Probing new physics in the laboratory and in space

Ovchynnikov, M.

Chapter 3

Probes from cosmology[1]

In this chapter, we consider two earliest messengers from cosmology: BBN and CMB, and study the impact of short-lived FIPs on corresponding observables. In the case of BBN, we concentrate on short-lived particles that decay hadronically, and derive analytically the bound on lifetimes that comes from the impact of mesons on the $p \leftrightarrow n$ conversion in the primordial plasma, see Sec. 3.1. In the case of CMB, we study the impact of short-lived FIPs on the effective number of degrees of freedom, N_{eff}, and in particular show that even if decaying mostly into neutrinos they may decrease N_{eff}, see Sec. 3.2. Finally, we apply the findings to the case of a particular model – HNLs, for which we first study their cosmological population (Sec. 3.3.4), and then derive constraints from BBN (Sec. 3.3.3.1) and CMB (Sec. 3.3.4).

3.1 BBN and hadronically decaying particles

In this section, we discuss bounds from BBN on hadronically decaying particles. We will first discuss the current measurements of the primordial abundance of helium, then derive the bounds from BBN on hadronically decaying particles, and then comment until which lifetimes the bound extends.

3.1.1 Measurements of ^{4}He abundance

Over the last 6 years five works determined the primordial ^{4}He abundance from stellar measurements [171–175]. The formal statistical errors of Y_p are at the level of $1-3\%$, however, the scatter between different groups is larger, see Fig. 1.

All these works determine astrophysical Helium abundance through measurements of recombination emission lines of ^{4}He and H in the metal-poor extragalactic ionized regions, then *linearly* extrapolating the measurements to zero metallicity. Given the high precision of the results, it is important to take into account various smaller effects: including

[1] Results of this chapter are presented in papers [37, 170]. The main contribution of Maksym Ovchynnikov is analytic and numeric estimates of the BBN and CMB bounds.

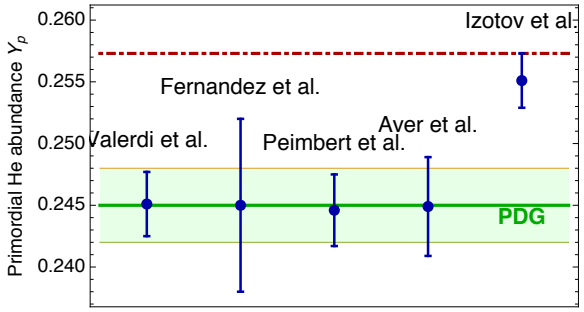

Figure 1: Measurements of Y_p of recent works [171–175]. The green shaded region is the PDG recommended value [52] (with $\pm 1\sigma$). The gray dashed line denotes the SBBN prediction $\bar{Y}_p = 0.247$ from [45]. The red dashed-dotted line is the maximal admissible value $Y_{p,\text{max}}$ on which we base our analysis.

[4]He fluorescent emission, different ion temperatures, spatial temperature fluctuations, and others [176, 177]. Additionally, while it is true that the metallicity and Helium abundance are positively correlated, the linear extrapolation to zero-metallicity may be prone to systematic uncertainties.

The value of Y_p predicted within the framework of SBBN is $\bar{Y}_p = 0.24709 \pm 0.00019$ (see, e.g., [45]). The effect of mesons leads to an increase of Y_p as compared to the SBBN value. Therefore, in order to get a conservative upper bound we assume that the maximally allowed Y_p is given by the 1σ deviation from the maximal value predicted by [171–175], which is $Y_{p,\text{max}} = 0.2573$. Note that this upper value significantly deviates from the PDG-recommended value [52] $Y_{p,\text{max}} = 0.248$ at 1σ. This translates to the bound

$$\frac{\Delta Y_p}{\bar{Y}_p} < 4.35\% \tag{3.1.1}$$

3.1.2 Bound on hadronically decaying particles

Sufficiently heavy FIPs can decay into mesons $h = \pi, K$, *etc*. Charged pions drive the $p \leftrightarrow n$ conversion via processes [178]

$$\pi^- + p \to n + \pi^0/\gamma, \quad \pi^+ + n \to p + \pi^0 \tag{3.1.2}$$

The cross section of these reactions is much larger than the cross section of weak interactions driven conversion processes:

$$\frac{\langle \sigma^{\pi}_{p\leftrightarrow n} v \rangle}{\langle \sigma^{\text{Weak}}_{p\leftrightarrow n} v \rangle} \simeq \frac{1}{G_F^2 m_p^2 T^2} \sim 10^{16} \left(\frac{1\,\text{MeV}}{T}\right)^2, \tag{3.1.3}$$

