How new physics affects primordial neutrinos decoupling: direct simulation Monte Carlo approach

Ovchynnikov, M.; Syvolap, V.

How new physics affects primordial neutrinos decoupling: Direct simulation Monte Carlo approach

Maksym Ovchynnikov[1,2,*] and Vsevolod Syvolap[3,†]

[1]*Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland*
[2]*Institut für Astroteilchen Physik, Karlsruher Institut für Technologie (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*
[3]*Instituut-Lorentz, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands*

(Received 24 September 2024; accepted 13 February 2025; published 12 March 2025)

Cosmological observations from big bang nucleosynthesis and the cosmic microwave background (CMB) offer crucial insights into the Early Universe, enabling us to trace its evolution back to lifetimes as short as 0.01 s. Upcoming CMB spectrum measurements will achieve unprecedented precision, allowing for more accurate extraction of information about the primordial neutrinos. This provides an opportunity to test whether their properties align with the predictions of the standard cosmological model or indicate the presence of new physics that influenced the evolution of the MeV-temperature plasma. A key component in understanding how new physics may have affected primordial neutrinos is solving the neutrino Boltzmann equation. In this paper, we address this question by developing a novel approach—neutrino direct simulation Monte Carlo (DSMC). We discuss it in depth, highlighting its model independence, transparency, and computational efficiency—features that current state-of-the-art methods lack. Then, we introduce a proof-of-concept implementation of the neutrino DSMC and apply it to several toy scenarios, showcasing key aspects of the primordial plasma's evolution in the presence of new physics.

DOI: 10.1103/PhysRevD.111.063527

I. INTRODUCTION

Primordial neutrinos are an important messenger from the Early Universe, bringing us information about the state of the Universe at times as early as $t \ll 1$ s. Their direct detection is significantly more challenging than that of primordial photons due to their tiny interaction cross section, which is governed by weak interactions. However, numerously populating the primordial plasma, they affected a number of cosmological observables. It makes it possible to indirectly extract information about their properties from precise cosmic measurements. In particular, they contribute to the number of ultrarelativistic (UR) degrees of freedom,

$$N_{\rm eff} = \frac{8}{7}\left(\frac{11}{4}\right)^{\frac{4}{3}}\frac{\rho_{\rm UR}-\rho_\gamma}{\rho_\gamma}, \tag{1}$$

where $\rho_{\rm UR}$ is the energy density of the ultrarelativistic species at the moment of cosmic microwave background (CMB) formation, and ρ_γ is the energy density of photons. This quantity influences the CMB and may be extracted from its measurements.

It is not only the total neutrino energy density that is important. Another essential property is the shape of the neutrino energy distribution function. It handles the neutron-to-proton conversion at MeV temperatures, which determines the onset of big bang nucleosynthesis (BBN), as well as affects baryon acoustic oscillations (BAO) [1,2]. The shape of the distribution may significantly modify the cosmological neutrino mass bound [3].

Assuming the standard cosmological history, based on the ΛCDM model, $N_{\rm eff}$ is fully represented by the neutrino, and its value is 3.043–3.044 [4–10]. The shape of the neutrino distribution is very close to the Fermi-Dirac distribution, with tiny distortions in the high-energy tail. Finally, there is no asymmetry between neutrinos and antineutrinos. Altogether, it serves as an input to the standard big bang nucleosynthesis model, which predicts the helium abundance $Y_p = 0.247 \pm 0.00017$ (see, e.g., [11,12]). These numbers agree with the current BBN and CMB observations. In particular, the measurements performed by the Planck collaboration [13] constrain $N_{\rm eff} = 2.99^{+0.34}_{-0.33}$ at 95% CL, whereas the primordial helium abundance measurements are in a range 0.233–0.2573, obtained by combining the observations from the works [14–21].

*Contact author: maksym.ovchynnikov@cern.ch
†Contact author: sivolapseva@gmail.com

However, the uncertainty window of these observations leaves room for sizable deviations from standard neutrino properties that may potentially originate from the presence of new physics at temperatures $T_{\rm EM} \lesssim 5$ MeV, when neutrinos start decoupling.[1] Examples of such scenarios include the presence of nonstandard neutrino interactions [22–25], a lepton asymmetry in the neutrino sector [26–32], a change in the expansion dynamics of the Universe, and the injection of nonthermal neutrinos by hypothetical long-lived particles, or LLPs [33–42]. The accuracy of the CMB measurements will be significantly improved with the future observations with Simons Observatory [43] (which has started collecting the data on June 2024) and CMB-S4 mission [44]. They will be able to measure $N_{\rm eff}$ with a percent precision, thus providing a unique potential to shed light on properties of the new physics or constrain it in case of the absence of deviations from ΛCDM.

Under certain approximations of neutrino oscillations, understanding the impact of the new physics effects on the neutrino properties requires solving the Boltzmann equation on the neutrino distribution function f_{ν_α}:

$$\frac{\partial f_{\nu_\alpha}}{\partial t} - pH\frac{\partial f_{\nu_\alpha}}{\partial p} = \mathcal{I}_{{\rm coll},\alpha}[f_{\nu_\alpha}, p]. \tag{2}$$

Here, p is neutrinos' momentum, H is the Hubble factor accounting for the expansion of the Universe, and $\mathcal{I}_{{\rm coll},\alpha}$ is the collision integral that takes care of the microscopic of the thermalization.

The main approach considered in literature is to reduce the integration inside $\mathcal{I}_{\rm coll}$ analytically as much as possible and convert the complex integrodifferential equation (2) into a system of the ordinary differential equations by discretizing the grid of the comoving momenta (see the pioneering work [45] as well as later realizations [7,46,47], and references therein). The method has also been used to study some well-motivated scenarios with LLPs such as heavy neutral leptons (HNLs) [33,35,39–42] and particles in late reheating scenarios [34,48,49].

However, several problems exist with this approach. First, it has a limited range of applicability, requiring analytic matrix elements for the processes and high reducibility of the dimensionality of the integration in $\mathcal{I}_{{\rm coll},\alpha}$. Second, even within the case studies, its computational complexity quickly grows if high-energy neutrinos are present in the system. For instance, depending on the grid density, solving the Boltzmann equation under the presence of HNLs with masses just $\simeq$200 MeV (injecting neutrinos with energies up to 100 MeV) may take days [39].

In addition, the method itself is very complex. The analytic reduction of the collision integral is highly nontrivial, the comoving grid density has to be adjusted to the model's parameters, and solver stability must be carefully verified. An indirect consequence of this is that there is the existing discrepancy between the predictions of various neutrino Boltzmann codes for the behavior of $N_{\rm eff}$ in the presence of the injection of high-energy neutrinos with energies well exceeding the plasma temperature. While some studies predict that injection of such neutrinos would increase $N_{\rm eff}$ [33,41], the others show the opposite [35,40,42].

In this paper, we address these issues by developing proof of principle of a novel approach to solving the neutrino Boltzmann equation based on the so-called direct simulation Monte Carlo (DSMC) [50–53]. Its basis is the numerical particle representation of the Boltzmann equation: one starts with a large number of particles obeying some initial condition in momentum and spatial spaces and then directly simulates their interactions to study the equilibration. Because of the straightforwardness of the method, DSMC directly calculates the linear functionals, e.g., the number and energy densities, velocities, etc., without any simplifications. The existing case studies describe the implementations of DSMC that efficiently simulate collisions of a number of particles as large as 10^8 [54,55]. As we will see, the simplicity of the scheme describing the interactions and the absence of momentum binning automatically release the DSMC approach from most of the problems described above.

This work also serves as the companion to the Letter [56], which presents a summary of the results.

The paper is organized as follows. In Sec. II, we review the properties of the primordial plasma around the neutrino decoupling, considering both the standard cosmological scenario and setups with new physics. Section III is devoted to a discussion on the existing approaches to solve the neutrino Boltzmann equation. In Sec. IV, we describe the basics of the DSMC approach and, in particular, why it may be well applicable to studying the dynamics of primordial neutrinos. Section V discusses the necessary modifications to the DSMC simulation required to study the primordial MeV plasma, and how they can be implemented. In Sec. VI, we present our proof-of-principle realization of the approach and different cross-checks we performed to validate it against well-defined scenarios. In Sec. VII, we apply the developed approach to a few case studies simplifying various physics setups, highlighting the variety of the applicability of the neutrino DSMC and the importance of using full Boltzmann equations. Finally, in Sec. VIII, we make conclusions.

II. PRIMORDIAL PLASMA AT MEV TEMPERATURES

As we discussed in the Introduction, throughout this study, we mainly focus on the temperature domain 1 MeV $\lesssim T_{\rm EM} \lesssim$ 5 MeV, where neutrinos are already partially decoupled at $T \simeq 5$ MeV [39], but still interact significantly with the electromagnetic (EM) particles and

[1]Here and below, $T_{\rm EM}$ denotes the temperature of the electromagnetic plasma.

themselves. This necessitates a detailed understanding of the dynamics of these interactions.

At these temperatures, the primordial plasma consists of light particles—neutrinos $\nu, \bar{\nu}$, electromagnetically (EM) interacting light particles (electrons e^-, positrons e^+, and photons γ), as well as baryons $B = p$, n. The thermal population of other particles, such as muons, τ leptons, mesons, and excited baryon states, can be safely neglected, as they are too heavy to be abundantly present at this epoch.

The homogeneous and isotropic Universe expands with the rate $H(t) = \dot{a}(t)/a(t)$, where $a(t)$ is the scale factor, and H is the Hubble parameter. Assuming spatial flatness and neglecting the dark energy contribution, we get

$$H(t) = \frac{1}{M_{\rm Pl}}\sqrt{\frac{8\pi}{3}\rho_{\rm Universe}}, \tag{3}$$

where $M_{\rm Pl}$ is the Planck mass, and $\rho_{\rm Universe}$ the total energy density of the Universe.

To understand the scaling of $\rho_{\rm Universe}$, we need to discuss different components of the primordial plasma and, in particular, their interactions.

A. EM plasma and nucleons

Let us first consider the EM plasma. Examples of the processes are Compton scattering and electron-positron annihilation into a pair of photons. The corresponding rate well exceeds the Hubble parameter for times $t \lesssim 10^4$ s, which includes the period we are interested in.[2] This means that the population of the EM particles can always be well described by just one quantity—the temperature of the EM plasma $T_{\rm EM} \equiv T$.

The distribution function $f_{e^\pm}$ of electrons and positrons is Fermi-Dirac, while for photons it is Bose-Einstein:

$$f_{e^\pm}(p,T) = \frac{1}{\exp\left[\frac{\sqrt{p^2+m_e^2}}{T}\right]+1}, \tag{4}$$

$$f_\gamma(p,T) = \frac{1}{\exp[\frac{p}{T}]-1}, \tag{5}$$

with the electron's mass $m_e \approx 0.511$ MeV. The temperature $T_{\rm EM}$ is related to the total energy density of the EM particles $\rho_{\rm EM}$ by the formula

$$\rho_{\rm EM}(T_{\rm EM}) = \rho_{e^\pm}(T_{\rm EM}) + \rho_\gamma(T_{\rm EM}). \tag{6}$$

Here, the energy densities of $e^\pm, \gamma$ are

$$\rho_{e^\pm}(T_{\rm EM}) = g_{e^\pm}\int \frac{d^3\mathbf{p}}{(2\pi)^3}\sqrt{p^2+m_e^2}\, f_{e^\pm}(p,T_{\rm EM}), \tag{7}$$

$$\rho_\gamma(T_{\rm EM}) = g_\gamma \int \frac{d^3\mathbf{p}}{(2\pi)^3}\, p f_\gamma(p,T_{\rm EM}), \tag{8}$$

with the factors $g_{e^\pm} = 4$ and $g_\gamma = 2$ staying for the spin and charge degrees of freedom.

Because of the electroneutrality of the Universe, we may neglect the chemical potential of electrons at the temperatures of interest $T_{\rm EM} \simeq 1$ MeV. Indeed, it is $\mu_{e^\pm}/T_{\rm EM} \sim \eta_B \simeq 10^{-9}$, where η_B is the baryon-to-photon ratio.[3]

In terms of $T_{\rm EM}$, the Hubble factor (3) can be rewritten as

$$H(T_{\rm EM}) \equiv \frac{T_{\rm EM}^2}{M_{\rm pl}^*}, \quad M_{\rm pl}^* \approx \frac{M_{\rm pl}}{1.66\sqrt{g_*(T_{\rm EM})}}, \tag{9}$$

where g_* the effective number of relativistic species: $g_* = \rho_{\rm Universe}/\frac{\pi^2}{30}T_{\rm EM}^4$, with g_i being the number of spin and charge degrees of freedom. Assuming the ΛCDM scenario and that all the species are in perfect equilibrium, we have $g_* \approx g_\gamma + 7/8(g_e + g_\nu) = 10.75$.

Finally, the number density of baryons B in the Early Universe, n_B, may be expressed in terms of the baryon-to-photon ratio η_B and the photon number density:

$$n_B(T_{\rm EM}) = \eta_B(T_{\rm EM}) n_\gamma(T_{\rm EM}). \tag{10}$$

Knowing the value of η_B during the CMB formation, $\eta_{B,\rm Planck} = 6.09 \times 10^{-10}$ [13], and the dynamics of the Universe expansion, the temperature dependence of η_B may be calculated as

$$\eta_B(T_{\rm EM}) = \eta_{B,\rm Planck} \times \left(\frac{a(T_{\rm EM,CMB})T_{\rm EM,CMB}}{a(T_{\rm EM})T_{\rm EM}}\right)^3, \tag{11}$$

where a is the scale factor of the Universe. The scaling of η_B comes from the behavior of the baryon number density, $n_B \propto a^{-3}$, and the number density of photons, $n_\gamma \propto T_{\rm EM}^3$. The resulting temperature-dependent factor stays for the entropy dilution of the Universe. In the standard cosmological scenario, Eq. (11) gives $\eta_B(T_{\rm EM} \simeq 1\ {\rm MeV}) \approx 1.67 \times 10^{-9}$.

The relative ratio between protons and neutrons, important for BBN, is handled by their weak interactions with neutrinos and $e^\pm$ particles, which drive the $p \leftrightarrow n$ conversion, so the baryons are coupled to the UR content of the plasma. However, because of the tiny number density and

[2]The decoupling of EM particles happens much later. In particular, the EM particles' thermalization time is much shorter than any relevant timescale for the temperatures above $T_{\rm EM} \gtrsim 1$ keV. At lower temperatures, for example, by injecting high-energy $e^\pm, \gamma$s, we have a chance for them to photodisintegrate primordial nuclei before the EM particles thermalize [57].

[3]The asymmetry, obviously, becomes non-negligible after e^+e^- annihilation, at $T_{\rm EM} \lesssim m_e$, but then electrons become irrelevant for the dynamics of the primordial plasma.

the absence of other hadrons in the plasma in the standard scenario, nucleons play a negligible role in the thermodynamics of the Universe at MeV temperatures.

B. Neutrinos

Let us now discuss neutrinos. Generically, there may be an asymmetry between neutrinos and antineutrinos, but the minimal cosmological setup assumes zero asymmetry.[4] Neutrinos interact with themselves and $e^{\pm}$ particles via the weak force. The dimensional estimate for the weak interaction rates gives

$$\Gamma_{\rm weak} \simeq n_\nu \cdot \langle \sigma v \rangle \sim G_F^2 T_{\rm EM}^5, \tag{12}$$

where we assumed that the neutrinos have the thermal equilibrium with the EM plasma and, for the moment, set the neutrino temperature $T_\nu = T_{\rm EM}$. Namely, $n_\nu \propto T_{\rm EM}^3$ is the neutrino number density, while $\langle \sigma v \rangle$ is the thermally averaged cross section, which scales as

$$\langle \sigma v \rangle \sim G_F^2 \langle s \rangle \sim G_F^2 T_{\rm EM}^2, \tag{13}$$

and $G_F \approx 1.167 \times 10^{-5}$ GeV^{-2} is the Fermi coupling. The important feature is that the cross section scales with the energies of the interacting particles (we will return to it in Sec. VII).

