\rho}_{\rm SFR} = f_\star d(f_{\rm coll})/dt$. We use an iterative process to determine the value of $\dot{\rho}_{\rm SFR}^{\rm III}$ that satisfies Eqs. (16)–(18).

We also multiply Eq. (16) by $(1 - x_i)$ to account for the photoheating. This term becomes important only toward the end of reionization at lower redshifts but prevents the pop-III photoionization term from resulting in an overly large optical depth contribution in the range of 10–15. The functional form of the above model is an approximate way to characterize the self-regulation of the pop-III stellar population in the early Universe. Simpler models regulated by the average ionized fraction (e.g., [83]) give very similar results. We would also comment that relations in Refs. [70,78] that we use to derive Eqs. (17) and (18) were empirically determined from CDM simulations. An approach based on simulations is most likely required to model the details of the pop-III star formation history in the presence of the isocurvature fluctuations.

However, not including any self-regularization leads to larger ionized fractions earlier in its evolution, which violate the observational constraints shown in Fig. 5, as well as the integrated optical depth from *Planck* (see below). Thus, some form of self-regularization is important to implement, but the exact details of the model do not change the quantitative picture that including the isocurvature fluctuations leads to a slower decrease of the ionized fraction at higher redshifts, compared to just pop-III star formation, which is illustrated in Fig. 5.

Figure 5 shows how the ionized fraction evolves in the redshift range probed by the measurements. Current observations from a variety of sources are plotted on Fig. 5: Lyα dark pixels ([84] in gray), Lyα emitters ([85–87] in brown), and quasar (QSO) damping wings ([88–90] in green). The fiducial model (black solid line) uses only pop-II photoionization rates, with $f_{\rm esc} = 0.2$ and no isocurvature fluctuations ($f_{\rm iso} = 0$). The effect of including axion isocurvature fluctuations (red lines) exhibits a distinctly

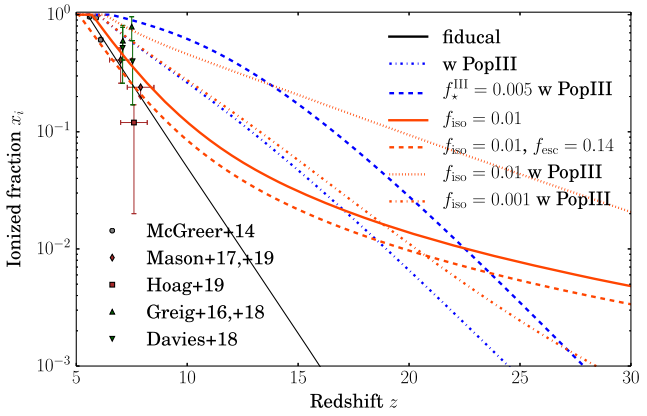

FIG. 5. The evolution of the ionized fraction of the intergalactic gas during reionization. The fiducial model assumes a given stellar efficiency described in Sec. IV. The effect of axion isocurvature fluctuations is shown for various values of $f_{\rm iso}$ and also varying assumptions about the escape fraction of pop-II stars (orange dashed line) or including a contribution from pop-III stars (blue and orange dot-dashed lines). Overplotted is a compilation of observational constraints on the ionized fraction coming from Lyα dark pixels (gray line), Lyα emitters (brown line), and QSO damping wings (green line).

longer tail of reionization, where the ionized fraction starts to increase much earlier and at a steadier rate than for the no isocurvature case. At lower redshift, where the ionized fraction can be currently estimated, the effect of the isocurvature fluctuations is slightly degenerate with the escape fraction of pop-II stars (red dashed line).

On the other hand, the effect of pop-III stars is prominent at higher redshifts (green dot-dashed line) and in tandem with the isocurvature fluctuations (dot-dashed red line) can create a boost to the ionized fraction such that it evolves much slower between redshifts of 25 and 10, potentially creating a strong observable signal of the isocurvature modes in the future observations. However, enhancing the star formation efficiency for pop-III stars to 0.005 as used in Ref. [73] increases the ionized fraction evolution even without isocurvature fluctuations (green dashed line in Fig. 5), making it not obvious that the astrophysics of star formation can be robustly disentangled from $f_{\rm iso}$. Nevertheless, at a high enough redshift all our isocurvature models cross the green-dashed line in Fig. 5 that corresponds to this extreme case of pop-III stellar efficiency. This is the unique signal of the isocurvature models in the ionization history, resulting from the nearly redshift-independent collapse fraction in such models.

Future observations by ground-based surveys (e.g., UKIDSS [91]; VIKING [92]; VHS [93]; UHS [94]) and wide-field surveys (e.g., Euclid, WFIRST, WEAVE, J-PAS) in combination with high signal-to-noise spectra from JWST would be more sensitive to the differences between the models. In particular, measuring the ionized fraction during the cosmic dawn epoch ($z > 15$) can lead to stronger constraints on the isocurvature fluctuations.

Another possibility of constraining the reionization process is utilizing the measurements of the CMB anisotropy, in particular, the effect of the CMB Thomson scattering off of free electrons. Since the redshift where this would occur ($z < 20$) is relatively closer than the surface of last scattering, this physical process affects predominantly large scales of the CMB fluctuations. The CMB constraints from the *Planck* satellite on the τ_e are very strong [95], as is shown by the gray band in Fig. 6. The axion isocurvature model has a different signal in the Thomson scattering optical depth, which primarily reflects the prolonged redshift evolution of the reionization process seen in Fig. 5. However, we note that reionization effects on the CMB are not just as a single number τ_e, as an earlier tail ionization creates polarization anisotropies at smaller scales [96,97]. An extended reionization is constrained by the *Planck* satellite to be $\tau_e(15, 30) < 0.007$, where this notation indicates the optical depth contributed between $z = 15$ and $z = 30$ [95]. [The *Planck* limits on the tail of reionization vary only slightly with the assumed priors and can lower the bound to $\tau(15, 30) < 0.006$ if flat priors are chosen on the positions of the knots on which τ is interpolated.]