Large cross section, absence of threshold and isotopic symmetry of these processes mean that if pions are present in the plasma in the amounts at least comparable with that of baryons, they drive the number densities of protons and neutrons to equal values, $n_n/n_p \simeq$

$\langle\sigma^{\pi}_{p\to n}v\rangle/\langle\sigma^{\pi}_{n\to p}v\rangle \simeq 1$.[2] The effect of kaons is qualitatively similar, but leads to a slightly different neutron-to-proton ratio (Appendix 3.B.1).

The impact of this effect on primordial ^{4}He abundance depends on how long mesons remain present in the plasma in significant amounts. Once mesons are created, they can *(i)* scatter and lose energy; *(ii)* decay; *(iii)* participate in $p \leftrightarrow n$ conversion. The corresponding rates are very different: at MeV temperatures and below, $\Gamma^h_{\rm scat} \gg \Gamma^h_{\rm decay} \gg \Gamma^h_{p\leftrightarrow n}$ (see [179]).

The instantaneous number density of mesons is an interplay between their production (via decays of FIPs) and their decays:

$$n_h^{\rm inst} = n_{\rm FIP}(T)\cdot {\rm Br}_{N\to h}\frac{\Gamma_{\rm FIP,dec}}{\Gamma_{h,\rm dec}} = n_N(T)\cdot {\rm Br}_{N\to h}\frac{\tau_h}{\tau_N}. \tag{3.1.4}$$

Here, ${\rm Br}_{{\rm FIP}\to h}$ is the branching of FIPs into mesons. $n_{\rm FIP}(T)$ is the number density of FIPs:

$$n_{\rm FIP}(T) = \left(\frac{a_{\rm dec}}{a(T)}\right)^3\cdot n^{\rm dec}_{\rm FIP}\cdot e^{-\frac{t(T)}{\tau_N}}, \tag{3.1.5}$$

where $n^{\rm dec}_{\rm FIP}$ is the FIP's number density at decoupling (i.e. when their interaction with plasma has been completely stopped), and $a(T)$ ($a_{\rm dec}$) is the scale factor at temperature T (correspondingly, at FIP decoupling).

The number of $p \leftrightarrow n$ reactions per nucleon occurring after time $t \gg \tau_{\rm FIP}$ (or below some corresponding temperature $T(t)$) is thus

$$N^h_{p\leftrightarrow n}(T) = \sum_h \int_{t(T)}^{\infty} dt\; n_h^{\rm inst}(T)\cdot\langle\sigma^h_{p\leftrightarrow n}v\rangle \approx \left(\frac{a_{\rm dec}}{a}\right)^3\frac{n^{\rm dec}_{\rm FIP}}{n_B}\cdot e^{-\frac{t(T)}{\tau_{\rm FIP}}}\cdot {\rm Br}_{{\rm FIP}\to h}\cdot P_{\rm conv}, \tag{3.1.6}$$

where n_B is the baryon number density, the sum goes over meson species and $P_{\rm conv}$ is the probability for a single meson to interact with nucleons before decaying:

$$P_{\rm conv} \simeq \frac{n_B\cdot\langle\sigma^h_{p\leftrightarrow n}v\rangle}{\Gamma^h_{\rm decay}}. \tag{3.1.7}$$

At $\mathcal{O}(1\text{ MeV})$ temperatures, $P_{\rm conv}\sim 10^{-2}-10^{-1}$, see Appendix 3.B.

[2] For each of the processes (3.1.2), there are no inverse reactions. Indeed, π^0 decays very fast, whereas γs quickly lose their energy. Therefore, the conversion (3.1.2) is highly non-equilibrium, and the corresponding value of n_n/n_p is not given by the usual Boltzmann exponent.

The meson driven conversion keeps the value $n_n/n_p \simeq 1$ roughly until a temperature T_0 when the number of reactions per nucleon drops below one,

$$N^h_{p\leftrightarrow n}(T_0) \simeq 1, \tag{3.1.8}$$

and weak SBBN reactions start to relax the n/p ratio down to its SBBN value, see Fig. 2 (left panel).