The rate becomes comparable to the Hubble expansion rate of the Universe already at $T_{\rm EM} \sim 1$ MeV. As a result, at these temperatures, the weak reactions are no longer able to maintain equilibrium in the neutrinos sector, and the latter gradually decouple [58]. The shape of their spectrum in ΛCDM closely follows the Fermi-Dirac one. Its temperature T_{ν_α} remains equal to the EM temperature until the annihilation of electron-positron pairs, which happens around $T_{\rm EM} \simeq m_e$. Then, their relation may be found from the entropy conservation law, giving $T_{\nu_\alpha} \approx (4/11)^{1/3} T_{\rm EM}$. Extended neutrino decoupling introduces a small correction to this relation, leading to the value of $N_{\rm eff}$ that is slightly larger than the instant decoupling result $N_{\rm eff} = 3$.

Neutrino interaction processes. Let us now discuss neutrino interactions in more detail. They include elastic scatterings off neutrinos and $e^{\pm}$ and annihilations:

$$\nu_\alpha + e^{\pm} \leftrightarrow \nu_\alpha + e^{\pm}, \quad \nu_\alpha + \bar{\nu}_\alpha \leftrightarrow e^- + e^+, \tag{14}$$

$$\nu_\alpha + \nu_\beta \leftrightarrow \nu_\alpha + \nu_\beta, \quad \nu_\alpha + \bar{\nu}_\alpha \leftrightarrow \nu_\beta + \bar{\nu}_\beta, \tag{15}$$

as well as charge-conjugated ones [39]. The other reactions include the electroweak corrections, such as subdominant $e^+e^- \to \nu_\alpha \bar{\nu}_\alpha \gamma$.

The MeV plasma is "flavor-asymmetric" in the sense that electrons and positrons are present in plasma, while μ and τ leptons are not. Given the structure of the charged current of weak interactions, which includes the lepton and the corresponding neutrino, the direct interaction rate of ν_e with $e^{\pm}$ is larger than the rate of the corresponding scatterings but with $\nu_{\mu,\tau}$. Because of this, one can naively expect that $\nu_{\mu,\tau}$s decouple earlier from the EM plasma, while ν_es are kept longer in equilibrium. However, besides the interactions (15), neutrinos also experience flavor transitions called oscillations. The oscillations generically appear because the neutrino charge eigenstates do not coincide with the mass eigenstates.

In the primordial plasma, the neutrino oscillation rate is severely affected by the dense medium. Namely, neutrinos acquire a correction to the self-energy caused by interactions with electrons and positrons [58]. It effectively translates to a potential $\mathcal{V}_{\rm eff}^{(\nu_\alpha)}$ in the Hamiltonian describing the propagation of neutrinos ν_α. The functional form of the potential is $\mathcal{V}_{\rm eff}^{(\nu_\alpha)} = C^\alpha \frac{G_F^2 T_{\rm EM}^4 E_\nu}{\alpha_{\rm EM}}$, where C^α is a neutrino-dependent constant.

If $\mathcal{V}_{\rm eff}^{(\nu_\alpha)}$ is higher than the energy splitting for different neutrino eigenstates $\Delta m^2/2E_\nu$, the mixing angle is effectively suppressed, and oscillations can be ignored. Therefore, the oscillations are absent at high temperatures and/or for high-energy neutrinos. In ΛCDM, they effectively turn on at $T_{\rm EM} \simeq 3$ MeV.

In total, because of oscillations, the interactions of three neutrino flavors with the EM plasma are similar. Therefore, the decoupling of ν_e, ν_μ, ν_τ occurs in a similar fashion.

C. How new physics may spoil properties of primordial plasma

There are various ways of introducing new physics to the primordial plasma. They will change the dynamics of the primordial plasma, in particular, departing the neutrino properties from the ΛCDM ones. To be specific, let us consider the scenario appearing in many well-motivated extensions of the Standard Model, adding LLPs (with mass $m \gg T_{\rm EM}$).

To significantly affect the Universe, such particles need to be out-of-equilibrium relics. Before decaying, they would increase the energy density of the Universe and, hence, modify the Hubble factor. After decaying, their influence gets split into many contributions. First, they still modify the dynamics of the Universe by introducing the dilution to the scale factor $a(T)$. It influences the behavior and value of η_B at MeV temperatures via Eq. (11).[5]

Second, their decay products may either constitute additional species ("dark radiation") or inject energy into

[4] In principle, the sector of the charged particles has the asymmetry, but it is at the level of the baryon-to-photon ratio, which can be neglected.

[5] In particular, the presence of additional energy at high temperatures compared to ΛCDM leads to an increase of the scale factor at the CMB epoch, $a(T_{\rm CMB})$. As $\eta_B(T_{\rm CMB})$ is fixed, $\eta_B(T \simeq 1\ {\rm MeV})$ from Eq. (11) must be larger than the ΛCDM value to compensate for the dilution.

the population of neutrinos and the EM particles. The EM population gets immediately thermalized, which results in an increase of $T_{\rm EM}$, while the neutrino injections cause the spectral distortions. Since neutrinos with different energies interact at different rates, much slower than the EM particles, the distortions will not disappear, affecting the total neutrino number and energy densities, as well as the $p \leftrightarrow n$ conversion rates.

Under such scenarios, the nucleons may also be involved in the thermodynamics of the Universe in a nontrivial way. Decaying LLPs may inject relatively long-lived mesons such as $\pi^{\pm}, K^{\pm}, K_L$. Before decaying, these particles experience numerous interactions with the SM plasma particles and themselves. Scattering off nucleons surprisingly becomes very efficient—the smallness of η_B is compensated by the largeness of the nucleon interaction cross section, driven by the strong force [38]. Because of these scatterings, the mesons change the distribution of their energy among the neutrino and EM sector [59], which leads to the impact on the time-temperature relation $t(T_{\rm EM})$ and neutrino properties.

III. EXISTING APPROACHES TO SOLVE THE ν BOLTZMANN EQUATION

In general, to study the thermalization of neutrinos, one has to solve the quantum kinetic equations (QKEs) for the neutrino density matrix [7,8,47,60–62]. However, for our purposes, it may be reasonable to approximate the oscillations by the temperature-dependent oscillation probabilities, $\langle P_{\alpha\beta}\rangle(E_\nu, T_{\rm EM})$, similarly to how this is done in [35,39]. Then, it may be possible to reduce the complexity by converting the QKEs into the Boltzmann equations for the neutrino distribution function f_{ν_α} in the momentum space:

$$\frac{\partial f_{\nu_\alpha}(E_\nu,t)}{\partial t} - E_\nu H \frac{\partial f_{\nu_\alpha}(E_\nu,t)}{\partial E_\nu} = \sum_\beta \langle P_{\beta\alpha}\rangle \cdot \mathcal{I}_{{\rm coll},\nu_\beta}[E_\nu, f_{\nu_\alpha}, f_{\nu_\beta}, T], \tag{16}$$

supplemented with the Friedmann equation describing the expansion of the Universe (and in particular H), the equation for the evolution of the EM plasma temperature $T_{\rm EM}$, and the equation governing the dynamics of LLPs in case they are present. Here, $E_{\nu_\alpha} = |\mathbf{p}_\nu|$ is the neutrino physical momentum. $\mathcal{I}_{{\rm coll},\nu_\beta}$ is the collision integral for the neutrino of the flavor β, which in general contains a source term from new physics particles, a neutrino-neutrino interaction term, and a neutrino-EM interaction term. It has the form [46]

$$\mathcal{I}_{{\rm coll},\nu_\alpha} = \frac{1}{2E_{\nu_\alpha}} \sum \int \prod_{i=2} \frac{d^3\mathbf{p}_i}{(2\pi)^3 2E_i} \prod_{f=1} \frac{d^3\mathbf{p}_f}{(2\pi)^3 2E_f} \times |\mathcal{M}|^2 F[f] (2\pi)^4 \delta^{(4)}\left(\sum_{i=1} p_i - \sum_{f=1} p_f\right). \tag{17}$$

The first summation encompasses all potential interaction processes involving ν_α, with $i = 1$ representing the neutrino itself [$p_1 \equiv (E_{\nu_\alpha}, \mathbf{p}_1)$]. The integral extends over all possible states of ν_α characterized by momentum p_1. Here, i and j denote the initial and final states of a given process, respectively. The term $|\mathcal{M}|^2$ represents the squared matrix element of the process (see Table 3 in Ref. [39] for the explicit expressions of $|\mathcal{M}|^2$ relevant to neutrino processes at MeV-scale temperatures). The factor $F[f]$ accounts for the statistical distribution within the medium and is given by

$$F[f] = \prod_{i=1}(1 \mp f_i) \prod_{f=1} f_f - \prod_{i=1} f_i \prod_{f=1}(1 \mp f_f), \tag{18}$$

where $f_{i,f}$ denote the momentum distributions for the ith and fth particles. Finally, the factor $(1 - f)$ corresponds to Pauli blocking for fermions, whereas $(1 + f)$ corresponds to Bose enhancement for bosons. Finally, the δ function ensures the conservation of the four-momentum in the process.

Depending on the scenario studied, there are two different state-of-the-art approaches to solving the Boltzmann equation (16). If the neutrinos injected by decays of new physics are close to thermal, $E_\nu \simeq 3.15 T_{\rm EM}$, or if decays are solely electromagnetic, then it may be possible to approximate the neutrino distribution by

$$f_{\nu_\alpha}(E_\nu, t) \approx f_{\rm FD}(E_\nu, T_{\nu_\alpha}(t)) = \frac{1}{\exp[\frac{E_\nu}{T_{\nu_\alpha}(t)}] + 1} \tag{19}$$

and consider an integrated version of the Boltzmann equations on the three neutrino temperatures $T_{\nu_\alpha}(t)$ [28,63]. In the limit of negligible electron mass, it may be possible to represent the energy transfer rates in the system as an analytic expression. Another example includes obtaining a correction to the neutrino high-energy tail caused by noninstant decoupling in the standard scenarios [64].[6]

In practice, once we add new physics, the assumption of the perfect thermality of the neutrino distribution is typically violated. The first reason is the energy dependence of the equilibration of neutrinos, as discussed in Sec. II. Neutrinos with different energies interact at very different rates, which leads to neutrino spectral distortions even if we simply heat the EM plasma.

To study the distortions, one needs to solve the Boltzmann equation (16) in the full generality. In the literature, this is done using the approach that we will call the discretization method. The algorithm is to analytically reduce the dimensionality of the integration in $\mathcal{I}_{\rm coll}$ to some

[6]It may be converted to the momentum-dependent correction to the neutrino temperature $T_\nu(p)$ that approaches $T_{\rm EM}$ at $p \gg 3.15 \cdot T_{\rm EM}$ and vanishes at small momenta.

integer k and then discretize the comoving momentum space $y = p \cdot a(t)$.[7] The integrodifferential Boltzmann equation is then converted into a system of ordinary differential equations (see [7,45–47] and references therein).

While the approach has been successfully used in the standard cosmological scenario (see, e.g., [7,10,46,47]), it has limitations when applying it to new physics scenarios. To understand this, let us assume that we inject neutrinos with high energy $E_{\nu,\max}$ in the temperature range from $T_{\rm EM} = T_{\rm ini}$ to $T_{\rm EM} = T_{\rm fin}$. Generically, the computational time of the discretization approach, required to evolve the system during this temperature range, scales as (see Appendix A)

$$t_{\rm comp} \propto E_{\nu,\max}^{k+2}. \tag{20}$$

Here, the factor k is the dimensionality of the collision integral after the analytic reduction of the integration:

$$\mathcal{I}_{{\rm coll},\nu_\alpha} = \int \prod_{i=1}^{k} d\xi_i F(\{\xi\}), \tag{21}$$

with $\{\xi\} = \{\xi_1, \ldots, \xi_k\}$ being integration variables, and F is some function depending on the distribution. The k value is bounded from below by the standard cosmological scenario case, which is $k = 2$ [45]. New physics may drive the computational time (20) to enormously large values, or simply destroy the whole approach via spoiling the reduction (21).

Indeed, first, the discretization approach requires simple analytic matrix elements in the neutrino source terms. In practice, this is not the case when we have hadronically decaying LLPs with mass $m \gg \Lambda_{\rm QCD}$. This is because quarks and hadrons appearing in the decays undergo subsequent showering and hadronization. The latter results in a complicated phase space structure which is hard to fit in the form of an analytic matrix element.

Second, even if simple analytic matrix elements do exist, the computational complexity quickly increases if we depart significantly from the standard cosmological case. For example, simply increasing the integration dimensionality from $k = 2$ to higher values may enormously increase the time of calculations. This is the case of, e.g., $2 \to 3$ scatterings with neutrinos such as the famous $e^+e^- \to \nu\bar{\nu}\gamma$. Another example is when there are n-body decays with $n > 3$, which are quite often for LLPs [65].

Finally, the computational time problem exists even in the most optimistic case $k = 2$. Let us assume injections of neutrinos with large energy $E_\nu \gg T_{\rm EM}$. They may appear in decays of heavy LLPs. Considering, e.g., $E_\nu \sim 1$ GeV would enlarge the computational time compared to the standard cosmological case (where we assume $E_{\nu,\max} = 20$ MeV) by a factor $\sim 50^4 \sim 10^7$ [Eq. (A3)], making any applications impossible in practice. Finally, depending on the energy density of the LLP, it may sizably contribute to the Universe's energy density. For the same temperature range, the scale factor would be larger than in the Standard Model case, which does not allow fixing the maximal comoving momentum in the grid $y_{\max}$.

To summarize, there is no adequate approach to studying the dynamics of primordial neutrinos in the presence of new physics while maintaining model independence, efficiency, and transparency.

IV. BASICS OF DSMC

Consider the Liouville equation for the N-particle probability distribution density $F_N(\mathcal{R}, \mathcal{V}, t)$, where $\mathcal{R}$, $\mathcal{V}$ is the set of coordinates and velocities of the particles, with a short-range potential $\Phi_{i,j}$ of binary interactions:

$$\frac{\partial F_N}{\partial t} + \sum_{i=1}^{N} \mathbf{v}_i \frac{\partial F_N}{\partial \mathbf{r}_i} + \sum_{1 \le i < j \le N} \Phi_{i,j} F_N = 0. \tag{22}$$

The DSMC approach approximately solves it using the following scheme (see [53,66,67] and references therein):

(1) Apply the $N-1$ space variable reduction $F_N \to \tilde{F}_N = \int F_N \prod_{s=2}^{N} d\mathbf{r}_s$.
(2) Switch to the iteration scheme by considering the equation on the time intervals $(t; t + \Delta t)$.
(3) Decompose the space domain $\mathcal{D}$ onto disconnected subdomains $\mathcal{D} = \cup_{l=1}^{M} \mathcal{D}^{(l)}$ ("cells"), populated by fixed amounts of particles during Δt.
(4) Split the evolution into three successive procedures within each time step: ballistic motion (free streaming in the absence of collisions), binary collisions within each $\mathcal{D}^{(l)}$, and then interchanging particles between cells as a result of the first two steps. These collisions may change the kinematics of particles, their types, and number (e.g., via the collision $2 \to n$).

Under an assumption that the system obeys ergodic conditions, the DSMC approach may be converted to an analog of the Bogoliubov-Born–Green-Kirkwood–Yvon hierarchy for $3 + 3N$ phase space, which reduces to the Boltzmann equation in the limit $N \to \infty$ and assuming the molecular chaos (i.e., that the velocities of colliding particles are statistically independent).