The limits on the tail of reionization are most constraining for models with an earlier star formation, in particular, if the contribution of pop-III stars is included. Of the models plotted in Fig. 6, the models with $f_{\rm iso} = 0.01$ and including pop-III star formation are clearly excluded, with $\tau_e(15, 30) = 0.018$ (dotted red line) as shown in Fig. 7. On the other hand, with the typical pop-III star formation rate, the current data are not excluding a lower value of $f_{\rm iso} = 0.001$, suggesting that lower $f_{\rm iso}$ values are

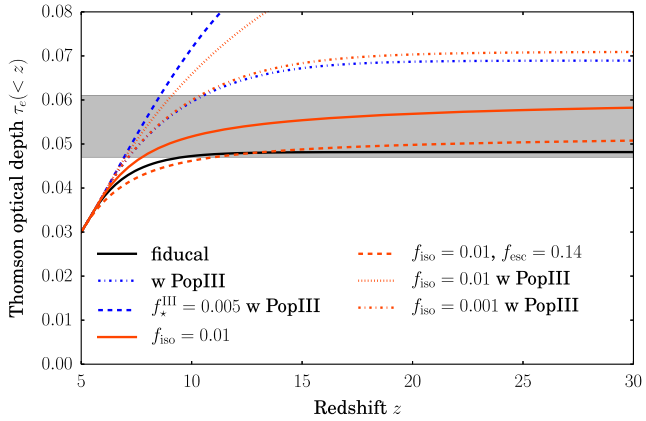

FIG. 6. Thomson optical depth to recombination, τ_e. The gray band shows the *Planck* 2018 constraints on τ_e. The models plotted are the same as in Fig. 5, with the black solid line representing the fiducial case using the stellar efficiency described in Sec. IV and $f_{\rm iso} = 0$, while the orange lines show the contribution for varying $f_{\rm iso}$. The models that are clearly discrepant by the current CMB constraints have either (1) pop-III photoionization and $f_{\rm iso} = 0.01$ or (2) high pop-III star formation efficiency and no isocurvature fluctuations (blue dashed line).

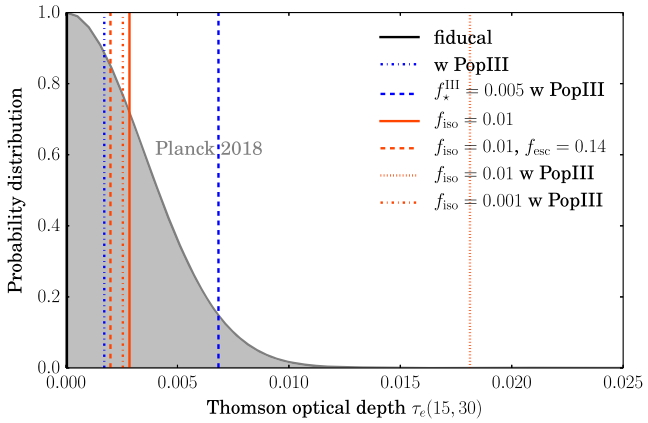

FIG. 7. Thomson optical depth contributed in redshift interval $15 < z < 30$, $\tau_e(15, 30)$, in both models and observations. The gray posterior shows the Gaussian that yields the *Planck* 2018 2σ upper bound of $\tau_e(15, 30) < 0.007$. The vertical lines show $\tau_e(15, 30)$ values for the same models that are plotted in Fig. 6: the fiducial case using the stellar efficiency described in Sec. IV and $f_{\rm iso} = 0$ (black solid line), adding pop-III photoionization rates (blue dot-dashed line), and varying $f_{\rm iso}$ (orange lines with differing line styles). Including even small pop-III star formation efficiencies can result in detectable $\tau_e(15, 30)$ for $f_{\rm iso} = 0.01$.

more degenerate with astrophysical uncertainties of early star formation. Along this line, increasing the star formation efficiency of pop-III stars to $f_{\rm star}^{\rm III} = 0.005$ (0.05) [98] leads to $\tau_e(15, 30) = 0.007$ (0.017) for $f_{\rm iso} = 0$. While tangential to the focus of this paper, this interestingly suggests that *Planck* is already constraining pop-III star efficiencies in some of the range typically used. The limits on the tail of reionization are most constraining for models with an earlier star formation, in particular, if the contribution of pop-III stars is included. Apart from the stellar efficiency, changing the escape fraction of photons from pop-II stellar population [$\tau_e(15, 30) = 0.002$—dashed red line] can also lower the predicted optical depth, making isocurvature models similar to the fiducial adiabatic dark matter model (solid black line). This effect can also lower the optical depth in pop-III models that have slightly higher τ_e compared to the CMB data (green and red dot-dashed lines).

The enhanced contribution to τ_e from the isocurvature fluctuations can be mimicked by astrophysical uncertainties: Similar effects can be observed by keeping $f_{\rm iso} = 0.01$ fixed but switching off the pop-III star formation (solid red line) or switching off isocurvature contribution but adding pop-III photoionization with the stellar efficiency of $f_\star^{\rm III} = 5 \times 10^{-4}$ (green dot-dashed line). However, differences may show up in the tail of the reionization, where the aforementioned two models differ by a factor of ≈ 2 in $\tau_e(15, 30)$. In particular, further increasing pop-III stellar efficiency by another order of magnitude to $f_\star^{\rm III} = 0.05$ results in too much ionization at early times—$\tau_e(15, 30) = 0.018$—which is ruled out by *Planck* CMB constraints. Such a high $\tau_e(15, 30)$ is similar to that for the case with low pop-III stellar efficiency and nonzero $f_{\rm iso}$ (see the red dot-dashed line in Fig. 7). However, the contribution to the ionization fraction comes from $z < 20$ in the case of high pop-III stellar efficiency, while the signal in isocurvature models is dominated by the contribution at $z > 20$.