However, if T_0 is close enough to the freeze-out of weak $p \leftrightarrow n$ processes, occurring roughly at $T_n \simeq 0.8$ MeV, the relaxation is not complete (Fig. 2, right panel). This leads to a positive correction $\Delta(n_n/n_p)$ as compared to the SBBN case, which translates to an increase of the ^{4}He abundance ΔY_p.

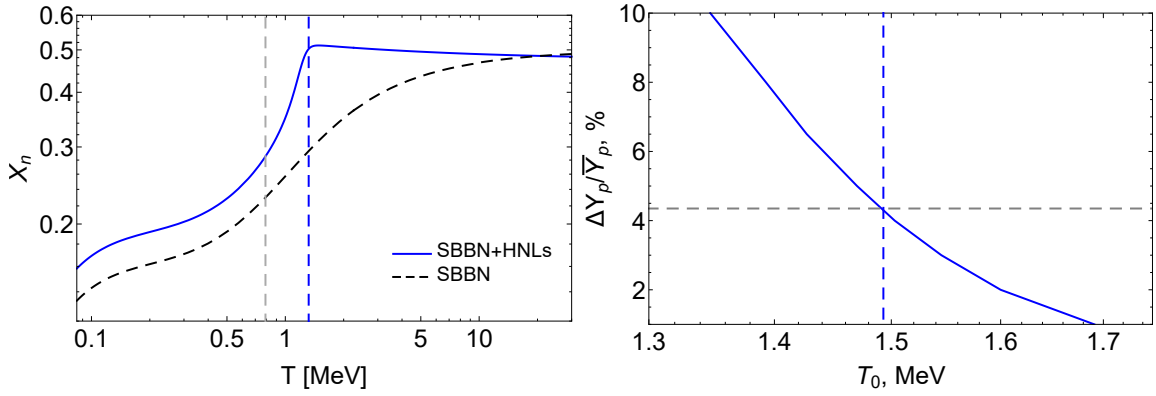

Figure 2: *Left panel*: temperature evolution of the neutron abundance $X_n = n_n/(n_n + n_p)$ in the presence of pions from decays of an HNL with mass $m_N = 400$ MeV and lifetime $\tau_N = 0.03$ s. Below $T \simeq 100$ MeV, pions drive the neutron abundance to $X_n \approx 0.5$. At temperatures $T_0 \simeq 1.3$ MeV (the blue vertical dashed line) pions disappear, and X_n starts relaxing towards its SBBN value but does not reach it. After the neutron decoupling (the gray vertical line) X_n evolves mainly due to the neutron decays. *Right panel*: a relation between the temperature T_0 (defined by Eq. (3.1.8)) and corrections to the ^{4}He abundance, as compared to the SBBN central value $\bar{Y}_p \approx 0.247$. It corresponds to the case of when only charged pions are present in plasma. The gray horizontal line corresponds to maximally allowed correction $\Delta Y_p/\bar{Y}_p = 4.35\%$ that we adopt in this work (see Appendix 3.1.1). The intersection of gray and colored lines defines the temperature $T_0^{\rm min}$.

In this way, the *upper bound* on the ^{4}He abundance $Y_{p,\rm max}$ is translated to the *lower bound* $T_0 \geq T_0^{\rm min}$. Together with the relations (3.1.6)–(3.1.8), this allow us to find an upper limit on the FIP lifetime $\tau_{\rm FIP}$:

$$\tau_{\rm FIP} \lesssim \frac{t(T_0^{\rm min})}{\ln\left[\sum_h \left(\frac{a_{\rm dec}}{a_0}\right)^3 \frac{n^{\rm dec}_{\rm FIP} P_{\rm conv} {\rm Br}_{N\to h}}{n_\gamma(T_0^{\rm min})\eta_B}\right]}. \tag{3.1.9}$$

Here, n_γ is the number density of photons, η_B is the baryon-to-photon ratio, and $t(T)$ is time-temperature relation. $t(T)$ is given by the Standard Model relation: $t(T) = \frac{M_*}{2T^2}$, with $M_* = \frac{M_{\rm Pl}}{1.66\sqrt{g_*}}$ the reduced Planck mass, where $g_*(T) \simeq 10.6$ for $T \simeq 1 - 2$ MeV.[3]