A. No-time-counter scheme

The central part of the DSMC approach is to simulate the evolution of particles within an individual cell. There are various methods [53,68–72]. The most efficient ones have $\mathcal{O}(N)$ computational complexity. Examples of the latter are

[7] In the standard cosmological scenario case, the comoving grid is convenient since it "freezes" the neutrino distribution: e.g., the peak of the energy distribution corresponds to the same y at different times.

no-time-counter (NTC), Majorant collision frequency, simplified Bernoulli trial, and others [72]. Here, we will discuss the NTC method, proposed in [71], which we will adapt for our purposes.

First, one defines the time step of the simulation Δt. It must be sufficiently small to resolve the characteristic interaction time in the system. It may be calculated as

$$\Delta t = \left(\frac{(\chi_{\text{particle}} \cdot \sigma v)_{\max} \cdot \mathcal{N}}{V_{\text{system}}} \right)^{-1}. \tag{23}$$

Here, $\mathcal{N}$ is the number of *computational* particles (those actually used in the simulation), on the opposite of the number of *physical* particles N. χ_{particle} is the particles' weight (see below); V_{system} is the system's volume; σ is the interaction cross section; and v is the relative velocity. The subscript "max" denotes finding the maximal value among the system.

Let us discuss the relation between N, $\mathcal{N}$, and macroscopic observables. The quantity N is fixed by the volume V_{system} to represent the number density of the ith species, $n_i = N_i/V_{\text{system}}$. In its turn, it is related to the number of computational particles actually used in the simulation, $\mathcal{N}$, as $N = \sum_{i=1}^{\mathcal{N}} \chi_i$, where χ_i are individual weights of the particles. They need to be introduced if we address some redundancy in the system by replacing multiple particles with a single one. For example, in the setup without charge asymmetries, there is no need to consider particles and antiparticles separately. We can replace electrons and positrons with a single particle having the weight $\chi_e = 2$.

Next, consider splitting the system's volume into cells. Let us assume that there are n_{cells} cells, each having the volume $V_{\text{cell}} = V_{\text{system}}/n_{\text{cells}}$. In the standard DSMC application cases, particular cells contain $N_{\text{cell}} \equiv \mathcal{N}/n_{\text{cells}}$ particles as low as $\mathcal{O}(10$–$20)$ and even lower, which is enough for simulating the evolution properly [53,72]. Within a particular cell, one samples randomly

$$N_{\text{sampled}} = \frac{N_{\text{cell}}(N_{\text{cell}} - 1)}{2} \frac{\omega_{\text{cell,max}} \Delta t}{V_{\text{cell}}} \tag{24}$$

pairs of particles to interact. Here, $\omega_{\text{cell,max}} = (\chi_{\text{particle}} \sigma v)_{\text{cell,max}}$ is the estimate of the maximum interaction cross section within the cell.

For each sampled pair, one accepts its interaction with the probability

$$P_{\text{acc}} = \frac{\omega}{\omega_{\text{cell,max}}}, \qquad \omega = (\chi_{\text{particle}} \sigma v)_{\text{pair}}. \tag{25}$$

If the interaction is accepted, one simulates the possible final states for the given pair and its scattering kinematics.

The complexity of the NTC scheme grows as $\mathcal{O}(N_{\text{cell}})$ [55]. This is achieved by the fact that $\omega_{\text{cell,max}} \Delta t / V_{\text{cell}}$ in the number of sampled events is typically $\ll 1$. The systems with the total number of particles $N \gg 10^6$ may be simulated within minutes, even on ordinary laptops. Such large values are already enough to reach the precision required in our studies.

The NTC method has been tested for various systems, including relativistic ones [73–77], which demonstrates its flexibility and coverage of the wide range of scenarios.

V. DSMC FOR NEUTRINOS

Let us now discuss how to apply the DSMC approach to study the evolution of primordial neutrinos.

As in the case of the state-of-the-art methods, we will first utilize the simplification coming from the properties of the Early Universe at the times of interest—its homogeneity and isotropy. Because of this, we may drop the spatial degrees of freedom and treat the system as effectively zero dimensional, with all interactions occurring at one point. Splitting the system into cells is a formal step to maintain performance because it allows parallelization for applying the NTC scheme. We will also neglect any cells' boundary interactions.

To accurately trace the thermalization of neutrinos, we represent their population by a set of individual particles characterized by the four-momentum, flavor, and particle-antiparticle type. Every interaction involving the neutrino [Eq. (15)] would modify its properties. Namely, it may change its four-momentum (if the interaction is elastic) and/or flavor (if it is the annihilation of the type $\nu_\alpha \bar{\nu}_\alpha \to \nu_\beta \bar{\nu}_\beta$). Finally, there are annihilation processes $\nu_\alpha \bar{\nu}_\alpha \leftrightarrow e^+ e^-$, which may lead to a change in the number of neutrinos.

Proceeding with the traditional DSMC method in the case of the primordial plasma with neutrinos is impossible, as it does not incorporate its fundamental features. These include the expansion of the Universe, the hierarchy between the equilibration rates in the neutrino and EM sectors, the Pauli principle, neutrino oscillations, and the presence of decaying particles. Below, we discuss these features and how we address them in detail (see also Fig. 1, showing the modification of the NTC scheme).

(1) *Expansion of the Universe*. From the DSMC's point of view, it simply represents an external force acting on the particles of the system, with an additional modification of increasing the system's (and cells') volume. These two effects may be simply accounted for by redshifting the total volume of the system V_{system} (and hence the cell's volume) as well as the individual energies E_i of the particles $\{i\}$, applied at each step of the simulation. Namely, at the beginning of the time step Δt, we calculate the Hubble factor H using Eq. (3), and then make use of the relation

$$V_{\text{system}} \to V_{\text{system}}(1 + 3H\Delta t), \qquad E_i \to \frac{E_i}{1 + H\Delta t} \tag{26}$$

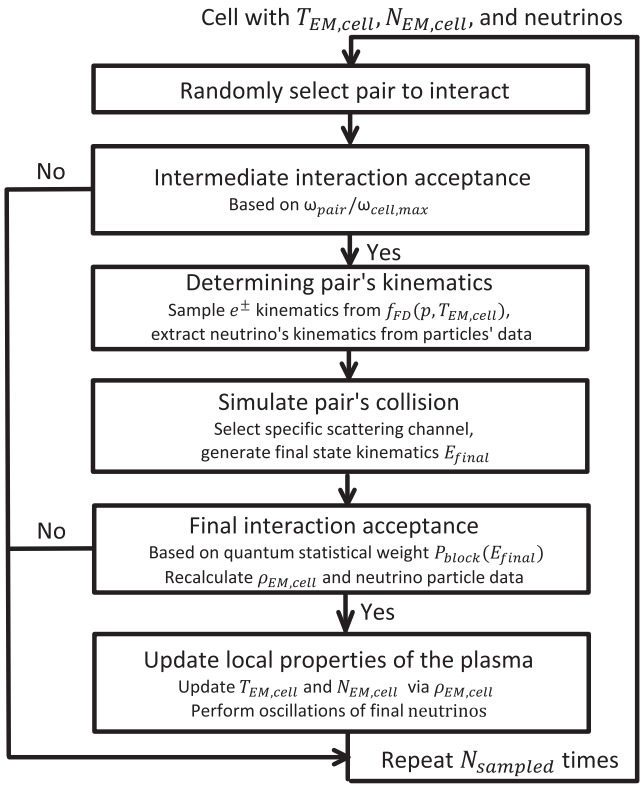

FIG. 1. The modification of the no-time-counter scheme, used to simulate the interactions within the system's cells within the direct simulation Monte Carlo approach, for describing interactions in the MeV primordial plasma. First, we sample $N_{\rm sampled}$ pairs to interact, Eq. (24). For each pair, we compute its interaction weight and make an intermediate decision on whether it will interact using the criterion (25). Then, we sample the kinematics of the interacting particles, generate the final states resulting from the collision, and make the final decision of whether the interaction takes place from the Pauli principle (29). Finally, we update the local properties of the plasma: the EM plasma temperature and the number of EM particles, as well as neutrino flavor distributions by the oscillation probabilities, Eq. (30).

provided that $H\Delta t \ll 1$. To account for this requirement, we modify the definition (23):

$$\Delta t = \min\left[0.01H^{-1}, \left(\frac{(\chi_{\rm particle}\cdot\sigma v)_{\rm max}\cdot N}{V_{\rm system}}\right)^{-1}\right]. \tag{27}$$

Here, 0.01 is an arbitrary small factor. Given varying H, volume $V_{\rm system}$, and neutrino energy E_ν throughout the evolution of the system, the value Δt gets updated at the beginning of each iteration during the simulation using the following formula for the maximal rate:

$$(\chi_{\rm particle}\sigma v)_{\rm max} = \frac{2G_F^2}{3\pi}\cdot 2E_{\nu,\rm max}\langle E_\nu\rangle, \tag{28}$$

where we used $\chi_{\rm particle} = 1$ (provided that we implemented particles and antiparticles separately) and the estimate for $(\sigma)_{\rm max}$ corresponding to the process $\nu_\alpha + \bar{\nu}_\alpha \to \nu_\alpha + \bar{\nu}_\alpha$, which is the fastest among all the neutrino interaction processes (15) [39]. The factor $2E_{\nu,\rm max}\langle E_\nu\rangle$ stays for the averaged invariant mass of the interacting high-energy neutrino with the rest of the neutrinos.

(2) *Properties of the EM plasma*. As we discussed previously in Sec. II, the reactions involving solely EM particles are orders of magnitude faster than those where neutrinos participate. As we have chosen the time step Δt comparable to the neutrino interaction rates, the EM particles may be viewed as a part of perfectly thermal plasma characterized by one parameter—temperature $T_{\rm EM}$. However, we then need to implement the response of any single interaction involving the EM particles on $T_{\rm EM}$.

To reach this, at the beginning of the iteration, we characterize the EM plasma with the energy density $\rho_{\rm EM}$, both globally (for the whole system) and locally (at the level of the individual cell). The global and cells' EM temperatures, which we denote by $T_{\rm EM}$ and $T_{\rm EM,cell}$ respectively, is related to the global and cell's energy densities of the EM plasma $\rho_{\rm EM}$, $\rho_{\rm EM,cell}$ by Eq. (6).

During the NTC routine, the local number of electrons and positrons $N_{e^\pm,\rm cell}$ per cell is calculated from the relation between $T_{\rm EM,cell}$ and the number density $n_{\rm EM}(T_{\rm EM,cell})$. The kinematics of any $e^\pm$ selected within the NTC algorithm is sampled from the Fermi-Dirac distribution $f_{\rm FD}(p, T_{\rm EM,cell})$. The change in $\rho_{\rm EM,cell}$ resulting from the accepted interaction leads to the update in $T_{\rm EM,cell}$ and $N_{e^\pm,\rm cell}$.

At the global level, once the simulations for all cells are performed, $\rho_{\rm EM,cell}$ are merged into the total energy density $\rho_{\rm EM,system}$, which allows obtaining the global temperature of the EM plasma.

(3) *Quantum statistics*. It enters the binary part of the collision integral (16) with fermionic final states F_1, F_2 having energies $E_{F_{1,2}}$ via multiplicative Pauli blocking factors,

$$P_{\rm block} = \left(1 - f_{F_1}(E_{F_1})\right)\times\left(1 - f_{F_2}(E_{F_2})\right), \tag{29}$$

where f is the energy distribution of the given final state. Thus, it suppresses interactions where the final states would occupy the high-populated part of the energy distribution (e.g., $E \lesssim T$ for the equilibrium shape distribution with the temperature T). To implement this, one should consider the local energy distributions for both EM particles and neutrinos and calculate $P_{\rm block}$. A possible simplification is, when calculating the blocking factor, to describe neutrino's distribution by the Fermi-Dirac function $f_{\rm FD}(T_{\nu_\alpha,\rm cell})$, where $T_{\nu_\alpha,\rm cell}$ is the local effective

neutrino temperature obtained in a way similar as we do for the EM plasma.[8]

(4) *Neutrino oscillations*. We incorporate them at the end of the iteration time step by changing each of the neutrino flavors according to the formula

$$\nu_\alpha(E_\nu) \to \sum_\beta \langle P_{\alpha\beta}\rangle(E_\nu, T_{\rm EM})\nu_\beta(E_\nu), \quad (30)$$

where $\langle P_{\alpha\beta}\rangle(E_\nu, T_{\rm EM})$ are averaged neutrino oscillation probabilities [also Eq. (16)]. For the neutrino oscillation parameters, we use the results from [78].

(5) *Presence of LLPs X and new interactions*. Let us start with discussing LLPs. Further, we assume that LLPs are nonrelativistic and decoupled at the temperatures of interest, which ideally matches the scope of this study. Decaying either into the EM plasma particles or neutrinos, they would heat the EM plasma temperature and distort the neutrino bath.

Having the initial condition for the X's number density, $n_X(T_{\rm ini})$ for some temperature $T_{\rm ini}$, at the beginning of the simulation, we add the amount N_X of X particles fixed in a way such that $N_X/V_{\rm system} = n_X(T_{\rm ini})$. Per each time step Δt, provided that it is much smaller than the LLP's lifetime τ_X, their number is evolved by the exponential distribution: $dN_X/dt = -N_X/\tau_X$.

For each decay, it is possible to obtain the energies of resulting neutrinos and calculate the amount of the EM energy using Monte Carlo simulations—the baseline approach for particle physics. This is a natural choice if one wants to maintain the model independence, as it is maximally general and may describe any process. In particular, exclusive decays (where we have well-defined "fixed" final states, e.g., $X \to 3\pi$) may be simulated on-flight by sampling the phase space of decay products using the analytic matrix element of the process. The phase space of hadronic decays in the LLP mass range $m \gg \Lambda_{\rm QCD}$ (such as $X \to q\bar{q}\nu$, where q is a quark) may be obtained by simulating them in Pythia8 [79] for a grid of masses and subsequently using the output particle's data in the form of events inside the DSMC code.

The Monte Carlo sampler must incorporate the interactions of the decay products with the primordial plasma, which may substantially redistribute their energy between the neutrino and EM sectors compared to the vacuum case. Namely, all electrically charged particles with lifetimes $\tau \gtrsim 10^{-10}$ s, such as muons, charged pions, and kaons, appearing in the MeV plasma may undergo kinetic energy loss via EM interactions, annihilation, and interactions with nucleons before decaying [80]. This evolution may again be implemented probabilistically, in the spirit of Monte Carlo simulations.

Absolutely similarly, it is possible to sample the energies for the nonstandard scattering processes, e.g., for the $2 \to 3$ scatterings $e^+e^- \to \nu_\alpha\bar{\nu}_\alpha\gamma$.

In order to finish the discussion of the approach, let us address the question of the number of particles per cell, $N_{\rm cell}$, entering Eq. (24). In our system, it is

$$N_{\rm cell} = N_{e^\pm,{\rm cell}} + 2\sum_\alpha N_{\nu_\alpha,{\rm cell}}. \quad (31)$$

Since statistical quantities, such as temperatures, are involved in simulating the interactions, it is not possible to use small $N_{\rm cell} \sim 10$, as it is typically done in the DSMC simulations. Instead, the values as large as $N_{\rm cell} = \mathcal{O}(100)$ should be considered (see Appendix C). As a bonus, such a large number also allows for avoiding various stochastic problems of the NTC method, including repeated interaction of the same pair [52].

Now, let us discuss the values of $N, N_{\rm cell}$ we use and the scaling of the DSMC simulation time with the maximal neutrino energy in the system $E_{\nu,\max}$. More details may be found in Appendix C, and here we make a summary.

The typical number of particles per neutrino flavor we consider in the setup is $N_\nu \simeq 10^6$, which results in $N = {\rm few} \times 10^6$. The standard number of particles per cell we have chosen is $N_{\rm cell} = 400$. These numbers are enough to keep the statistical noise at the level of 0.1% in the absence of high-energy neutrinos.