On the other hand, further increasing the amount of isocurvature power by a factor of 5 ionizes the Universe to 10% early on [$z \sim 29(46)$ for $f_{\rm iso} = 0.05(0.1)$], leading to large values of $\tau_e(15, 30) \sim 0.04(0.09)$. Such models are clearly ruled out by the current CMB data, despite the astrophysical uncertainties. At a high enough level of $f_{\rm iso}$, the statement that such models are excluded by the CMB holds over the range of pop-III efficiencies considered. In our models, this transition happens in the range of $f_{\rm iso} = 0.01$–0.1.

Neglecting pop-III contribution also lowers the effect of isocurvature modes. This occurs because the minimal mass ($M_{\rm min}$) that contributes to the pop-II photoionization rates [Eq. (15)] is typically $\sim 10^8$ $M_\odot$, requiring a large $f_{\rm iso}$ to have an appreciable effect on these mass scales (see Fig. 2). On the other hand, the minimal mass for pop-III photoionization rates ($M_{\rm min}^{\rm III}$) is generally 2 orders of magnitude lower than for pop-II stars ($\sim 10^6$ $M_\odot$) and, thus, more sensitive to smaller values of $f_{\rm iso}$.

Since some contribution from the pop-III star formation is expected, values of $f_{\rm iso}$ of the order of 10^{-2} are excluded with the current measurements already, which corresponds to the ALP mass limit of $m_a > 10^{-18}$ eV. Current and future CMB observations (e.g., CLASS, LiteBIRD) aim to put more stringent constraints on τ_e approaching the cosmic variance limit of $\sigma_\tau = 0.002$ [99,100]. The sensitivity of measurements of the tail of reionization via statistics like $\tau_e(15, 30)$ likely can be improved even more significantly over *Planck* with future missions than this improvement in σ_τ [100], although we expect that measuring even higher redshift contributions like $\tau_e(25, 40)$ would be needed to be able to disentangle astrophysics and improve constraints on $f_{\rm iso}$.

Finally, we note that early ionization (which is likely also associated with x-ray and ultraviolet backgrounds) would shape the high-redshift 21 cm emission signal [101]. The 21 cm signal is potentially sensitive to much lower star formation rate densities via these emissions than the ionizing emissions this section has focused on [77]. Section VI discusses another effect that may be even more constraining for this signal.

VI. CMB RECOMBINATION AND THE DARK AGES THERMAL HISTORY

As illustrated in Fig. 4, the presence of white-noise isocurvature fluctuations leads to the formation of dark matter halos much earlier than in the standard scenario. These early dark matter halos are moving supersonically relative to the gas, with an rms Mach number of ≈ 2 and

with a Maxwellian distribution [102]. Some regions can even be moving hypersonically at $z \gtrsim 500$ (i.e., with relative velocities of $\gtrsim 10\ \mathrm{km\,s^{-1}}$ so that the shocks can ionize the gas). Furthermore, a $10^4\ \mathrm{M}_\odot$ dark matter halo will lose its velocity relative to the dark matter within a Hubble time [103], potentially ionizing and heating the gas in the Universe if enough of these halos are present.

We first investigate the effect of shock ionization on the cosmic microwave background from such hypersonic motion. Even percent-level differences in the global $z \sim 500$ recombination history that result from this ionization could have a detectable effect on the cosmic microwave background [104]. However, while we found that the shocks in a large fraction of the Universe at $z > 500$ would often heat the gas sufficiently for it to start to collisionally ionize, ionization would quickly sap out the thermal energy of the gas, leaving it at insufficient temperatures to collisionally ionize further. We found that, because of this cost to ionization, even the strongest shocks would ionize the gas only to ∼1%. This small ionization, coupled with the fact that (for viable $f_{\rm iso}$) only a fraction of dark matter has collapsed into the $M_h \gtrsim 10^3\ \mathrm{M}_\odot$ at $z \gtrsim 500$ halos that generate significant shocks, results in the recombination history being negligibly affected.

We next turn to the heating imparted by such shocks. If the heating occurs early enough, it could also affect the recombination history, as the recombination rate depends inversely on the temperature. Our calculations suggest that such heating does not occur at early enough times to be relevant for recombination. Another observable is the cosmological 21 cm signal. When the 21 cm signal is in absorption as is anticipated $15 \lesssim z \lesssim 30$, its amplitude is inversely proportional to the gas temperature [101]. We show below that this shock heating could be important for this 21 cm signal.

A simple estimate for the amount of shocking uses that we know how much energy is dissipated into the gas via dynamical friction, a frictional force from the gas that acts to decelerate the supersonically streaming dark matter halos. Namely, halos more massive than $\sim 10^5 - 10^6\ \mathrm{M}_\odot$ should lose all of their relative velocity to the baryons in a Hubble time at $z \sim 20$ [77]. Some of this dynamical energy should go into shocks (and if all of the energy goes into shocks, we would expect to heat the Universe by $\langle \mathcal{M}^2 \rangle \sim 4$). We estimate the effect of shock heating on the thermal history by solving

$$\frac{dT}{dt} = \overbrace{-2HT}^{\text{adiabatic}} + \overbrace{\frac{8\pi^2 x_i T_\gamma^4 \sigma_T (T_\gamma - T)}{45 m_e (1+x_i)}}^{\text{Compton}} - \underbrace{\frac{\mu m_p}{3M_P^4} \langle \zeta_s (v_{b\text{-dm}}) v_{b\text{-dm}}^{-1} \rangle \int_{M_{\rm min}}^{M_{\rm max}} dM_h M_h^2 \frac{dn}{dM_h}}_{\text{shock heating}}, \quad (19)$$