Let us rewrite the logarithmic factor in (3.1.9) as

$$\left(\frac{a_{\rm dec}}{a_0}\right)^3 \frac{n_{\rm FIP,dec}}{n_\gamma(T_0^{\rm min})} = \frac{n_{\rm FIP,dec}}{n_\gamma(T_{\rm dec})} \cdot \zeta, \tag{3.1.10}$$

where $\zeta = \left(\frac{a_{\rm dec} T_{\rm dec}}{a_0 T_0^{\rm min}}\right)^3$ is the "entropy dilution" factor. For example, if FIPs were in thermal equilibrium and decoupled while being ultrarelativistic, $T_{\rm dec} \gg m_{\rm FIP}$, we have $n_{\rm FIP,dec}/n_\gamma(T_{\rm dec}) \approx \mathcal{O}(1)$. The dilution factor ζ is a product of the SBBN value times the value induced by FIPs during their evolution:

$$\zeta = \left(\frac{a_{\rm dec}^{\rm SBBN} T_{\rm dec}}{a_0^{\rm SBBN} T_0^{\rm min}}\right)^3 \times \left(\frac{a_0^{\rm SBBN} T_0^{\rm min}}{a_0 T_0^{\rm min}}\right)^3 \equiv \zeta_{\rm SBBN} \times \zeta_{\rm FIP}, \tag{3.1.11}$$

where we used that $a_{\rm dec} \approx a_{\rm dec}^{\rm SBBN}$; this approximation is valid since at temperatures $T_{\rm dec}$ there are many SM particles, and FIPs only contribute a small fraction to the total energy density of the Universe. In SBBN at temperatures $T \gtrsim 1$ MeV, all particles are at local equilibrium, which defines the dynamics of the scale factor and hence the value of $\zeta_{\rm SBBN}$:

$$\zeta_{\rm SBBN} \approx \frac{g_*(T_0^{\rm min})}{g_*(T_{\rm dec})} \simeq \frac{1}{8}, \tag{3.1.12}$$

where we used that $a_{\rm SBBN}(T) \propto g_*^{-1/3}(T) \cdot T^{-1}$. Decays of heavy FIPs violate the thermal equilibrium at $\mathcal{O}(1\text{ MeV})$, and the scaling (3.1.12) changes. For GeV-scale particles with lifetimes $\tau_{\rm FIP} \sim 0.01\text{ s} - 0.1\text{ s}$ that were in thermal equilibrium, the factor $\zeta_{\rm FIP}$ reaches $\mathcal{O}(0.1)$, see Appendix 3.3.1 using HNLs as an example.

The simple analytic estimate leads to the model-independent bound on FIPs that decay hadronically:

$$\tau_{\rm FIP} \lesssim \frac{0.023 \left(\frac{1.5\text{ MeV}}{T_0^{\rm min}}\right)^2 \text{ s}}{1 + 0.07 \ln\left[\frac{P_{\rm conv}}{0.1} \frac{{\rm Br}_{{\rm FIP}\to h}}{0.4} \frac{2n_{\rm FIP,dec}}{3n_\gamma(T_{\rm dec})} \cdot 24 \left(\frac{a_{\rm dec} T_{\rm dec}}{a_0 T_0^{\rm min}}\right)^3\right]}. \tag{3.1.13}$$

The presence of mesons increases the ^{4}He abundance. Therefore, in order to fix $T_0^{\rm min}(m_{\rm FIP})$, we need to adopt an upper bound on the primordial ^{4}He abundance, $Y_{p,\rm max}$, that is consistent with measurements [52]. The smallest error bars come from measuring Y_p in

[3] This is indeed the case for short-lived FIPs with $\tau_{\rm FIP} \ll 0.1$ s.

low-metallicity interstellar regions and extrapolating its value to zero metallicity (pioneered in [180]). Several groups [171–175] have determined Y_p using this method, albeit with different data and assumptions. The resulting scatter between results is larger than the reported error bars. We treat this difference as an additional systematic uncertainty and adopt the maximal value $Y_{p,\text{max}} = 0.2573$ (see Appendix 3.1.1). The maximally allowed relative deviation is therefore

$$\Delta Y_p / Y_{p,\text{SBBN}} \approx 4.35\%. \tag{3.1.14}$$

To relate ΔY_p and T_0^{min}, we study how the n_n/n_p ratio is relaxed below T_0. The relaxation occurs solely via the SBBN reaction,

$$\frac{dX_n}{dt} = \Gamma^{\text{SBBN}}_{p\to n}(1 - X_n) - \Gamma^{\text{SBBN}}_{n\to p} X_n, \quad X_n = \frac{n_n}{n_n + n_p}, \tag{3.1.15}$$