The scaling of the computational time with $E_{\nu,\max}$ is *linear* to *quadratic* (in some marginal cases, as we comment on below). The scaling comes from the unavoidable linear dependence of the number of time steps on $E_{\nu,\max}$ [Eq. (27)] and the possible scaling $N(E_{\nu,\max})$. The latter may be required to maintain a computationally large enough number of high-energy injected neutrinos to avoid fluctuations in the microscopy of thermalization. As far as $E_{\nu,\max} \lesssim 1$ GeV, N may be kept constant, and the scaling of the simulation time is linear. If $E_{\nu,\max} \gtrsim 10$ GeV, one would need to increase N to keep the number of injected neutrinos large enough to avoid fluctuations. However, the increase is *linear* with $E_{\nu,\max}$. As a result, in this worst-case scenario, the scaling of the running time becomes $E^2_{\nu,\max}$—still much better than the scaling of the discretization approach, Eq. (20).[9]

[8]The actual neutrino distribution is, of course, nonthermal, and we use this approximation only when calculating $P_{\rm block}$. Since the deviations from the thermality we study are not very large without loss of generality, we believe that the approximation is accurate.

[9]There may be, in principle, an additional slowdown coming from the need to distribute particles into cells at the beginning of each iteration. We have checked that this splitting only costs a tiny fraction of the whole time independently of the value of the number of computational particles $\mathcal{N}$ and, hence, does not add anything on top of the expected scaling.

VI. CURRENT IMPLEMENTATION

We have implemented a simplified version of the DSMC method described above, which serves as proof of principle.[10]

The main approximation of the *current* implementation is that we have neglected the electron mass m_e when describing the population of the EM particles; this is done to simplify the sampling of $e^{\pm}$ particles and relate the total energy of the EM plasma to its temperature.

Although keeping m_e finite is necessary to know the final value of neutrino-to-EM energy densities ratio ($= N_{\rm eff}$), it is irrelevant for studying the main topic of this work—nonequilibrium dynamics of neutrinos at the times when they start decoupling, and qualitative behavior such as the sign of the correction $\Delta N_{\rm eff} = N_{\rm eff} - N_{\rm eff}^{\Lambda\rm CDM}$. This is because of two reasons. First, m_e does not affect the dynamics at MeV temperatures (the domain of interest of this study), since it can be simply neglected compared to the typical electrons' energies of $E_e \approx 3.15 \cdot T_{\rm EM}$. To validate this statement, we compare the predictions of DSMC at MeV temperatures with the approaches keeping finite electron mass and next-to-leading order (NLO) QED corrections, and find a perfect agreement. Second, including it at lower temperatures cannot change the sign of $\Delta N_{\rm eff}$, modifying only its magnitude $\Delta N_{\rm eff}$.

We will include the electron mass and its QED corrections [5,81,82] in future papers delivering the full DSMC implementation.

The implementation is written in *Mathematica*. However, low-level routines, such as simulations of interactions and manipulations with cells, are compiled in C++. This approach allows combining moderate performance with symbolic calculations, which are needed when dealing with describing kinematics and deriving the matrix elements of various processes. Also, it makes it possible to use existing realizations of Monte Carlo sampling of decays of LLPs, such as SensCalc [83]. In the next revisions, we will write a part of the code in native C++ and use it as a library inside *Mathematica*.

On a laptop with CPU `AMD Ryzen AI 9 HX 370`, the running time required to produce most of the plots below is $\lesssim$5 min; it varied only mildly depending on the setup, including the energies of the neutrinos included in the system. In particular, in order to produce the neutrino distributions shown in Fig. 5, we spent only 30 s. We expect significant improvement, possibly by an order of magnitude, in the running time after optimizing the code and/or rewriting some of its modules in native C++. Finally, with the implementation, we maintain the approximate linear scaling of the computational time with N, as expected from the basics of the NTC approach.

[10]The code may be provided upon request.

To validate the developed neutrino DSMC, we have studied its predictions in the case of well-established scenarios, including the following:

(1) *Approaching thermal equilibrium*. In the absence of Universe expansion, independently of the initial conditions, neutrinos have to reach thermal equilibrium with the EM particles. In particular, their differential distribution in the number and energy densities, which we will plot throughout the paper, must be

$$\frac{dn_\nu}{dE_\nu} = \frac{g_\nu}{2\pi^2} f_{\rm FD}(E_\nu, T_\nu) \times E_\nu^2, \tag{32}$$

$$\frac{d\rho_\nu}{dE_\nu} = \frac{g_\nu}{2\pi^2} f_{\rm FD}(E_\nu, T_\nu) \times E_\nu^3, \tag{33}$$

where $T_\nu = T_{\rm EM}$ is the neutrino temperature, two powers of E_ν come from the phase space, and one in Eq. (33) from the definition of ρ_ν. Finally, $f_{\rm FD}$ is the Fermi-Dirac distribution [Eq. (5)], with g_ν being the lepton charge degree of freedom [Eq. (9)].

Equation (33) automatically implies that, in equilibrium, the ratio of the energy densities of the neutrino and EM plasmas is

$$\left(\frac{\rho_\nu}{\rho_{\rm EM}}\right)_{\rm eq} = \frac{7/8 \cdot g_\nu}{7/8 \cdot g_e + g_\gamma} = \frac{21}{22}, \tag{34}$$

where we have used Eq. (6) and assumed $T_{\rm EM} \gg m_e$.

(2) *Energy transition rates*. Consider the initial setup where the distribution function of neutrinos is fixed by $f_{\rm FD}$, parametrized with the temperature $T_{\nu_\alpha} \neq T_{\rm EM}$. During the equilibration and in the absence of expansion, the energy transition rates between the neutrino and EM sectors must match the well-known analytic result from [28] (where we turn off the expansion as well).

(3) *Expansion and decoupling*. If including the expansion of the Universe in the previous setup, we should consistently recover the decoupling of neutrinos, which prevents their population from full thermalization, as well as reproduce the results of [28].

Details may be found in Appendix D. In addition, we have performed tests that are not present in the paper. Those include the evolution of neutrinos and antineutrinos (the evolution must preserve the lepton symmetry up to Monte Carlo fluctuations) and independence on the exact simulation setup (e.g. number of simulation cells, the total number of particles, etc.). We believe that it proves that our approach fulfills the requirements to be accepted as a valid method for treating the evolution of neutrinos.

VII. CASE STUDIES

To demonstrate the potential of various implications of the DSMC method, we will consider several toy case

studies specified by the initial conditions on the neutrino distribution functions. These setups have two applications. On the one hand, they mimic distinct scenarios with new physics and thus provide useful insights into the dynamics of the primordial plasma. On the other hand, they will comprehensively demonstrate the performance and flexibility of the neutrino DSMC approach.

First, we investigate the evolution of a system where neutrinos initially possess an equilibrium energy distribution with a temperature $T_{\nu_\alpha} \neq T_{\rm EM}$ (see Sec. VII A). This setup encompasses two distinct scenarios. The first scenario arises when energy is injected exclusively into the electromagnetic (EM) sector, resulting in $T_{\rm EM} > T_{\nu_\alpha}$. The second scenario occurs when nearly thermal neutrinos are introduced into the neutrino sector, as explored in [28]. These cases can be analyzed using the integrated Boltzmann equation developed in [28,63]. Nevertheless, we will demonstrate that, even within these simplified setups, deviations from the thermal shape of the neutrino distribution emerge, leading to discrepancies between the solutions of the unintegrated and integrated approaches to the neutrino Boltzmann equation, particularly in the determination of $N_{\rm eff}$.

Second, we will consider injections of high-energy monochromatic neutrinos (Sec. VII B). This scenario represents the case of two-body decays of heavy LLPs, such as neutrinophilic scalars [84], majorons [85], $B-L_\alpha$ mediators [86], and relics in late reheating scenarios [34]. In the context of our studies, it is preferable over the realistic continuous injections by decaying LLPs because of the transparency of the analysis; despite the simplicity, understanding the dynamics of instant injections provides qualitative insights for the continuous decays, which we explore in our companion Letter [56].[11]

We will consider high injection temperatures, $T_{\rm EM} \gtrsim 1$ MeV. We will show that in the case of sufficiently large neutrino energy, such that $E_\nu \gg T_{\rm EM}$, these injections would result in a decrease in the neutrino-to-EM energy densities ratio compared to the standard cosmological scenarios. This setup will also serve to demonstrate that the performance of the DSMC does not depend on the neutrino energy (supporting the initial expectations) and to cross-check it by comparing the neutrino evolution with the predictions of the discretization codes.

Finally, we will study injections of neutrinos from decays of different long-lived SM particles, such as muons, charged pions, and kaons (Sec. VII C). This case corresponds to a common scenario of LLPs with complex decay chains, which may not decay into neutrinos directly but instead decay into such heavy states. Examples are, e.g., a decay of the Higgs-like scalars into $\pi^+\pi^-/K^+K^-$, the dark photon decay into $2\pi/3\pi/4\pi$, and decays of HNLs into $\pi\mu$ [87]. Another illustrative case is the decay into quarks, where we have a high multiplicity of meson states. We will show that, independently of the decaying particle (or the fraction of their energy placed to the neutrino plasma right after decay), the ratio (35) decreases below the equilibrium value. This case also study demonstrates the flexibility of our approach, which may handle any decay chain with complicated kinematics.

To make the illustrative analysis for this and other studies performed in this paper, we introduce the quantity

$$\delta\rho_\nu = \left(\frac{\rho_\nu}{\rho_{\rm EM}}\right)^{-1}_{\rm eq} \frac{\rho_\nu}{\rho_{\rm EM}} - 1. \tag{35}$$

Throughout the section, we will compare the predictions of DSMC with the modified discretization approach from [7], which we develop for the work [59]. The brief description of the approach can be found in Appendix B. It serves two important purposes. First, the overall agreements between DSMC and this approach serve as a very robust cross-check of our method, showing how well it traces the dynamics of neutrinos. In particular, in the discretization approach, we keep finite electron mass and include LO QED corrections. Finding the very good agreement, we validate our approximation of neglecting the electron mass in our proof-of-principle study.

Second, we will compare the performances of the two methods, highlighting the setups where the discretization approach becomes inapplicable in practice.

A. From equilibrium spectral shapes to distortions

Let us consider a system with neutrinos having an equilibrium shape of energy distributions, but the temperatures of these distributions differ from the EM plasma temperature.

We will study how the equilibration of this initial condition evolves in time, to identify the possible deviations from the description dynamics of the equilibration following Ref. [28], where we turn off the electron mass in order to compare apples with apples. These deviations genuinely appear from the nonthermal distortions in the neutrino sector (invisible within the method of [28]). It is because the interaction rates of different parts of the neutrino spectrum are energy dependent (Sec. II).

We will consider the particular initial condition where neutrinos have the same temperature $T_{\nu_\alpha} = 3.5$ MeV, and the EM plasma has a lower temperature $T_{\rm EM} = 3$ MeV.

The resulting evolution of $\delta\rho_\nu$, as predicted by the DSMC approach and the method from [28], is shown in Fig. 2. From the figure, we see that in terms of $\delta\rho_\nu$, the two descriptions match at the initial stages, while the deviations

[11]For the instant injections, the computational time of both the discretization approach and DSMC does not include scaling of the number of time steps with the injected neutrino energy $E_{\nu,\rm max}$ (Appendix A). This is because they quickly lose their energy, and the time step required to resolve their thermalization no longer depends on $E_{\nu,\rm max}$.

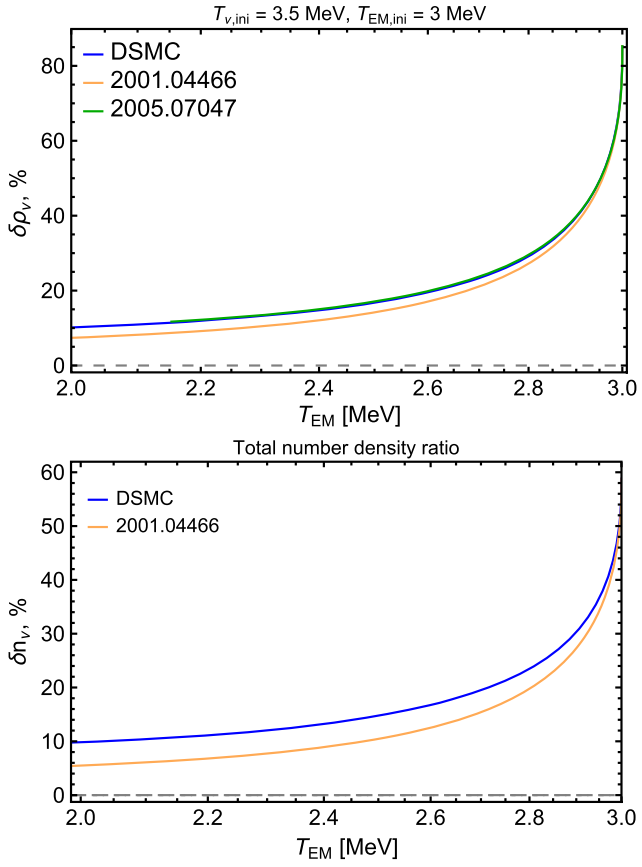

FIG. 2. The evolution of neutrinos and the EM plasma energy densities ratio under the scenario where the neutrino distribution shape is thermal [Eq. (19)], but has temperature T_ν different from the EM plasma T_{EM}. For the initial setup, we consider $T_\nu = 3.5$ MeV and $T_{\mathrm{EM}} = 3$ MeV. Top panel: the energy densities ratio $\delta\rho_\nu$, given by Eq. (35). The blue line shows the result of our DSMC approach, the green line denotes the prediction of the modified code from [7,59], whereas the orange line is obtained using the method of integrated neutrino Boltzmann equations from [28], which assumes that the shape of the neutrino distribution is perfectly thermal throughout the whole evolution. Bottom panel: the analog of $\delta\rho_\nu$ but for the number densities of neutrinos and the EM particles, highlighting the deviation from the thermality of the neutrino spectrum throughout the evolution.

appear once the system develops, signaling the accumulating neutrino spectral distortions. They get frozen throughout the evolution because of the decoupling of neutrinos. The same conclusion holds in the opposite case of the initial condition $T_{\mathrm{EM}} > T_\nu$. The accumulation of the distortions is easily visible if considering the analog of $\delta\rho_\nu$ but for the number density of neutrinos, which we show in the bottom panel of the figure. It demonstrates that for this setup, the distortions build up because of the suppression of the $\nu\bar{\nu} \to e^+e^-$ annihilation rate, which keeps the extra energy stored in the neutrino sector.

Therefore, we conclude that the integrated Boltzmann approach may provide insufficient accuracy even in cases where there are no direct distortions of the neutrino spectrum (see further discussion of this point in Ref. [59]).

The DSMC predictions perfectly agree with the discretization method from [7,59]. Both approaches work reasonably fast—within a minute, but the discretization approach is $\mathcal{O}(2)$ times faster. This is explained by the smallness of the maximal neutrino energy—the scenario for which the discretization works well.

B. Instant neutrino injection

Let us proceed to a different scenario in which there are injections of nonthermal neutrinos with $E_\nu \gg T_{\mathrm{EM}}$. For this setup, the integrated Boltzmann approach is completely inapplicable, as high-energy neutrinos have a much larger rate of interactions than their thermal counterparts, severely influencing the dynamics of the thermalization even if their amount is low.

We will study the injection of monochromatic neutrinos with energy $E_{\nu,\mathrm{inj}}$ at temperature $T_{\mathrm{EM}} = 3$ MeV, and consider three different values $E_{\nu,\mathrm{inj}} = 20, 70, 500$ MeV. We will analyze both the evolution of $\delta\rho_\nu$ and the neutrino spectrum shape.