where σ_T is the Thomson cross section, m_p is the mass of hydrogen atom, $v_{b\text{-dm}}$ is the velocity difference between dark matter and baryons, M_h is the halo mass, and $\rho_{\rm dm}$ is the density of dark matter. The Compton cooling term owes to the scattering of CMB photons, which is negligible below redshift $z = 200$. The "shock heating" term in Eq. (19) follows from the power generated from dynamical friction, taking the expression in Ref. [105] but dropping the factor of the Coulomb logarithm. The motivation for dropping this logarithm is that the resulting expression accounts only for gas that intersects within the Bondi-Hoyle radius for accretion ($r_{\rm BH} = 2GM/v_{b\text{-dm}}^2$ [106], and see Ref. [107]), which is the gas whose trajectory would be deflected to the origin (in the absence of pressure) and, hence, is most likely to shock. We conservatively assume the shock heating has efficiency ζ_s at thermalizing its energy, and we take $\zeta_s = 0.1$ motivated by entropy increase calculated in planar shocks with Mach numbers of $\mathcal{M} = 2$. Finally, $M_{\rm max}$ is set to the halo mass whose timescale to lose its energy by dynamical friction is much less than the age of the Universe, as once a halo reaches this mass, it will likely have decelerated and no longer contribute to the heating. We take $10^6\ \mathrm{M}_\odot$ as the maximum mass. The minimum mass is set by where the halo viral radius equals $r_{\rm BH}$, which we find is $M \approx 10^4\ \mathrm{M}_\odot$. It is worth stressing that the shock heating effect is most sensitive to the maximum mass. If we make the maximum mass a factor of 10 smaller ($10^5\ \mathrm{M}_\odot$), the temperature difference will be about 3 times smaller in Fig. 8, which we think reflects the level of uncertainty.

Our simple estimates show that the shock heating effects from axion halos starts to become significant around redshift $z = 20$ as shown in Fig. 8 for $f_{\rm iso} \gtrsim 10^{-4}$. Models predict a global 21 cm absorption feature at ∼80 MHz, corresponding to absorption at $z \sim 15$–20 [101], the same

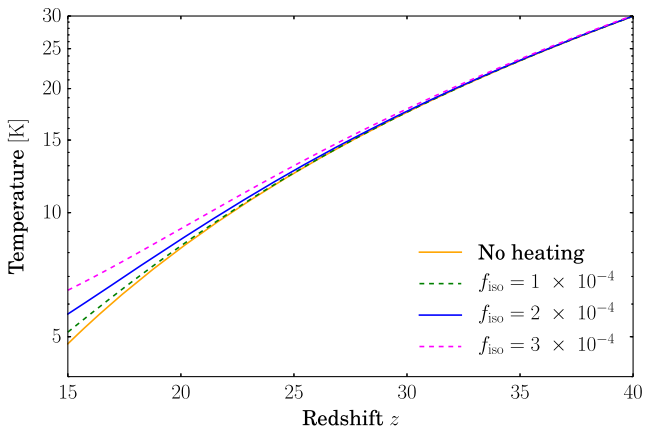

FIG. 8. Rough estimates for the evolution of gas temperature at different $f_{\rm iso}$, with larger $f_{\rm iso}$ leading to more shock heating. Our estimates suggest that $f_{\rm iso} \sim 10^{-4}$ lead to percent-level or greater shock heating. Percent-level heating would manifest in a qualitatively different high-redshift 21 cm signal, with significant baryon acoustic oscillation peaks.

signal purported to be detected by EDGES [108]. This absorption dip is inversely proportional to the gas temperature. Thus, a detection of the full amplitude of this dip should at a minimum be used to discern shock heating at the $\mathcal{O}(1)$ level, requiring for us $f_{\rm iso} \sim 5 \times 10^{-4}$. Such heating would be hard to disentangle from x-ray heating from the first supernovae and black holes [101]. However, efforts to detect fluctuations in the 21 cm have a potentially smoking gun signal for this heating. Since a change in temperature is tied to the relative velocity between the baryons and the dark matter ($v_{b\text{-dm}}$), and this relative velocity is modulated by the acoustic physics in the early Universe, any heating could result in large acoustic oscillations in the signal. McQuinn and O'Leary [77] showed that even just ∼3% changes in the temperature that are tied to $v_{b\text{-dm}}$ would lead to order-unity acoustic features in the 21 cm signal at $k \sim 0.1$ Mpc^{-1} [77], qualitatively changing the 21 cm signal. Our estimates in Fig. 8 suggest that heating at the few percent level occurs for $f_{\rm iso} \gtrsim 10^{-4}$, although we illustrate the rough constraint in Fig. 1 at $f_{\rm iso} = 3 \times 10^{-4}$. These acoustic features are quite distinct from the smoother continuum of fluctuations from the extra star formation would create, which were referenced as a potential observable in Sec. V.

VII. CONCLUSIONS

One possible candidate for the dark matter is that it is an ultralight scalar field that is generated in the early Universe in a similar manner to that for the QCD axion, making it an ALP. Most previous studies have concentrated on how the ultralight ALP's quantum pressure suppresses the small-scale growth of the adiabatic fluctuations from inflation or on how its relaxation can lead to solitonic cores [109–111]. However, if the symmetry breaking that establishes the ALP occurs after inflation ends, this leads to white isocurvature fluctuations in the ALP energy density. The parameter space where the postinflationary scenario can occur are for symmetry-breaking scales of $10^{13} - 10^{16}$ GeV for the particle mass ranges that are probed by the large-scale structure observables considered here ($m_A \sim 10^{-13} - 10^{-20}$ eV). The higher values for the symmetry-breaking scale (and lower values for the mass) push against limits from searches for inflationary B modes. This paper focused on how these isocurvature fluctuations could influence various observations of early structure formation.

Figure 1 summarizes our resulting constraints on the ALP mass m_a and isocurvature fluctuation amplitude $f_{\rm iso}$—defined in the traditional manner as their ratio with adiabatic fluctuations at a wave number of 0.05 Mpc^{-1} (but we also report constraints in terms of the more natural nonlinear wave number $k_{\rm nl}$). The solid lines show the relation between the axion mass and $f_{\rm iso}$. Different colors represent different parameterizations of the evolution of the axion mass with temperature after it commences oscillations. The simplest model, and also most conservative in terms of mass constraints, is the $n = 0$ case, where the mass was set at early times. For an ALP coupled to an asymptotically free sector (in analogy to the QCD axion), leading to a mass that increases in size until the cosmic temperature falls below the sector's confinement scale, the value of n is nonzero (with $n = 4$ approximating the evolution of the QCD axion). As n increases above 4, the sensitivity of our results to n becomes weak.