The option $E_{\nu,\mathrm{inj}} = 20$ MeV primarily serves to demonstrate the necessity of using the unintegrated Boltzmann approach in case of nonthermal distortions. The second setup $E_{\nu,\mathrm{inj}} = 70$ MeV is central—it will show the qualitative impact of large neutrino energies on N_{eff}. We will use it to compare with the discretization codes from [7,40,41], predicting contradictive behavior of the sign of $N_{\mathrm{eff}} - N_{\mathrm{eff}}^{\Lambda\mathrm{CDM}}$ in the presence of high-energy neutrinos. Finally, the highest energy case $E_{\nu,\mathrm{inj}} = 500$ MeV highlights the performance of our setup—the running time and precision are almost independent of the neutrino energy.

1. Injection of 20 MeV neutrinos

Consider the injection of 20 MeV neutrinos. We assume equal injection among the three neutrino flavors, with the total injected energy density $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 5\%$. Here and below, we include the Hubble expansion of the Universe, but turn off the neutrino oscillations.

The evolution of the resulting $\delta\rho_\nu$ is shown in Fig. 3, where we, as usual, also include the prediction of the integrated Boltzmann approach. Both approaches predict a monotonic decrease of $\delta\rho_\nu$. In particular, at late temperatures, when the expansion prevents equilibrating, we end up with the value of $\delta\rho_\nu$ close to 0. However, the rate of decrease of $\delta\rho_\nu$ predicted the neutrino DSMC is much faster. This is explained by the fact that, compared to thermal particles, the injected high-energy neutrinos have a larger probability of interacting with the EM sector and, hence, transporting their energy.

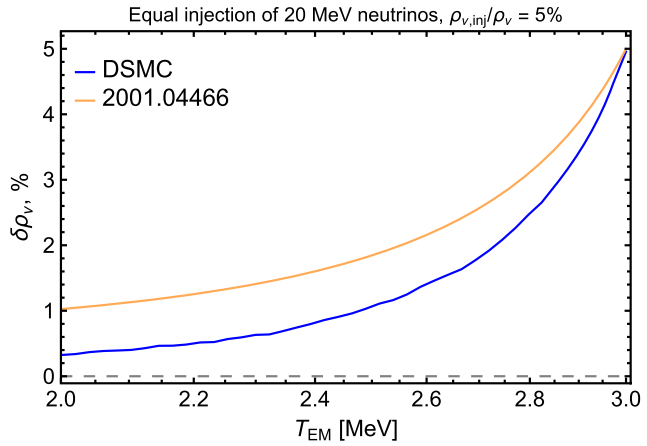

FIG. 3. The behavior of the ratio (35) under the injection of 20 MeV neutrinos equally to all neutrino flavors at the temperature $T_{\mathrm{EM}} = 3$ MeV. The total injected energy density is $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 5\%$. The blue line shows the prediction of the DSMC method, whereas the orange one corresponds to the integrated Boltzmann approach from [28].

2. Injection of 70 MeV neutrinos

Let us now proceed with the 70 MeV injection. We will consider several setups here. The first one is with equal injection among the neutrino flavors and a large $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 30\%$. It serves as a very illustrative demonstration of the qualitative features of the evolution of $\delta\rho_\nu$. The two others are with the smaller injected energy $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 5\%$ and two different injection patterns: equal energy distribution among the flavors, and the injection solely into the sector of electron neutrinos. We will use them to compare with the predictions of different discretization codes from the literature.

Figure 4, upper panel, shows the evolution of $\delta\rho_\nu$ for the 30% injection setup. Now, there is a qualitative difference in its behavior between the integrated and DSMC approaches. The former results in the naively expected monotonic decrease of $\delta\rho_\nu$, whereas according to the latter, it first rapidly drops below zero, where it then freezes out. Without the expansion of the Universe, it would have been a decrease of $\delta\rho_\nu$ to negative values, and then a slow monotonic reaching $\delta\rho_\nu \to 0$ from below.

To understand this counterintuitive result, let us remind of Sec. II and highlight two important properties of the plasma: (i) EM particles instantly equilibrate between themselves, and (ii) weak interaction rates grow with the invariant mass of colliding particles. Because of this, the injected nonthermal neutrinos quickly "knock out" thermal neutrinos by the interactions

$$\nu_{\mathrm{inj}}\bar{\nu}_{\mathrm{thermal}} \to e^+e^-, \quad \nu_{\mathrm{inj}}\nu_{\mathrm{thermal}} \to \nu\nu. \tag{36}$$

The first process pumps the injected energy and a fraction of the energy of the thermal population to the EM sector.

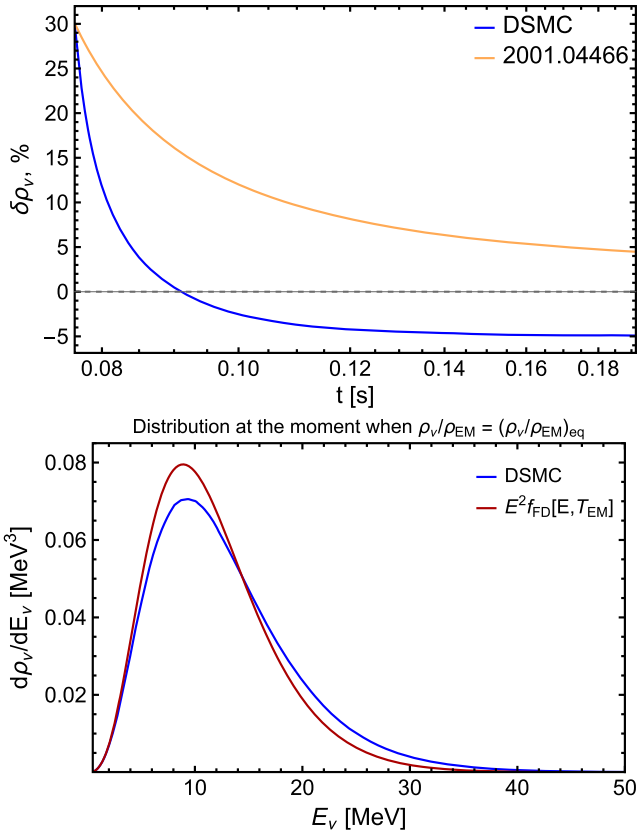

FIG. 4. The temporal evolution of the plasma after the injection of neutrinos with energies $E_\nu = 70$ MeV and the overall energy density $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 30\%$. The other parameters of the setup are similar to the one considered in Fig. 3. Top panel: the behavior of $\delta\rho_\nu$ with temperature, where we show the predictions of the DSMC (the blue curve) and the integrated approach from Ref. [28] (the orange curve). Bottom panel: comparison of the shape of the neutrino energy distribution for the system from Fig. 4 at the moment when $\delta\rho_\nu = 0$ during the equilibration, as obtained with the DSMC simulation (the blue curve) and assuming the equilibrium neutrino spectrum (the red curve).

The rate of these processes is much higher than the rate of the same processes when only thermal particles are involved. Knocking out thermal neutrinos determines the shape of the neutrino spectrum during these interactions: compared to the equilibrium spectrum f_{FD}, it is underabundant in small energies and overabundant in large energies.

The snapshot of the neutrino spectrum at the moment when $\delta\rho_\nu = 0$ is shown in the lower panel of Fig. 4. Then, we have equilibrium amounts of energies in the EM and neutrino sectors. However, while the EM plasma has a perfect thermal spectrum, the neutrino spectrum has a shift to higher energies.

The further dynamics of $\delta\rho_\nu$ depends on the balance between the energy transfer rates $\nu \to \mathrm{EM}$ and $\mathrm{EM} \to \nu$. Because of the energy dependence of the weak processes' rate, the overabundance of the high-energy neutrino leads

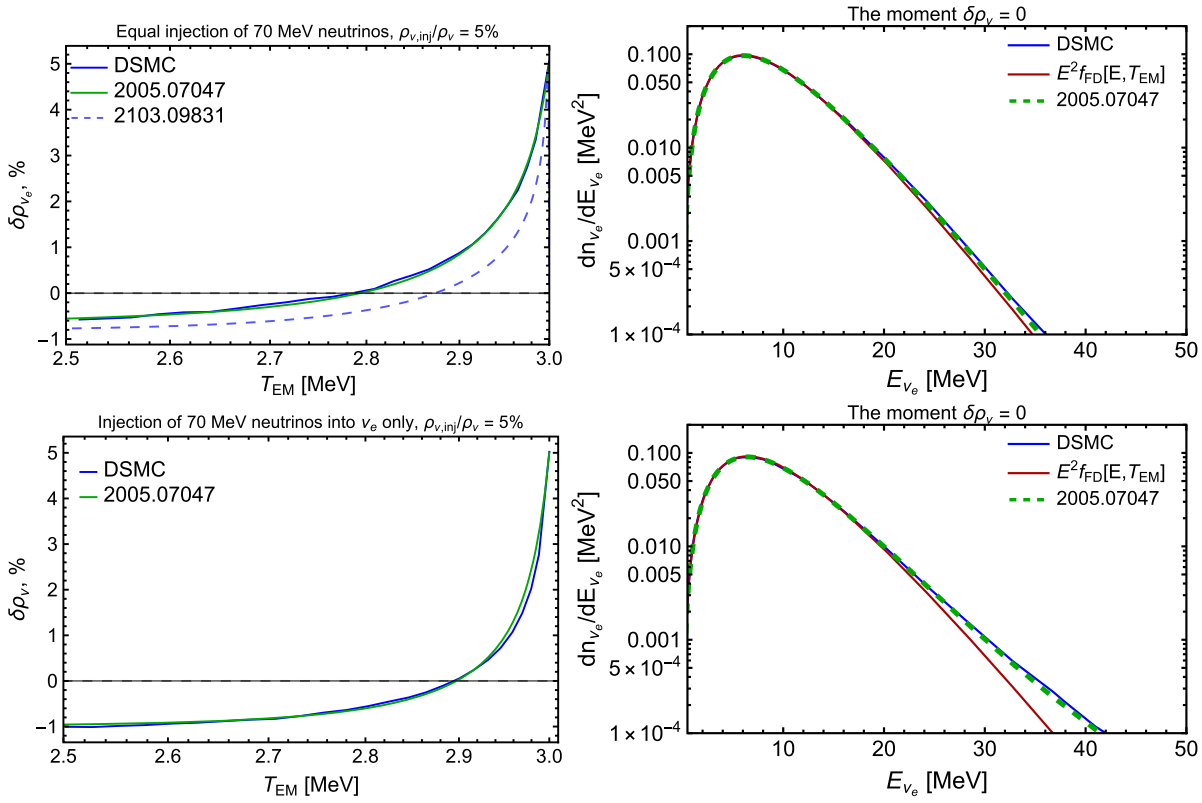

FIG. 5. Comparison of the DSMC approach with the discretization codes for the setup of injection of 70 MeV neutrinos at $T_{\mathrm{EM}} = 3$ MeV. Two configurations are considered: equal injection among the flavors (the top panels) and the injection solely into ν_e (the bottom panels). In both cases, the injected energy fraction is $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{total}} = 5\%$. The left plots show the evolution of $\delta\rho_\nu$, given by Eq. (35). In the plots, the blue lines are the DSMC predictions, the green lines denote the calculation by the discretization approach from [7] (see also [80]), and the dashed blue line is the result obtained in [40] (see text for discussions). The right plots are snapshots of the electron neutrino distribution spectrum at the temperature when $\delta\rho_\nu = 0$. In addition to the results from the DSMC and Ref. [59], we include the plot of the equilibrium neutrino distribution given by $E_\nu^2 f_{\mathrm{FD}}(E_\nu, T_{\mathrm{EM}})$ (solid red lines).

to the faster transfer $\nu \to$ EM than EM $\to \nu$, where we have thermal electrons. As a result, $\delta\rho_\nu$ continues falling below zero until neutrino-induced heating of the EM plasma temperature and/or the expansion of the Universe turn the negative energy transfer from the neutrino sector to zero.

Since the sign of $\delta\rho_\nu$ is associated with the sign of the correction ΔN_{eff}, we conclude that the injection of such high-energy neutrinos is associated with a decrease in N_{eff} below its ΛCDM value. This conclusion holds in the case when the EM plasma temperature is high enough during the neutrino injection, such that the interactions between the neutrinos and the EM plasma are possible.

A similar result has been obtained in our previous work [40], which considered a setup with the injection of 70 MeV neutrinos but with a smaller amount within the discretization approach. The same behavior has been observed when considering the cosmological impact of HNLs decaying mainly into neutrinos (see also Refs. [35,42]). These results, however, contradicted Ref. [41] (see also [33]), which studied the same setup with HNLs with masses below the pion mass and found that N_{eff} may only increase. Given that all of these studies are based on the discretization method, the discrepancy became an open question. Our approach is completely independent and, therefore, resolves it.

We finish this discussion by directly comparing our method with the discretization codes. Let us consider the setup when we inject 70 MeV neutrinos with the amount $\rho_{\nu,\mathrm{inj}}/\rho_{\nu,\mathrm{tot}} = 5\%$. Figure 5 shows the evolution of $\delta\rho_\nu$ and neutrino spectra snapshot according to DSMC and the discretization codes from [7,40], where for the latter we take the results shown in Fig. 7 from Appendix A. In the discretization codes, the electron mass effects are included.

We see a very similar behavior of the evolution predicted by DSMC and the discretization method from [59], both in terms of $\delta\rho_\nu$ and the spectrum. The tiny discrepancy may be explained by the fact that we have neglected the electron mass in the DSMC calculations. On the other hand, the discrepancy between DSMC and Ref. [40] is somewhat larger. This is explained by the fact that the caption of Fig. 7 in Ref. [40] wrongly mentions the setup other than the one actually used to make the plot. Unfortunately, the information about the true setup has been lost.

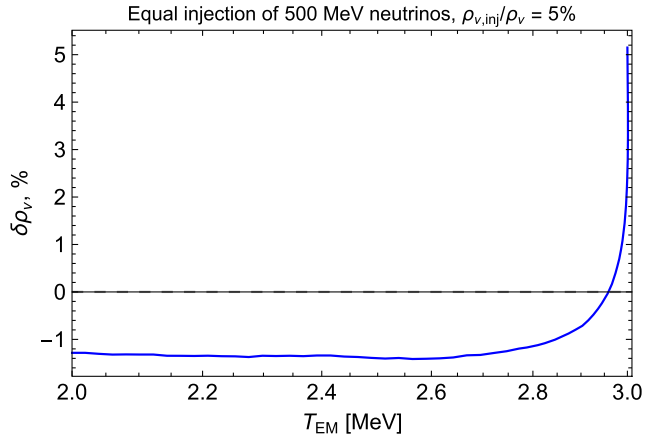

FIG. 6. The same setup as in Fig. 3 but under an injection of 500 MeV neutrinos.

The running time of the DSMC simulation required to obtain Fig. 5 is within a minute. In contrast, the discretization approach we used required ≃10 min.

3. Injection of 500 MeV neutrinos

Let us finalize this case study by considering the injection of 500 MeV neutrinos. The behavior of $\delta\rho_\nu$ is shown in Fig. 6; it resembles the features shown in the case of the injection of 70 MeV neutrinos.

The more important point is the performance of the DSMC setup. The running time required to simulate this setup was roughly the same as that for simulating 20 and 70 MeV neutrinos. The 500 MeV case is already unrealistic to study using the discretization codes, as the running time would grow by a factor of $> (500/70)^3 \simeq 400$. On the other hand, the running time of DSMC is roughly the same as for the 70 MeV scenario.

C. Decays of long-lived SM particles

Let us now proceed with a more complicated case, when neutrinos are not injected directly in the decay chain but emerge via the evolution of heavy primary decay products Y, which may be muons or long-lived mesons such as $\pi^\pm, K^\pm, K_L$.

In the primordial plasma, Ys experience a nontrivial evolution once being injected. The interactions they are involved in include kinetic energy loss, interactions with nucleons, annihilation with themselves, and decays, see [80] for more details. The decay products generically involve neutrinos. This evolution influences their energy distribution among the neutrino and EM sectors.

Our approach for simulating this redistribution is the following. We first inject these particles into the plasma and then decay them using Monte Carlo techniques. For the case of charged decay products, we transfer all of the kinetic energy to the EM plasma and then decay them at rest. This is because the energy loss rate is much faster than any other relevant process in the MeV plasma. This simplified description follows the state-of-the-art studies [36,39]; the rest of the interactions discussed above will be added in the future. To simulate the phase space of the decay chain, we use SensCalc [83], a tool calculating the event rate with the decaying LLPs at various laboratory experiments. It contains a module handling LLP decay chains and, in particular, the decays of different SM particles. We have modified it to incorporate the evolution of mesons and muons in the primordial plasma. In absolutely the same way, it may be used to simulate decays of the LLPs, with these mesons appearing among the final states.

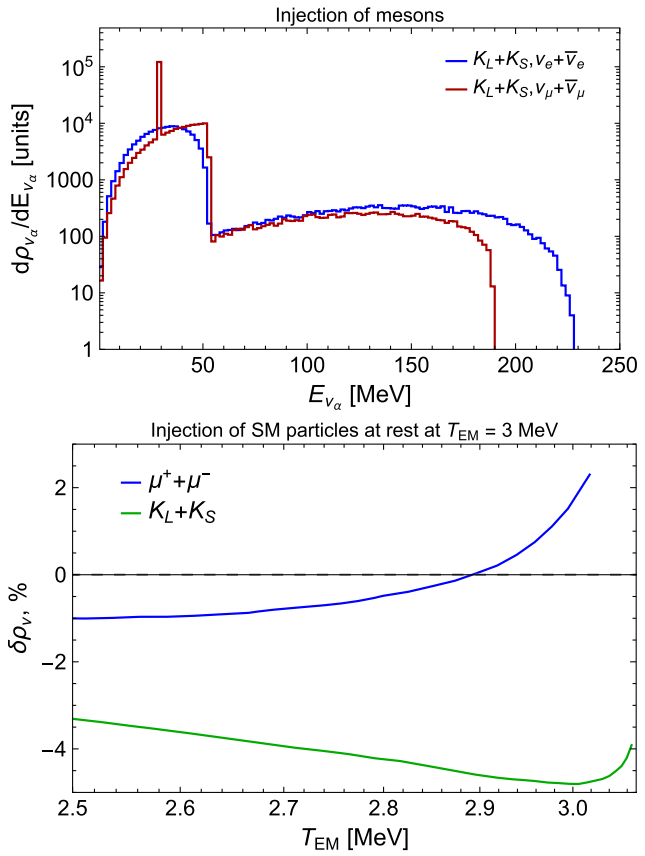

FIG. 7. Impact of injection of heavy SM particles in the primordial plasma. Top panel: the distribution of electron and muon neutrinos produced by decays of $K_L K_S$ pairs. When simulating their decay, we used the module of SensCalc tool [83]. For the chain of the decay products, we account for instant kinetic energy loss by charged particles. The continuous extension of the spectrum to ≃200 MeV is caused by the direct decay of kaons into neutrinos. The increase at $E_\nu = 50$ MeV follows from decays of secondary muons stopped in the plasma, whereas the sharp increase at $E_\nu \approx 34$ MeV originates from decays of secondary pions. Bottom panel: the evolution of the quantity $\delta\rho_\nu$ under the injection of $\mu^+\mu^-$ (the blue curve) and $K_L K_S$ (the green curve) in the primordial plasma at temperature $T_{\mathrm{EM}} = 3$ MeV. The curves start at different temperatures $T_{\mathrm{EM}} \neq 3$ MeV because EM decays of these particles reheat the EM plasma.

The neutrinos from Y decays have a nontrivial spectrum. For instance, the neutrino distribution from decays of neutral kaons $K_L + K_S$ is shown in Fig. 7. They have the following main decay modes:

$$K_S \to 2\pi^0, \quad K_S \to \pi^+\pi^-, \tag{37}$$

$$K_L \to 3\pi^0, \quad K_L \to \pi^+\pi^-\pi^0, \quad K_L \to \pi^\pm l^\mp \nu_l. \tag{38}$$

The neutral pions instantly decay into photons, just heating the EM plasma, whereas $\pi^\pm, \mu^\pm$ particles lose kinetic energy before decaying:

$$\pi^+ \to \mu^+ \nu_\mu, \quad \mu^+ \to e^+ \nu_e \bar{\nu}_\mu. \tag{39}$$

The spectrum of neutrinos from all of these particles has the high-energy part with $E_\nu \gg T_{\rm EM} = \mathcal{O}(1{\rm MeV})$, and we expect the same behavior of $\delta\rho_\nu$ as in the case of the injections of high-energy neutrinos. Clarifying this question is important since many past studies [29,36,37] treated these injections using the semianalytic integrated Boltzmann approach (see, however, Refs. [39,42,59]).

The evolution of $\delta\rho_\nu$ under the injection of $\mu^+\mu^-$ and $K_L K_S$ is shown in Fig. 7. Let us start with the case of the muons. They inject $1/3$ of their energy into the EM plasma, with the rest going to the nonthermal neutrino population. Completely similar to the instant neutrino injection case, $\delta\rho_\nu$, being initially positive, instantly decreases below the ΛCDM value. This finding contradicts the studies [36,37], which considered the scenario of decays of Higgs-like scalars into two muons and found that it increases $N_{\rm eff}$ even in the regime of small scalar lifetimes $\mathcal{O}(0.1\ {\rm s})$.

The $K_L K_S$ case is also interesting. Decaying, they put most of their energy into the EM plasma sector, so we start with a negative $\delta\rho_\nu$. However, the presence of very high-energy neutrinos with $E_\nu = 100$–200 MeV leads to a further slight drop of $\delta\rho_\nu$, and then it tries to approach the equilibrium.

VIII. CONCLUSIONS

Upcoming CMB observations will reach unprecedented precision, which may be used to discover or constrain new physics that was present in the primordial plasma at temperatures as large as a few MeV. To reach this goal, we have to understand the dynamics of the Early Universe in the presence of new physics. It requires solving the neutrino Boltzmann equation across a variety of scenarios, including long-lived relics, nonstandard neutrino interactions, and lepton asymmetry in the neutrino sector.

Current state-of-the-art methods are limited in scope and face computational challenges when neutrino evolution deviates significantly from the standard scenario. These limitations arise from the complex phase space of interactions, the presence of high-energy neutrinos, and the lack of analytic matrix elements—features that are common in systems with new physics. Furthermore, the complexity of implementing these methods makes it difficult to extend them to include various new physics models, even within the range of applicability.

In this paper, we have presented an approach that is potentially free from all these limitations. It is based on the direct simulation Monte Carlo method to solve the Boltzmann equation, see Sec. IV. The traditional version of the DSMC approach is applied to rarefied gases and cannot be used to study the Early Universe. Fundamental modifications are required, such as including the Universe expansion, the hierarchy between weak and electromagnetic interaction rates, the Pauli principle, neutrino oscillations, and the presence of decaying particles. We have discussed these features and how to include them in the DSMC in Sec. V.

In Sec. VI, we have described our current proof-of-principle implementation of the DSMC approach for neutrinos that incorporates these modifications. We have validated it by conducting cross-checks within well-understood physics scenarios (see also Appendix D). In Sec. VII, we have demonstrated the performance and flexibility of DSMC by applying its prototype to several toy scenarios that mimic real scenarios: the equilibration of the neutrinos and EM plasma initially having different temperatures (Sec. VII A), injection of high-energy neutrinos (Sec. VII B), and injection of metastable particles including muons, pions, and kaons, which have a complicated decay chain including neutrinos (Sec. VII C).

Using these simple scenarios, we have found that the instant injection of high-energy neutrinos into a plasma with temperature $T_{\rm EM} \gtrsim 1$ MeV always decreases the neutrino-to-electromagnetic energy densities ratio, which leads to a negative change in $N_{\rm eff}$ compared to the standard cosmological scenario (see, in particular, Figs. 5 and 7). Being extended to the case of continuous neutrino injections in our companion Letter [56], this finding resolves the previously existing discrepancy between different state-of-the-art approaches in predictions about the dynamics of $N_{\rm eff}$ in the presence of high-energy neutrinos.

Our current neutrino DSMC code is rather proof of principle, limited by the efficiency of the implementation and some approximations. Once these problems are overcome, it will result in a powerful independent method of solving neutrino Boltzmann equations. We leave this for future work.

ACKNOWLEDGMENTS

We thank Stefan Stefanov for the in-depth review of the implementation of the proof-of-principle DSMC approach for neutrinos, and Fabio Peano, Luís Olivera e Silva, and Kyrylo Bondarenko for discussions at the early stages of this project. We also thank Kensuke Akita for helping with cross-checks of the DSMC approach, in particular, for adapting his approach to solve the neutrino Boltzmann

equations for our needs. M. O. received support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement No. 860881-HIDDeN.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX A: SCALING OF THE COMPUTATIONAL TIME OF THE DISCRETIZATION APPROACH

Let us first repeat the formula (20) for the scaling of the computational time of the discretization approach. We are interested in the domain from $t(T_{\rm ini})$ to $t(T_{\rm fin})$, when neutrinos with the energy $E_{\nu,\rm max}$ are injected. The time is

$$t_{\rm computation} \propto N_y^{k+1} \times N_{\rm time\ step}, \tag{A1}$$

where N_y is the number of comoving momentum bins, and $N_{\rm time\ step}$ is the number of time steps covering the time domain. k is the minimal possible dimensionality of the collision integral under analytic reduction, Eq. (21).

The power N_y^{k+1} follows from the following considerations. Given N_y bins, we have N_y equations governing the evolution of the corresponding distribution modes. Next, each of the equations contains the collision integral, which, under the discretization, is represented as the product of k connected summations over momentum modes:

$$\mathcal{I}_{{\rm coll},\nu_\alpha} = \int \prod_{i=1}^{k} d\xi_i F(\{\xi\}) = \prod_{i=1}^{k} \sum_{y_i}^{N_y} \Delta\xi_i F(\{\xi\}). \tag{A2}$$

To obtain the scaling $t_{\rm computation}(E_{\nu,\rm max})$, we need to relate N_y and $N_{\rm time\ step}$ to $E_{\nu,\rm max}$.

(i) In practice, for an arbitrary new physics model, the only robust choice of the binning is linear: $N_y = E_{\nu,\rm max}/\Delta E$, with ΔE being the bin width (see a discussion below). ΔE must be kept constant to preserve the accuracy throughout the neutrino evolution.

(ii) Next, the time step must be sufficiently small to resolve the neutrino equilibration rate. Consider the neutrino with the highest possible energy $E_{\nu,\rm max}$. Its thermalization rate may be estimated using the Fermi theory as $\Gamma \sim n_\nu \cdot G_F^2 \langle s \rangle \sim G_F^2 T^4 E_{\nu,\rm max}$, where T is the plasma temperature. Therefore, as far as high-energy neutrinos are present in the plasma, the corresponding time step scales as $\Delta t \sim \Gamma^{-1} \propto E_{\nu,\rm max}^{-1}$. It means that to cover some fixed domain of time from $t(T_{\rm in})$ and down to some moment $t(T_{\rm fin})$, one would need $N_{\rm time\ step} \propto E_{\nu,\rm max}$ time steps.

Therefore, the complexity grows as

$$t_{\rm computation} \propto E_{\nu,\rm max}^{k+2}. \tag{A3}$$

Let us now briefly return to the choice of the binning. In principle, one may consider a different grid structure other than linear, e.g., logarithmic, or more exotic choices, see, e.g., [47]. However, they would generically cause problems with energy conservation, accuracy, and stability for any beyond-the-standard scenario [39,59]. This is because, throughout the evolution, different comoving modes are populated mostly. The reason is that the spectrum of decaying LLPs is fixed in terms of physical momentum p, but varies in time if switching to the comoving momenta $y = a(t) \cdot p$. It is very unrealistic in practice to find an adjustment for a generic LLP: it would arbitrarily modify the dependence $a(t)$ and also decay into neutrinos with different energy distributions.

Another issue of the logarithmic grid is when there are two-body decays into neutrinos. Their energy distribution is just a δ function. In the comoving space, its argument moves towards different momenta y. Any binning other than linear would harm the accuracy when trying to resolve this peak.

APPENDIX B: DETAILS ON THE DISCRETIZATION APPROACH

The discretization approach we use to compare with DSMC is discussed in Ref. [59] (see also Ref. [7]). It utilizes NLO QED corrections and includes three-flavor neutrino oscillations following Ref. [39].

Within the solver, we first introduce the following dimensionless variables:

$$x = m_e a, \quad y = pa, \quad z = T_{\rm EM} a, \tag{B1}$$

normalizing $z \to 1 (a \to 1/T_{\rm EM})$ at the high-temperature limit. The quantities x, y, z characterize time, momentum, and photon temperature, respectively. Then, we discretize the comoving momentum. The discretization is linear: for the neutrino (electromagnetic) momentum grid y_i, we use 200 (80) grid points with $y_{\rm min} = 0.01(0.01)$ and $y_{\rm max} = \max[a_{\rm stop} m_X/2, 40](40)$, where $a_{\rm stop}$ is the estimate of the final scale factor, and m_X is the LLP's mass.

The Boltzmann solver is written in Python with SciPy, NumPy, and Numba libraries as in [8]. To solve the ordinary differential equations on the discretized neutrino modes, we use the LSODA method in SOLVE_IVP distributed in SciPy. By considering the setups with continuous injections of neutrinos of energy $E_{\nu,\rm max}$, we have recovered the approximate scaling (20).[12]

[12]The scaling is slightly worse because of the need for computing the Jacobian of the system of ordinary differential equations computed within LSODA.

APPENDIX C: CONVERGENCE OF THE DSMC ALGORITHM

In order to have robust DSMC simulations, we have to use large enough numbers of simulated particles N and particles per cell, $N_{\rm cell}$ (here and below, we assume that the computational particles match the physical particles). The former is needed to overcome the Monte Carlo noise—random fluctuations of macroscopic observables around their expected values. The latter is crucial because the simulation in each cell involves the calculation of temperatures of the EM plasma and effective neutrino temperatures. If the number of particles of the given type (EM particles, or neutrinos ν_α) per cell is too small, the temperature may have large statistical fluctuations, adding noise on top.

In Fig. 8, we test the system's behavior with neutrinos and EM particles that are initially in perfect equilibrium. We consider two setups:

(1) The one with $N_{\rm cell} = 400$ and N varying from $N = 6 \times 10^4$ to $N = 3 \times 10^7$.
(2) Another one with $N = 5 \times 10^3$ and the number of neutrinos per cell $N_{\nu,\text{per cell}}$ ranging from 30 to 400.

Our goal is to define the setup with the noise at the subpercent level, which keeps the system in the dynamic equilibrium.

For the first setup, we observe the random fluctuations of the quantity $\delta\rho_\nu$, defined by Eq. (35), within 2%–3% around zero. The simulation with $N = 3 \times 10^7$ has fluctuations at the level $\mathcal{O}(0.1\%)$, which roughly corresponds to the scaling of the fluctuations as $1/\sqrt{N}$. It is well enough for our purposes.

We can reach large N either considering a single DSMC simulation with this N, or, equivalently, averaging over n simulations with the number of particles N/n. This flexibility allows using DSMC even on laptops without large RAM and simultaneously accumulating large N.

For the second setup, considering small values of $N_{\rm cell}$, we not only gain additional fluctuations but constantly drive $\delta\rho_\nu$ towards negative values. Its origin is rounding the number of the EM particles after updating the local EM cell temperature. If this number is tiny (which is the case when $N_{\rm cell}$ is small), rounding causes a statistically significant effect. The problem gradually disappears once $N_{\rm cell} \gtrsim 100$. We will consider $N_{\rm cell} = 400$, because it provides a balance between the performance of the code and the quality of statistical sampling.