The cosmological observables presented in this paper are sensitive to different axion masses m_A or, equivalently, different levels of $f_{\rm iso}$, with the smaller scale the observable is sensitive to, the stronger the constraint. Our strongest present constraint comes from the Lyα forest power spectrum measurements at high redshifts (orange dashed line). The lower bound on the ALP mass from the Lyα forest is $m_A > 2 \times 10^{-17}$ eV for $n = 0$ (and $m_A > 10^{-13}$ eV for $n = 4$). Apart from being currently the most constraining bound, the Lyα analysis is also the least affected by uncertainties in the astrophysics of the existing probes we investigated.

Another potential probe is high-redshift galaxy observations. We find that only for m_A already ruled out by the Lyα forest is the observed luminosity function qualitatively changed in a manner that could potentially be disentangled from more mundane astrophysical explanations. However, smaller mass (and higher redshift) galaxies than can be observed directly are more substantially boosted by isocurvature fluctuations. Such diminutive galaxies may be observable via their effect on the ionized fraction evolution during the reionization epoch. We find that a particularly interesting observable is the CMB, which is sensitive to the high-redshift tail of reionization. This tail can be substantially more extended in models with white isocurvature fluctuations. While we find that the ionization fraction in models where galaxies form via the traditional route (in halos massive enough that the gas can cool atomically) show only qualitatively different trends for m_A already ruled out by the forest, models that include pop-III stars (even for much lower efficiencies for their formation than is commonly assumed) could lead to a small residual ionization to extremely high redshifts. Thus, future CMB efforts could potentially probe a m_A range similar to that of the Lyα forest.

Finally, the shock heating of the gas due to supersonically moving axion minihalos during the cosmic Dark Ages and cosmic dawn could lead to even stronger constraints, potentially excluding ALP masses of $m_A < 10^{-16}$ eV for $n = 0$. This shocking would suppress the depth of the absorption trough in the global 21 cm signal (as probed by, e.g., EDGES and PRIZM). The caveat is that x-ray heating could have a similar effect [112,113]. However, even percent-level changes in the mean temperature from shock heating will manifest in distinct baryon acoustic oscillation features in the 21 cm brightness temperature fluctuations that trace the relative baryon-dark matter velocity field. These oscillations are potentially a smoking gun of shock

heating from a dramatic enhancement in the number of minihalos.

Some low-redshift small-scale structure probes could complement the probes discussed here. First, local observations of Milky Way tidal streams could lead to detection of small subhalos in the mass range 10^8–10^5 $M_\odot$ [114,115], with some uncertainty in whether the lowest values of 10^5 $M_\odot$ can be disentangled from astrophysical uncertainties, as encounters with these subhalos open up gaps in these streams. This places the sensitivity of the galactic streams somewhere in the range of isocurvature amplitudes of $f_{\rm iso} = 0.001$–0.01, potentially pushing the constraints lower than the current Lyα bound and comparable to our most optimistic reionization constraints.

In addition, Miralda-Escudé [116] recently showed that the microlensing caustics of stars on a cluster macrolens could even be sensitive to the minute value of $M_{H(m_A)}$ for the QCD axion of $\sim 10^{-12}$ $M_\odot$, where $M_{H(m_A)}$ is the mass within the horizon at $T_{\rm osc}$. In particular, these microlensing caustics are perturbed by these axion structures, deviating from the smooth profile otherwise expected. This constraint can also be translated to our scenario. Miralda-Escudé [116] showed this method is sensitive to $10^{-13} < M_{H(m_A)} < 10^{-6}$ $M_\odot$, which translates to the bounds on the ALP mass of $10^{-15} < m_A < 10^{-11}$ eV for $n = 0$ ($10^{-11} - 10^{-6}$ eV for $n = 4$). Since the sensitivity falls off on both sides of the ALP mass range, this makes the microlensing of stars complementary to the signatures of early structure formation considered in this paper. Future observations with HST or JWST should be able to push forward this exciting science [117,118].

Lastly, a postinflation ALP may affect the properties of black holes. Studies of black hole superradiance [119–122]—the gravitational production of an ALP halo from the free energy in black hole spin—have excluded the existence of ALPs with $10^{-14} < m_A < 10^{-11}$ eV from measurements of finite stellar black hole spins. The measurements of supermassive black hole spin can potentially exclude a wide mass range $m_A < 10^{-16}$ eV [121] but inferring the black hole masses over a broad mass range. The bounds from superradiance are also valid only in the limit of $f_A > 10^{14}$ GeV and no self-interaction [121]. Furthermore, the earlier structure formation sourced by a postinflation ALP could potentially produce the seeds that grow into the highest-mass black holes, ameliorating somewhat the difficulty in having sufficient time for these seeds to grow into the highest-redshift quasars (e.g., [123]).

ACKNOWLEDGMENTS

We thank Akshay Ghalsasi for helpful conversations and Erik Anson for running tests of the universality of the mass function in cosmologies near our white case. V. I. and M. M. thank U.S. National Science Foundation Grant No. AST-1514734, and M. M. and H. X. thank the University of Washington Royalty Research Grant program. H. X. is also supported in part by the U.S. Department of Energy, under Award No. DE-SC0011637. V. I. acknowledges support by the Kavli Foundation.