Having defined the stable setup in the perfect equilibrium case, we can now pose the question of whether it is stable if injecting high-energy neutrinos. Let us assume that these neutrinos have the energy $E_{\nu,\rm inj}$ carry the fraction Δ of the energy density of the thermal plasma; in practice, Δ is fixed by the initial energy density of the decaying LLP. Fixing the number of thermal particles N and estimating their mean energy as $\langle E\rangle = 3.15 \cdot T_{\rm EM}$, we can derive the number of injected neutrinos $N_{\rm injected}$:

$$N_{\rm injected} = \frac{3.15 \cdot \Delta \cdot N \cdot T_{\rm EM}}{E_{\nu,\rm inj}}. \tag{C1}$$

For the given $N, T_{\rm EM}, E_{\nu,\rm inj}, N_{\rm injected}$ has to be large enough. Otherwise, fluctuations are possible: if the number of injected neutrinos is too small, they may all annihilate into the EM plasma particles during the first interaction.

In practice, we have found that for $N \gtrsim 10^6$, the number $N_{\rm inj}$ for which the fluctuations are manageable is $N_{\rm inj} \gtrsim 100$. For instance, for $T_{\rm EM} = 3$ MeV and $\Delta = 0.05$, we satisfy this requirement as far as $E_{\nu,\rm inj} \lesssim 10$ GeV.

Even if we cross this extreme limit, we may overcome fluctuations if increasing N *linearly* with $E_{\nu,\rm inj}$. In this case, the scaling of the computational time of the DSMC algorithm would be $t_{\rm computation} \propto E^2_{\nu,\rm inj}$, where one power comes from the number of time steps and another from

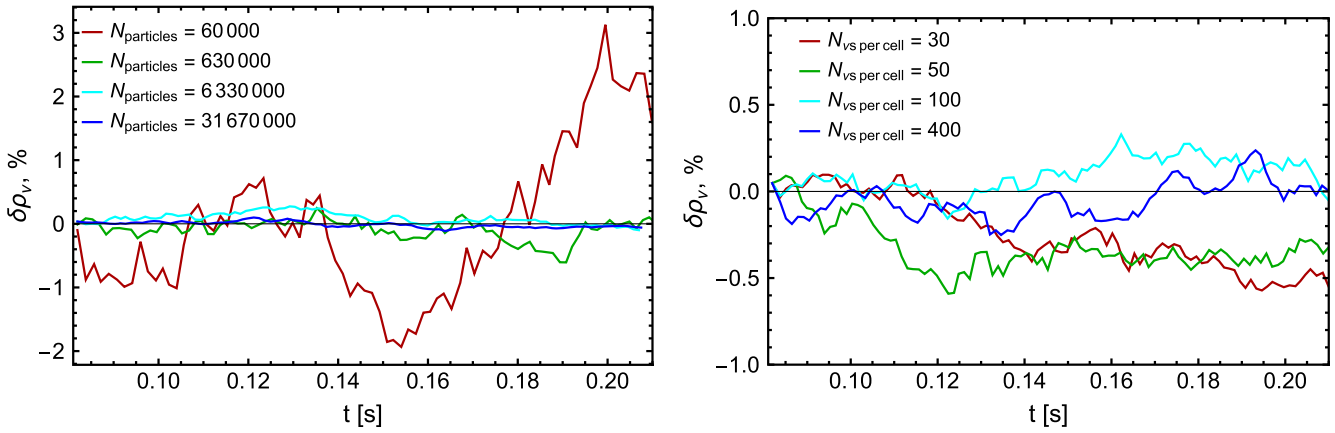

FIG. 8. The temporal evolution of the quantity $\delta\rho_\nu$ when varying numbers of neutrinos per cell $N_{\rm cell,\nu}$ and particles in the system N. Left panel: fixing $N_{\rm cell,\nu} = 400$ and varying N. Right panel: fixing $N = 3 \times 10^6$ and varying $N_{\rm cell,\nu}$.

increasing N. This is still way better than the scaling (A3) of the discretization approach, which is *at least* $\propto E^4_{\nu,\mathrm{inj}}$.

APPENDIX D: CROSS-CHECKS

1. Approaching thermal equilibrium

To test whether the DSMC simulation brings the system of neutrinos and EM particles to the dynamical equilibrium defined by Eqs. (33) and (34), we will use the following setup:

(i) The Universe contains neutrinos and antineutrinos of all flavors together with electrons, positrons, and photons.
(ii) The expansion of the Universe is absent. Therefore, the total energy density of the system is constant, the particles' momenta do not experience redshift and are subject only to their interactions. Such an assumption is needed to allow neutrinos to thermalize fully, making interpreting the results more transparent.
(iii) The initial distribution function of neutrinos consists of two components that are the same for all flavors:
  (1) The equilibrium component, which has Fermi-Dirac distribution with the temperature $T^{\mathrm{ini}}_\nu = 3$ MeV.
  (2) The nonequilibrium component—neutrinos with an arbitrary energy distribution, with the energy density constituting some fraction $\ll 1$ of the equilibrium energy density.

The first subscenario we consider is where there are no nonequilibrium neutrinos, so the system is initially in the equilibrium state. If at least one component of the DSMC simulation is implemented incorrectly, the system will escape the equilibrium, tending to the false ground state. A prominent example is when the cross sections are taken to be velocity independent; then, the distribution of the system tends to the fake-equilibrium spectrum $d\rho_\nu/dE_\nu \sim E^2_\nu \times f_{\mathrm{FD}}$ instead of the correct $E^3_\nu f_{\mathrm{FD}}$ (for relativistic particles with Boltzmann statistics, such an issue has been encountered and explained in [77]). Another issue may be if the maximal interaction weight $\omega_{\max}$ in the acceptance criterion of the pair's interaction (25) is not actually the maximal one. Then, the system falls into the state with $\delta\rho_\nu < 0$.

Our DSMC implementation passes this test, see Fig. 9.

Next, we consider two nontrivial initial conditions: different temperatures of neutrinos and EM particles, and the addition of nonequilibrium neutrinos. The relaxation of the neutrino distribution to the equilibrium one for such scenarios is shown in Fig. 10. Its results are in perfect agreement with the theoretical expectations.

2. Energy transition rates

In this scenario, we will reproduce the semianalytical result of [28,63], where the evolution of neutrinos in the Early Universe was studied under an assumption that every moment of time, the shape of their energy distribution is thermal. The energy transition rates were calculated analytically in terms of the temperatures of neutrinos and EM plasma T_{ν_α}, T. The Boltzmann equations are reduced to the simple system of differential equations on T_{ν_α}, T. For our simulation, the following setup will be used:

(i) The Universe's content is neutrinos and antineutrinos of all flavors together with electrons, positrons, and photons.
(ii) The simulation is altered such that neutrino distributions always have the shape (33) at each simulation step. Basically, we treat neutrinos in exactly the same way as the EM particles in the full DSMC simulation.
(iii) The expansion of the Universe is not included to concentrate on the energy exchange rates.
(iv) As in the whole study, the electron mass is set to zero.

The example of the resulting evolution of the energy density of the neutrino plasma is presented in Fig. 11, where the almost perfect correspondence between theoretical predictions and simulation can be seen. Such reproduction of the energy evolution behavior confirms that averaged energy transition rates are computed correctly.

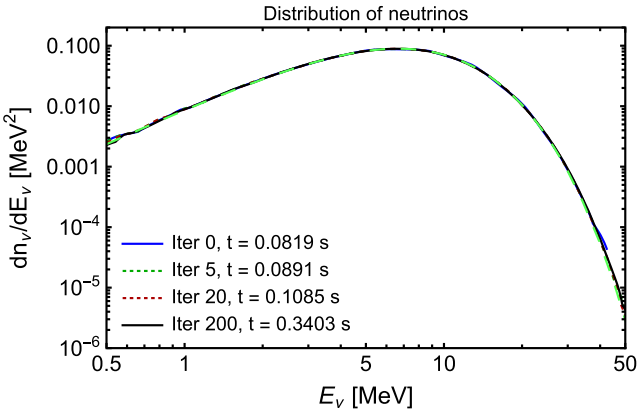

FIG. 9. The evolution of the neutrino distribution dn_ν/dE_ν averaged over all flavors under the assumption of fully equilibrium initial conditions (33) and (34). The "Iter no." curves correspond to the number of the iteration. No significant changes are developed throughout the simulation. The minor changes are related to the quality of the sampler of the kinematics of the electrons via the Fermi-Dirac distribution. The dashed green line shows the analytic Fermi-Dirac distribution with the temperature equal to the temperature of the electromagnetic plasma T_{EM}.

3. Expansion and decoupling

In the third cross-check, we will follow the previous setup, but with the expansion of the Universe included. Because of initial difference between temperatures of neutrino and EM plasma, we expect some remaining

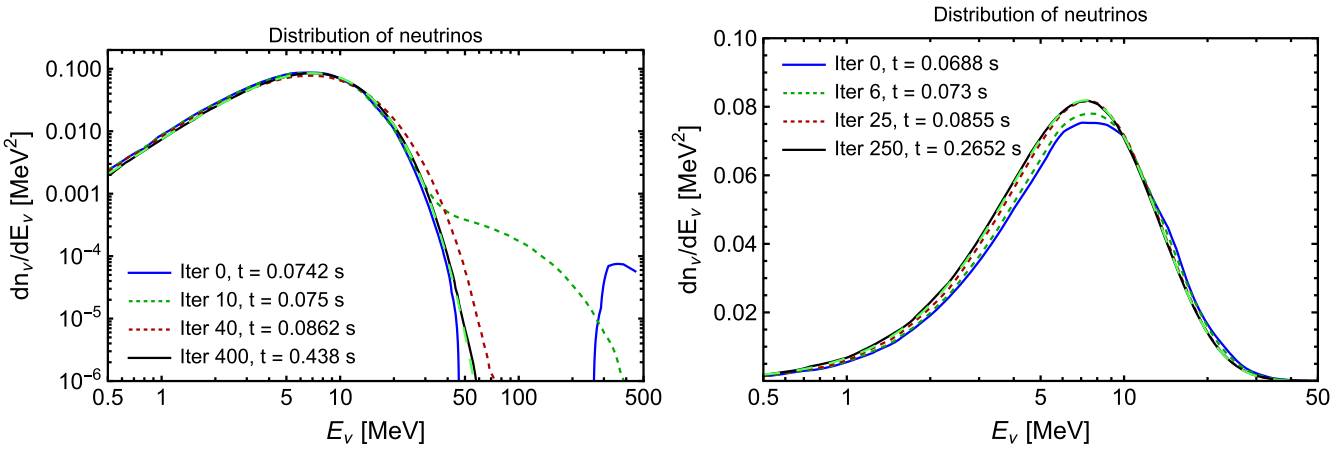

FIG. 10. Evolution of the neutrino distribution function dn_ν/dE_ν averaged over all flavors under different initial setups, showing how DSMC drives it towards thermal equilibrium with the EM sector. Left panel: with equilibrium neutrinos and EM plasma at temperature $T_{\rm EM} = 3$ MeV and nonequilibrium neutrinos with energies randomly distributed in the range 300 MeV $< E_\nu <$ 450 MeV. Their total energy density is related to the total energy of the equilibrium part as $\rho_{\nu_\alpha}^{\rm non-eq}/\rho_\nu^{\rm eq} = 0.15$. The nonequilibrium part of the spectra rapidly loses its energy in the first steps of simulation, leading to the distortions of the spectra at high energies, which are eventually equilibrated. The plot shows the snapshots of the binned neutrino distribution function as obtained at different iterations of the DSMC simulation. The iteration 0 corresponds to the initial setup, while the iteration 400 is the final state. For comparison, the long-dashed green line shows the Fermi-Dirac distribution $dn_\nu/dE_\nu = E_\nu^2 f_{\rm FD}(E_\nu, T_{\rm EM,final})$, being the thermal equilibrium of neutrinos with the EM plasma with the final temperature $T_{\rm EM,final} \approx 3.15$ MeV. Right panel: with equilibrium neutrinos having temperature $T_{\nu_\alpha} = 3.5$ MeV and EM plasma at temperature $T_{\rm EM} = 3$ MeV. The meaning of the lines is the same, while the number of iterations is 250.

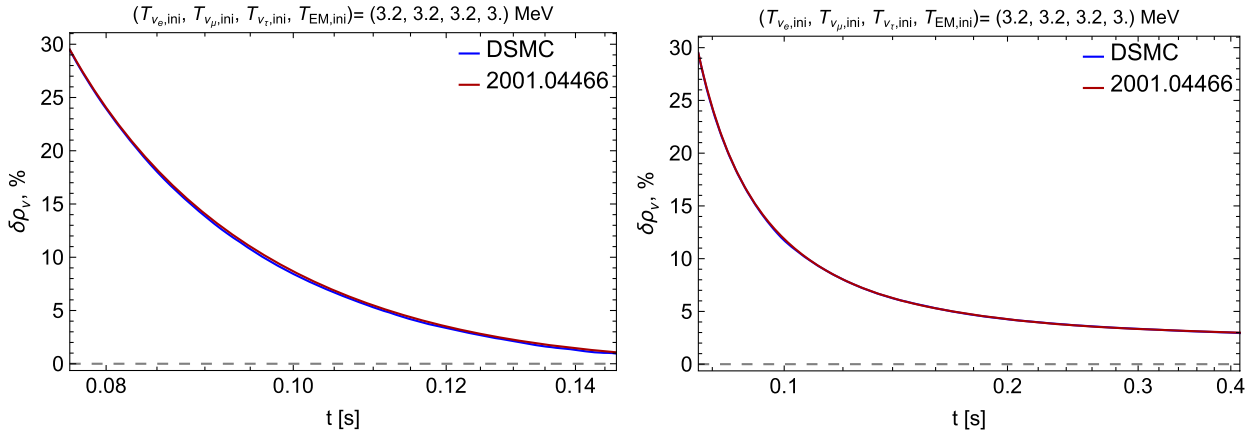

FIG. 11. The evolution of the ratio of the neutrino energy density to the EM energy density in DSMC simulation compared to the theoretical prediction from [28], under an assumption that the shape of the neutrino distribution function is always thermal at each step of the simulation. The initial conditions for the setup are $T_{\nu_i} = 3.2$ MeV for every flavor and the temperature of the EM plasma is $T_{\rm EM} = 3$ MeV. Left panel: not including the expansion of the Universe. Because of the absence of expansion, the ratio approaches to the exact SM value. Right panel: expansion included.

inequality between them, since the start of the simulation occurs close to the temperature of the neutrino decoupling. In similar terms, we present the example of such comparison in Fig. 11.

[1] S. Bashinsky and U. Seljak, Neutrino perturbations in CMB anisotropy and matter clustering, Phys. Rev. D **69**, 083002 (2004).

[2] D. Baumann, F. Beutler, R. Flauger, D. Green, A. Slosar, M. Vargas-Magaña, B. Wallisch, and C. Yèche, First constraint on the neutrino-induced phase shift in the spectrum of baryon acoustic oscillations, Nat. Phys. **15**, 465 (2019).

[3] J. Alvey, M. Escudero, N. Sabti, and T. Schwetz, Cosmic neutrino background detection in large-neutrino-mass cosmologies, Phys. Rev. D **105**, 063501 (2022).

[4] G. Mangano, G. Miele, S. Pastor, and M. Peloso, A Precision calculation of the effective number of cosmological neutrinos, Phys. Lett. B **534**, 8 (2002).

[5] J. J. Bennett, G. Buldgen, M. Drewes, and Y. Y. Y. Wong, Towards a precision calculation of the effective number of neutrinos $N_{\rm eff}$ in the standard model I: The QED equation of state, J. Cosmol. Astropart. Phys. 03 (2020) 003; 03 (2021) A01.

[6] J. J. Bennett, G. Buldgen, P. F. De Salas, M. Drewes, S. Gariazzo, S. Pastor, and Y. Y. Y. Wong, Towards a precision calculation of $N_{\rm eff}$ in the standard model II: Neutrino decoupling in the presence of flavour oscillations and finite-temperature QED, J. Cosmol. Astropart. Phys. 04 (2021) 073.