[1] G. Jungman, M. Kamionkowski, and K. Griest, Phys. Rep. **267**, 195 (1996).
[2] S. Weinberg, Phys. Rev. Lett. **40**, 223 (1978).
[3] F. Wilczek, Phys. Rev. Lett. **40**, 279 (1978).
[4] E. W. Kolb and M. S. Turner, *The Early Universe*, Vol. 69 (Addison-Wesley, Redwood City, CA, 1990).
[5] J. Preskill, M. B. Wise, and F. Wilczek, Phys. Lett. **120B**, 127 (1983).
[6] L. F. Abbott and P. Sikivie, Phys. Lett. **120B**, 133 (1983).
[7] M. Dine and W. Fischler, Phys. Lett. **120B**, 137 (1983).
[8] D. J. E. Marsh, Phys. Rep. **643**, 1 (2016).
[9] L. Hui, J. P. Ostriker, S. Tremaine, and E. Witten, Phys. Rev. D **95**, 043541 (2017).
[10] P. Svrcek and E. Witten, J. High Energy Phys. 06 (2006) 051.
[11] Y. Akrami, *et al.* (Planck Collaboration), arXiv:1807.06211.
[12] The maximum temperature is larger (in some models by orders of magnitude) than the reheat temperature (e.g., [13]).
[13] E. W. Kolb, A. Notari, and A. Riotto, Phys. Rev. D **68**, 123505 (2003).
[14] I. I. Tkachev, Phys. Lett. B **376**, 35 (1996).
[15] L. Kofman, A. Linde, and A. A. Starobinsky, Phys. Rev. Lett. **76**, 1011 (1996).
[16] A. Arvanitaki, S. Dimopoulos, S. Dubovsky, N. Kaloper, and J. March-Russell, Phys. Rev. D **81**, 123530 (2010).
[17] In this strongly interacting "axiverse" scenario, any postinflation ALP likely cannot have multiple nondegenerate vacua to avoid a domain wall catastrophe. Thus, the ALPs with nondegenerate vacua would come into existence before inflation and have a small misalignment angle coherent over the cosmological volume so that they do not overclose the Universe, which perhaps could occur because of the anthropic principle [18]. For our results to apply, of course, the ALPs that dominate the dark matter density would have to come into existence after inflation.
[18] F. Wilczek, arXiv:hep-ph/0408167.
[19] T. Kibble, Phys. Rep. **67**, 183 (1980).
[20] C. J. Hogan and M. J. Rees, Phys. Lett. B **205**, 228 (1988).
[21] J. Preskill, M. B. Wise, and F. Wilczek, Phys. Lett. **120B**, 127 (1983).