[7] K. Akita and M. Yamaguchi, A precision calculation of relic neutrino decoupling, J. Cosmol. Astropart. Phys. 08 (2020) 012.

[8] J. Froustey, C. Pitrou, and M. C. Volpe, Neutrino decoupling including flavour oscillations and primordial nucleosynthesis, J. Cosmol. Astropart. Phys. 12 (2020) 015.

[9] M. Cielo, M. Escudero, G. Mangano, and O. Pisanti, $N_{\rm eff}$ in the standard model at NLO is 3.043, Phys. Rev. D **108**, L121301 (2023).

[10] M. Drewes, Y. Georis, M. Klasen, L. P. Wiggering, and Y. Y. Y. Wong, Towards a precision calculation of $N_{\rm eff}$ in the standard model III: Improved estimate of NLO corrections to the collision integral, J. Cosmol. Astropart. Phys. 06 (2024) 032.

[11] O. Pisanti, A. Cirillo, S. Esposito, F. Iocco, G. Mangano, G. Miele, and P. D. Serpico, PArthENoPE: Public algorithm evaluating the nucleosynthesis of primordial elements, Comput. Phys. Commun. **178**, 956 (2008).

[12] C. Pitrou, A. Coc, J.-P. Uzan, and E. Vangioni, Precision big bang nucleosynthesis with improved Helium-4 predictions, Phys. Rep. **754**, 1 (2018).

[13] N. Aghanim *et al.* (Planck Collaboration), Planck 2018 results. I. Overview and the cosmological legacy of Planck, Astron. Astrophys. **641**, A1 (2020).

[14] N. Aghanim *et al.* (Planck Collaboration), Planck 2018 results. VI. Cosmological parameters, Astron. Astrophys. **641**, A6 (2020).

[15] Y. I. Izotov, T. X. Thuan, and N. G. Guseva, A new determination of the primordial He abundance using the He I λ10830 Å emission line: cosmological implications, Mon. Not. R. Astron. Soc. **445**, 778 (2014).

[16] E. Aver, K. A. Olive, and E. D. Skillman, The effects of He I λ10830 on helium abundance determinations, J. Cosmol. Astropart. Phys. 07 (2015) 011.

[17] A. Peimbert, M. Peimbert, and V. Luridiana, The primordial helium abundance and the number of neutrino families, Rev. Mex. Astron. Astrofis. **52**, 419 (2016).

[18] V. Fernández, E. Terlevich, A. I. Díaz, R. Terlevich, and F. F. Rosales-Ortega, Primordial helium abundance determination using sulphur as metallicity tracer, Mon. Not. R. Astron. Soc. **478**, 5301 (2018).

[19] M. Valerdi, A. Peimbert, M. Peimbert, and A. Sixtos, Determination of the primordial helium abundance based on NGC 346, an H II region of the small magellanic cloud, Astrophys. J. **876**, 98 (2019).

[20] E. Aver, D. A. Berg, A. S. Hirschauer, K. A. Olive, R. W. Pogge, N. S. J. Rogers, J. J. Salzer, and E. D. Skillman, A comprehensive chemical abundance analysis of the extremely metal poor Leoncino Dwarf galaxy (AGC 198691) Mon. Not. R. Astron. Soc. **510**, 373 (2021).

[21] A. Matsumoto *et al.*, EMPRESS. VIII. A new determination of primordial he abundance with extremely metal-poor galaxies: A suggestion of the lepton asymmetry and implications for the Hubble tension, Astrophys. J. **941**, 167 (2022).

[22] M. Archidiacono and S. Hannestad, Updated constraints on non-standard neutrino interactions from Planck, J. Cosmol. Astropart. Phys. 07 (2014) 046.

[23] F. Forastieri, M. Lattanzi, and P. Natoli, Constraints on secret neutrino interactions after Planck, J. Cosmol. Astropart. Phys. 07 (2015) 014.

[24] P. F. de Salas, S. Gariazzo, P. Martínez-Miravé, S. Pastor, and M. Tórtola, Cosmological radiation density with non-standard neutrino-electron interactions, Phys. Lett. B **820**, 136508 (2021).

[25] Y. Du and J.-H. Yu, Neutrino non-standard interactions meet precision measurements of $N_{\rm eff}$, J. High Energy Phys. 05 (2021) 058.

[26] A. D. Dolgov, S. H. Hansen, S. Pastor, S. T. Petcov, G. G. Raffelt, and D. V. Semikoz, Cosmological bounds on neutrino degeneracy improved by flavor oscillations, Nucl. Phys. **B632**, 363 (2002).

[27] E. Grohs, G. M. Fuller, C. T. Kishimoto, and M. W. Paris, Lepton asymmetry, neutrino spectral distortions, and big bang nucleosynthesis, Phys. Rev. D **95**, 063503 (2017).

[28] M. Escudero Abenza, Precision early universe thermodynamics made simple: $N_{\rm eff}$ and neutrino decoupling in the standard model and beyond, J. Cosmol. Astropart. Phys. 05 (2020) 048.

[29] G. B. Gelmini, M. Kawasaki, A. Kusenko, K. Murai, and V. Takhistov, Big bang nucleosynthesis constraints on sterile neutrino and lepton asymmetry of the Universe, J. Cosmol. Astropart. Phys. 09 (2020) 051.

[30] J. Froustey and C. Pitrou, Primordial neutrino asymmetry evolution with full mean-field effects and collisions, J. Cosmol. Astropart. Phys. 03 (2022) 065.

[31] M. Escudero, A. Ibarra, and V. Maura, Primordial lepton asymmetries in the precision cosmology era: Current status and future sensitivities from BBN and the CMB, Phys. Rev. D **107**, 035024 (2023).

[32] J. Froustey and C. Pitrou, Constraints on primordial lepton asymmetries with full neutrino transport, Phys. Rev. D **110**, 103551 (2024).

[33] A. D. Dolgov, S. H. Hansen, G. Raffelt, and D. V. Semikoz, Heavy sterile neutrinos: Bounds from big bang nucleosynthesis and SN1987A, Nucl. Phys. **B590**, 562 (2000).
[34] S. Hannestad, What is the lowest possible reheating temperature?, Phys. Rev. D **70**, 043506 (2004).
[35] O. Ruchayskiy and A. Ivashko, Restrictions on the lifetime of sterile neutrinos from primordial nucleosynthesis, J. Cosmol. Astropart. Phys. 10 (2012) 014.
[36] A. Fradette and M. Pospelov, BBN for the LHC: Constraints on lifetimes of the Higgs portal scalars, Phys. Rev. D **96**, 075033 (2017).
[37] A. Fradette, M. Pospelov, J. Pradler, and A. Ritz, Cosmological beam dump: Constraints on dark scalars mixed with the Higgs boson, Phys. Rev. D **99**, 075004 (2019).
[38] A. Boyarsky, M. Ovchynnikov, O. Ruchayskiy, and V. Syvolap, Improved big bang nucleosynthesis constraints on heavy neutral leptons, Phys. Rev. D **104**, 023517 (2021).
[39] N. Sabti, A. Magalich, and A. Filimonova, An extended analysis of heavy neutral leptons during big bang nucleosynthesis, J. Cosmol. Astropart. Phys. 11 (2020) 056.
[40] A. Boyarsky, M. Ovchynnikov, N. Sabti, and V. Syvolap, When feebly interacting massive particles decay into neutrinos: The $N_{\rm eff}$ story, Phys. Rev. D **104**, 035006 (2021).
[41] L. Mastrototaro, P. D. Serpico, A. Mirizzi, and N. Saviano, Massive sterile neutrinos in the early Universe: From thermal decoupling to cosmological constraints, Phys. Rev. D **104**, 016026 (2021).
[42] H. Rasmussen, A. McNichol, G. M. Fuller, and C. T. Kishimoto, Effects of an intermediate mass sterile neutrino population on the early Universe, Phys. Rev. D **105**, 083513 (2022).
[43] P. Ade *et al.* (Simons Observatory Collaboration), The Simons Observatory: Science goals and forecasts, J. Cosmol. Astropart. Phys. 02 (2019) 056.
[44] K. N. Abazajian *et al.* (CMB-S4 Collaboration), CMB-S4 science book, first edition, arXiv:1610.02743.
[45] S. Hannestad and J. Madsen, Neutrino decoupling in the early universe, Phys. Rev. D **52**, 1764 (1995).
[46] E. Grohs, G. M. Fuller, C. T. Kishimoto, M. W. Paris, and A. Vlasenko, Neutrino energy transport in weak decoupling and big bang nucleosynthesis, Phys. Rev. D **93**, 083522 (2016).
[47] S. Gariazzo, P. F. de Salas, and S. Pastor, Thermalisation of sterile neutrinos in the early Universe in the $3+1$ scheme with full mixing matrix, J. Cosmol. Astropart. Phys. 07 (2019) 014.
[48] T. Kanzaki, M. Kawasaki, K. Kohri, and T. Moroi, Cosmological constraints on neutrino injection, Phys. Rev. D **76**, 105017 (2007).
[49] T. Hasegawa, N. Hiroshima, K. Kohri, R. S. L. Hansen, T. Tram, and S. Hannestad, MeV-scale reheating temperature and thermalization of oscillating neutrinos by radiative and hadronic decays of massive particles, J. Cosmol. Astropart. Phys. 12 (2019) 012.
[50] G. A. Bird, Monte Carlo simulation of gas flows, Annu. Rev. Fluid Mech. **10**, 11 (1978).
[51] G. Bird, Monte Carlo simulation of gas flows, Annu Rev Fluid Mech **10**, 11 (2003).
[52] E. Roohi and S. Stefanov, Collision partner selection schemes in DSMC: From micro/nano flows to hypersonic flows, Phys. Rep. **656**, 1 (2016).
[53] S. Stefanov, On the basic concepts of the direct simulation Monte Carlo method, Phys. Fluids **31**, 067104 (2019).
[54] G. Goos, J. Hartmanis, and J. van Leeuwen, Computational science—ICCS 2002, Lect. Notes Comput. Sci. **2331**, 342 (2002); Plasma Phys. Controlled Fusion **50**, 124034 (2008).
[55] M. A. Gallis, J. R. Torczynski, S. J. Plimpton, D. J. Rader, and T. Koehler, Direct simulation Monte Carlo: The quest for speed, Technical Report, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), 2014.
[56] M. Ovchynnikov and V. Syvolap, companion Letter, Primordial neutrinos and new physics: Novel approach to solving neutrino Boltzmann equation, Phys. Rev. Lett. **134**, 101003 (2025).
[57] M. Kawasaki, K. Kohri, and T. Moroi, Big-bang nucleosynthesis and hadronic decay of long-lived massive particles, Phys. Rev. D **71**, 083502 (2005).
[58] A. D. Dolgov, Neutrinos in cosmology, Phys. Rep. **370**, 333 (2002).
[59] K. Akita, G. Baur, M. Ovchynnikov, T. Schwetz, and V. Syvolap, Dynamics of metastable standard model particles from long-lived particle decays in the MeV primordial plasma, arXiv:2411.00931.
[60] B. H. J. McKellar and M. J. Thomson, Oscillating doublet neutrinos in the early universe, Phys. Rev. D **49**, 2710 (1994).
[61] G. Sigl and G. Raffelt, General kinetic description of relativistic mixed neutrinos, Nucl. Phys. **B406**, 423 (1993).
[62] A. Vlasenko, G. M. Fuller, and V. Cirigliano, Neutrino quantum kinetics, Phys. Rev. D **89**, 105004 (2014).
[63] M. Escudero, Neutrino decoupling beyond the standard model: CMB constraints on the dark matter mass with a fast and precise $N_{\rm eff}$ evaluation, J. Cosmol. Astropart. Phys. 02 (2019) 007.
[64] A. D. Dolgov, S. H. Hansen, and D. V. Semikoz, Non-equilibrium corrections to the spectra of massless neutrinos in the early universe, Nucl. Phys. **B503**, 426 (1997).
[65] J. Beacham *et al.*, Physics beyond colliders at CERN: Beyond the standard model working group report, J. Phys. G **47**, 010501 (2020).
[66] G. Bird, Direct simulation and the Boltzmann equation, Phys. Fluids **13**, 2676 (1970).
[67] M. Ivanov and S. Rogasinskii, *Theoretical analysis of traditional and modern schemes of the DSMC method*, Rarefied gas dynamics (VCH Verlagsgesellschaft mbH, Weinheim, Germany, 1991), p. 629.
[68] G. Bird, The velocity distribution function within a shock wave, J. Fluid Mech. **30**, 479 (1967).
[69] K. Koura, Transient Couette flow of rarefied binary gas mixtures, Phys. Fluids **13**, 1457 (1970).
[70] K. Koura, Null-collision technique in the direct-simulation Monte Carlo method, Phys. Fluids **29**, 3509 (1986).
[71] G. Bird, Perception of numerical methods in rarefied gasdynamics, Prog. Astronaut. Aeronaut. **117**, 211 (1989).
[72] E. Roohi and S. Stefanov, Collision partner selection schemes in dsmc: From micro/nano flows to hypersonic flows, Phys. Rep. **656**, 1 (2016).
[73] D. P. Schmidt and C. J. Rutland, A new droplet collision algorithm, J. Comput. Phys. **164**, 62 (2000).

[74] J.-S. Wu and K.-C. Tseng, Analysis of micro-scale gas flows with pressure boundaries using direct simulation Monte Carlo method, Comput. Fluids **30**, 711 (2001).
[75] J.-S. Wu and Y.-Y. Lian, Parallel three-dimensional direct simulation Monte Carlo method and its applications, Comput. Fluids **32**, 1133 (2003).
[76] A. Venkattraman, A. A. Alexeenko, M. Gallis, and M. Ivanov, A comparative study of no-time-counter and Majorant collision frequency numerical schemes in DSMC, AIP Conf. Proc. **1501**, 489 (2012).
[77] F. Peano, M. Marti, L. Silva, and G. Coppa, Statistical kinetic treatment of relativistic binary collisions, Phys. Rev. E **79**, 025701 (2009).
[78] M. C. Gonzalez-Garcia, M. Maltoni, and T. Schwetz, NuFIT: Three-flavour global analyses of neutrino oscillation experiments, Universe **7**, 459 (2021).
[79] C. Bierlich *et al.*, A comprehensive guide to the physics and usage of Pythia 8.3, SciPost Phys. Codebases **2022**, 8 (2022).
[80] K. Akita, G. Baur, M. Ovchynnikov, T. Schwetz, and V. Syvolap, Dynamics of metastable Standard Model particles from long-lived particle decays in the MeV primordial plasma, arXiv:2411.00931; New physics decaying into metastable particles: impact on cosmic neutrinos, arXiv:2411.00892.
[81] A. F. Heckler, Astrophysical applications of quantum corrections to the equation of state of a plasma, Phys. Rev. D **49**, 611 (1994).
[82] N. Fornengo, C. W. Kim, and J. Song, Finite temperature effects on the neutrino decoupling in the early universe, Phys. Rev. D **56**, 5123 (1997).
[83] M. Ovchynnikov, J.-L. Tastet, O. Mikulenko, and K. Bondarenko, Sensitivities to feebly interacting particles: Public and unified calculations, Phys. Rev. D **108**, 075028 (2023).
[84] K. J. Kelly and Y. Zhang, Mononeutrino at DUNE: New signals from neutrinophilic thermal dark matter, Phys. Rev. D **99**, 055034 (2019).
[85] J. E. Kim, Light pseudoscalars, particle physics and cosmology, Phys. Rep. **150**, 1 (1987).
[86] P. Ilten, Y. Soreq, M. Williams, and W. Xue, Serendipity in dark photon searches, J. High Energy Phys. 06 (2018) 004.
[87] K. Bondarenko, A. Boyarsky4, D. Gorbunov, and O. Ruchayskiy, Phenomenology of GeV-scale heavy neutral leptons, J. High Energy Phys. 11 (2018) 032.