[22] G. Efstathiou and J. R. Bond, Mon. Not. R. Astron. Soc. **218**, 103 (1986).
[23] A. Vaquero, J. Redondo, and J. Stadler, J. Cosmol. Astropart. Phys. 04 (2019) 012.
[24] E. W. Kolb and I. I. Tkachev, Phys. Rev. D **49**, 5040 (1994).
[25] L. Dai and J. Miralda-Escudé, Astron. J. **159**, 49 (2020).
[26] D. J. E. Marsh, D. Grin, R. Hlozek, and P. G. Ferreira, Phys. Rev. D **87**, 121701 (2013).
[27] M. Feix, J. Frank, A. Pargner, R. Reischke, B. M. Schäfer, and T. Schwetz, J. Cosmol. Astropart. Phys. 05 (2019) 021.
[28] P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett *et al.* (Planck Collaboration), Astron. Astrophys. **594**, A13 (2016).
[29] R. K. Sheth and G. Tormen, Mon. Not. R. Astron. Soc. **329**, 61 (2002).
[30] J. S. Bagla, N. Khandai, and G. Kulkarni, arXiv:0908.2702.
[31] S. Weinberg, Phys. Rev. Lett. **40**, 223 (1978).
[32] F. Wilczek, Phys. Rev. Lett. **40**, 279 (1978).
[33] A. Vaquero, J. Redondo, and J. Stadler, J. Cosmol. Astropart. Phys. 04 (2019) 012.
[34] That the isocurvature transfer function limits to unity at high k is true for the dark matter-ALP transfer function, whereas the total matter transfer function will be lower due to the effects of Jeans smoothing on the baryons.
[35] A. Lewis and S. Bridle, Phys. Rev. D **66**, 103511 (2002).
[36] A. A. Meiksin, Rev. Mod. Phys. **81**, 1405 (2009).
[37] M. McQuinn, Annu. Rev. Astron. Astrophys. **54**, 313 (2016).
[38] V. Iršič and M. McQuinn, J. Cosmol. Astropart. Phys. 04 (2018) 026.
[39] R. Cen, J. Miralda-Escudé, J. P. Ostriker, and M. Rauch, Astrophys. J. Lett. **437**, L9 (1994).
[40] J. Miralda-Escudé, R. Cen, J. P. Ostriker, and M. Rauch, Astrophys. J. **471**, 582 (1996).
[41] L. Hernquist, N. Katz, D. H. Weinberg, and J. Miralda-Escudé, Astrophys. J. Lett. **457**, L51 (1996).
[42] U. Seljak, A. Makarov, P. McDonald, and H. Trac, Phys. Rev. Lett. **97**, 191303 (2006).
[43] M. Viel, J. Lesgourgues, M. G. Haehnelt, S. Matarrese, and A. Riotto, Phys. Rev. D **71**, 063534 (2005).
[44] V. Iršič, M. Viel, M. G. Haehnelt, J. S. Bolton, and G. D. Becker, Phys. Rev. Lett. **119**, 031302 (2017).
[45] T. Kobayashi, R. Murgia, A. De Simone, V. Iršič, and M. Viel, Phys. Rev. D **96**, 123514 (2017).
[46] R. Murgia, G. Scelfo, M. Viel, and A. Raccanelli, Phys. Rev. Lett. **123**, 071102 (2019).
[47] N. Afshordi, P. McDonald, and D. N. Spergel, Astrophys. J. Lett. **594**, L71 (2003).
[48] J. S. Bullock and M. Boylan-Kolchin, Annu. Rev. Astron. Astrophys. **55**, 343 (2017).
[49] R. Barkana, Z. Haiman, and J. P. Ostriker, Astrophys. J. **558**, 482 (2001).
[50] F. Pacucci, A. Mesinger, and Z. Haiman, Mon. Not. R. Astron. Soc. **435**, L53 (2013).
[51] We have checked that the results are not sensitive to the choice of the mass function by also investigating a mass function specifically calibrated to simulations at high redshift [52].
[52] H. Trac, R. Cen, and P. Mansfield, Astrophys. J. **813**, 54 (2015).
[53] Z. Lukić, K. Heitmann, S. Habib, S. Bashinsky, and P. M. Ricker, Astrophys. J. **671**, 1160 (2007).
[54] S. R. Furlanetto, J. Mirocha, R. H. Mebane, and G. Sun, Mon. Not. R. Astron. Soc. **472**, 1576 (2017).
[55] J. McBride, O. Fakhouri, and C.-P. Ma, Mon. Not. R. Astron. Soc. **398**, 1858 (2009).
[56] C. A. Correa, J. S. B. Wyithe, J. Schaye, and A. R. Duffy, Mon. Not. R. Astron. Soc. **450**, 1514 (2015).
[57] R. J. McLure, J. S. Dunlop, R. A. A. Bowler, E. Curtis-Lake, M. Schenker, R. S. Ellis, B. E. Robertson, A. M. Koekemoer, A. B. Rogers, Y. Ono *et al.*, Mon. Not. R. Astron. Soc. **432**, 2696 (2013).
[58] R. A. A. Bowler, J. S. Dunlop, R. J. McLure, and D. J. McLeod, Mon. Not. R. Astron. Soc. **466**, 3612 (2017).
[59] R. J. Bouwens, G. D. Illingworth, P. A. Oesch, M. Trenti, I. Labbé, L. Bradley, M. Carollo, P. G. van Dokkum, V. Gonzalez, B. Holwerda *et al.*, Astrophys. J. **803**, 34 (2015).
[60] R. J. Bouwens, P. A. Oesch, I. Labbé, G. D. Illingworth, G. G. Fazio, D. Coe, B. Holwerda, R. Smit, M. Stefanon, P. G. van Dokkum *et al.*, Astrophys. J. **830**, 67 (2016).
[61] R. J. Bouwens, P. A. Oesch, G. D. Illingworth, R. S. Ellis, and M. Stefanon, Astrophys. J. **843**, 129 (2017).
[62] R. Smit, R. J. Bouwens, M. Franx, G. D. Illingworth, I. Labbé, P. A. Oesch, and P. G. van Dokkum, Astrophys. J. **756**, 14 (2012).
[63] This effect may weaken our constraints from the galaxy luminosity function if lower mass galaxies are substantially dust absorbed.
[64] F. Haardt and P. Madau, Astrophys. J. **746**, 125 (2012).
[65] W. H. Press and P. Schechter, Astrophys. J. **187**, 425 (1974).
[66] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophys. J. **379**, 440 (1991).
[67] The full dependence on redshift and virial temperature for the adiabatic case is roughly $f_{\rm coll}(>T_{\rm vir}) = {\rm erfc}[2.04 \times Z_{10}\{\ln(4.6 \times Z_{10} T_{{\rm vir},4}^{-1})\}^{-1/2}]$, but the logarithmic dependence adds only a small correction to the redshift evolution.
[68] G. Sun and S. R. Furlanetto, Mon. Not. R. Astron. Soc. **460**, 417 (2016).
[69] J. M. Shull, A. Harness, M. Trenti, and B. D. Smith, Astrophys. J. **747**, 100 (2012).
[70] E. Visbal, Z. Haiman, and G. L. Bryan, Mon. Not. R. Astron. Soc. **453**, 4456 (2015).
[71] V. Bromm, R. P. Kudritzki, and A. Loeb, Astrophys. J. **552**, 464 (2001).
[72] M. Trenti and M. Stiavelli, Astrophys. J. **694**, 879 (2009).
[73] E. Visbal, Z. Haiman, and G. L. Bryan, Mon. Not. R. Astron. Soc. **475**, 5246 (2018).
[74] The star formation rate density in our model peaks at around 2×10^{-4} $M_\odot\,{\rm yr}^{-1}\,{\rm Mpc}^{-3}$ at a redshift of 15 and falls off towards higher redshifts (e.g., 10^{-6} $M_\odot\,{\rm yr}^{-1}\,{\rm Mpc}^{-3}$ at a redshift of 35), behavior quantitatively very similar to that found in Ref. [70]. This is true despite different star formation efficiency assumed in our model compared to Ref. [70], because the minimum mass in which molecular cooling can lead to pop-III star formation is lower in our model, compared to that of Ref. [70]. In Ref. [70], the

numeric value of the minimum mass is obtained from CDM simulations and corresponds to roughly $T_{\rm vir} = 1000$ K. See Eq. (17), as the minimum does not just set the absolute minimum but also what halos are affected by the Lyman-Werner background.
[75] Z. Haiman, M. J. Rees, and A. Loeb, Astrophys. J. **476**, 458 (1997).
[76] Z. Haiman, T. Abel, and M. J. Rees, Astrophys. J. **534**, 11 (2000).
[77] M. McQuinn and R. M. O'Leary, Astrophys. J. **760**, 3 (2012).
[78] R. H. Mebane, J. Mirocha, and S. R. Furlanetto, Mon. Not. R. Astron. Soc. **479**, 4544 (2018).
[79] M. E. Machacek, G. L. Bryan, and T. Abel, Astrophys. J. **548**, 509 (2001).
[80] J. H. Wise and T. Abel, Astrophys. J. **671**, 1559 (2007).
[81] M. Ricotti, N. Y. Gnedin, and J. M. Shull, Astrophys. J. **560**, 580 (2001).
[82] J. L. Johnson, T. H. Greif, and V. Bromm, Astrophys. J. **665**, 85 (2007).
[83] V. Miranda, A. Lidz, C. H. Heinrich, and W. Hu, Mon. Not. R. Astron. Soc. **467**, 4050 (2017).
[84] I. D. McGreer, A. Mesinger, and V. D'Odorico, Mon. Not. R. Astron. Soc. **447**, 499 (2015).
[85] C. A. Mason, T. Treu, M. Dijkstra, A. Mesinger, M. Trenti, L. Pentericci, S. de Barros, and E. Vanzella, Astrophys. J. **856**, 2 (2018).
[86] A. Hoag, M. Bradač, K. Huang, C. Mason, T. Treu, K. B. Schmidt, M. Trenti, V. Strait, B. C. Lemaux, E. Q. Finney *et al.*, Astrophys. J. **878**, 12 (2019).
[87] C. A. Mason, A. Fontana, T. Treu, K. B. Schmidt, A. Hoag, L. Abramson, R. Amorin, M. Bradač, L. Guaita, T. Jones *et al.*, Mon. Not. R. Astron. Soc. **485**, 3947 (2019).
[88] B. Greig, A. Mesinger, I. D. McGreer, Z. Haiman, and R. A. Simcoe, Mon. Not. R. Astron. Soc. **466**, 1814 (2017).
[89] B. Greig, A. Mesinger, and E. Bañados, Mon. Not. R. Astron. Soc. **484**, 5094 (2019).
[90] F. B. Davies, J. F. Hennawi, E. Bañados, Z. Lukić, R. Decarli, X. Fan, E. P. Farina, C. Mazzucchelli, H.-W. Rix, B. P. Venemans *et al.*, Astrophys. J. **864**, 142 (2018).
[91] A. Lawrence, S. J. Warren, O. Almaini, A. C. Edge, N. C. Hambly, R. F. Jameson, P. Lucas, M. Casali, A. Adamson, S. Dye *et al.*, Mon. Not. R. Astron. Soc. **379**, 1599 (2007).
[92] A. Edge, W. Sutherland, K. Kuijken, S. Driver, R. McMahon, S. Eales, and J. P. Emerson, Messenger **154**, 32 (2013).
[93] R. G. McMahon, M. Banerji, E. Gonzalez, S. E. Koposov, V. J. Bejar, N. Lodieu, R. Rebolo (VHS Collaboration), Messenger **154**, 35 (2013).
[94] S. Dye, A. Lawrence, M. A. Read, X. Fan, T. Kerr, W. Varricatt, K. E. Furnell, A. C. Edge, M. Irwin, N. Hambly *et al.*, Mon. Not. R. Astron. Soc. **473**, 5113 (2018).
[95] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo *et al.* (Planck Collaboration), arXiv:1807.06209.
[96] W. Hu and G. P. Holder, Phys. Rev. D **68**, 023001 (2003).
[97] C. H. Heinrich, V. Miranda, and W. Hu, Phys. Rev. D **95**, 023513 (2017).
[98] This is the efficiency one expects from assuming that each 10^5 $M_\odot$ halo hosts one (ten) 100 $M_\odot$ stars, and it further takes the efficiency to scale with halo mass.
[99] E. Di Valentino, T. Brinckmann, M. Gerbino, V. Poulin, F. R. Bouchet, J. Lesgourgues, A. Melchiorri, J. Chluba, S. Clesse, J. Delabrouille *et al.*, J. Cosmol. Astropart. Phys. 04 (2018) 017.
[100] D. J. Watts, G. A. Addison, C. L. Bennett, and J. L. Weiland, Astrophys. J. **889**, 130 (2020).
[101] S. R. Furlanetto, S. P. Oh, and F. H. Briggs, Phys. Rep. **433**, 181 (2006).
[102] D. Tseliakhovich and C. Hirata, Phys. Rev. D **82**, 083520 (2010).
[103] R. M. O'Leary and M. McQuinn, Astrophys. J. **760**, 4 (2012).
[104] T. R. Slatyer, N. Padmanabhan, and D. P. Finkbeiner, Phys. Rev. D **80**, 043526 (2009).
[105] E. C. Ostriker, Astrophys. J. **513**, 252 (1999).
[106] H. Bondi and F. Hoyle, Mon. Not. R. Astron. Soc. **104**, 273 (1944).
[107] Our expression for the heating power from each halo is equal to the cross section for Bondi-Hoyle accretion times the kinetic energy density of the accreted gas times the velocity offset.
[108] J. D. Bowman, A. E. E. Rogers, R. A. Monsalve, T. J. Mozdzen, and N. Mahesh, Nature (London) **555**, 67 (2018).
[109] H.-Y. Schive, T. Chiueh, and T. Broadhurst, Nat. Phys. **10**, 496 (2014).
[110] J. Veltmaat, J. C. Niemeyer, and B. Schwabe, Phys. Rev. D **98**, 043509 (2018).
[111] P. Mocz, A. Fialkov, M. Vogelsberger, F. Becerra, M. A. Amin, S. Bose, M. Boylan-Kolchin, P.-H. Chavanis, L. Hernquist, L. Lancaster *et al.*, Phys. Rev. Lett. **123**, 141301 (2019).
[112] R. Barkana, Nature (London) **555**, 71 (2018).
[113] A. Fialkov and R. Barkana, Mon. Not. R. Astron. Soc. **486**, 1763 (2019).
[114] J. Bovy, D. Erkal, and J. L. Sanders, Mon. Not. R. Astron. Soc. **466**, 628 (2017).
[115] A. Bonaca, D. W. Hogg, A. M. Price-Whelan, and C. Conroy, Astrophys. J. **880**, 38 (2019).
[116] L. Dai and J. Miralda-Escudé, arXiv:1908.01773.
[117] W. Chen, P. L. Kelly, J. M. Diego, M. Oguri, L. L. R. Williams, A. Zitrin, T. L. Treu, N. Smith, T. J. Broadhurst, N. Kaiser *et al.*, Astrophys. J. **881**, 8 (2019).
[118] A. A. Kaurov, L. Dai, T. Venumadhav, J. Miralda-Escudé, and B. Frye, Astrophys. J. **880**, 58 (2019).
[119] A. Arvanitaki, M. Baryakhtar, and X. Huang, Phys. Rev. D **91**, 084011 (2015).
[120] M. Baryakhtar, R. Lasenby, and M. Teo, Phys. Rev. D **96**, 035019 (2017).
[121] M. J. Stott and D. J. E. Marsh, Phys. Rev. D **98**, 083006 (2018).
[122] H. Davoudiasl and P. B. Denton, Phys. Rev. Lett. **123**, 021102 (2019).
[123] M. A. Latif and A. Ferrara, Pub. Astron. Soc. Aust. **33**, e051 (2016).