albeit with the altered initial condition $X_n(T_0) = X_n^h \simeq 1/2$. ($\Gamma^{\text{SBBN}}_{p\leftrightarrow n}(t)$ are SBBN rates, see [45]). Non-SBBN value of $X_n(T_0)$ *is the dominant effect* of short-lived HNLs on Y_p. At temperatures $T \lesssim T_0$, for HNLs with lifetimes $\tau_N \lesssim 0.02$ s, all other quantities that are relevant for BBN dynamics – η_B, time-temperature relation, the nuclear reactions chain – remain the same as in SBBN, which is because most of HNLs are no longer left in the plasma at these temperatures (see also Appendix 3.B.1). As a result, a value of $X_n(T_0)$ is translated into ΔY_p via

$$\frac{\Delta Y_p}{Y_{p,\text{SBBN}}} = \frac{\Delta X_n(T_{\text{BBN}})}{X_{n,\text{SBBN}}(T_{\text{BBN}})}, \tag{3.1.16}$$

where $T_{\text{BBN}} \approx 84$ keV is the temperature of the onset of nuclear reactions in SBBN [45].

To obtain the bound (3.1.13), we considered exclusively meson-driven $p \leftrightarrow n$ processes for $T > T_0^{\text{min}}$ and only weak SBBN processes for $T < T_0^{\text{min}}$. We also solved numerically the equation (3.1.15) for the neutron abundance in the presence of both mesons-driven and SBBN $p \leftrightarrow n$ conversion rates in Appendix 3.B.1 using HNLs as an example model, and obtained results being in perfect agreement with the bound (3.3.24). We have also repeated our analysis for the case of the GeV-mass scalar that mixes with the Higgs and found an excellent agreement with [181, 182].

> We conclude that BBN may constrain hadronically decaying FIPs with lifetimes as small as $\tau_{\text{FIP}} \simeq 0.02$ s.

3.1.3 Limits of applicability of the bound

Eq. (3.1.13) defines the lower bound on FIP lifetimes that may be constrained from the meson-driven ^{4}He overproduction. Our simplified approach is limited by lifetimes for which FIPs or their decay products survive until the onset of nuclear reactions. In this case, the dynamics of nuclear reactions gets changed by

1. meson-driven $p \leftrightarrow n$ conversion and nuclear dissociation processes;
2. change of time-temperature relation by FIPs;
3. change of η_B during nuclear reaction;
4. photo-dissociation processes by high-energetic photons originating from EM decays of FIPs.

Among these effects, the effect which firstly manifests with the increase of the lifetime is the meson-driven nuclear dissociation. Indeed, a change of η_B and $t(T)$ requires a FIP to contribute to the energy density significantly, while the effect of mesons only requires the amount of meson-driven reactions to be comparable with n_B. As for the photo-dissociation, it becomes relevant only from lifetimes of order $\tau_{\rm FIP} \gtrsim 10^4$ s, which is the time scale at which photons with energies large enough to dissociate deuterium no longer instantly disappear because of the annihilation $\gamma + \gamma_{\text{SM plasma}} \to e^+ + e^-$ (see, e.g., [182]). Let us now estimate the upper bound on the FIP lifetimes at which the simple analysis presented above is valid. The ^{4}He threshold-less dissociation processes with mesons are (see [178])

$$\pi^- + {}^4\text{He} \to T + n, \quad \pi^- + {}^4\text{He} \to D + 2n, \quad \pi^- + {}^4\text{He} \to p + 3n \tag{3.1.17}$$

To estimate the lifetimes at which the processes (3.1.17) can be neglected, we compare the number density of mesons *available for the dissociation* with the number density of ^{4}He nuclei:

$$n^h_{\text{He diss}}(T_{\text{BBN}}) \ll n_{\text{He}}(T_{\text{BBN}}), \tag{3.1.18}$$

c.f. Eq. (3.1.8). Here, $n^\pi_{\text{He diss}}$ is defined via

$$n^\pi_{\text{He diss}}(T_{\text{BBN}}) = n_{\text{FIP}} \cdot \text{Br}_{\text{FIP}\to\pi^-} \cdot P_{\text{He diss}}, \tag{3.1.19}$$

Here,

$$n_{\text{FIP}}(T) = \left(\frac{a_{\text{SBBN}}(t_{\text{dec}})}{a_{\text{SBBN}}(T)}\right)^3 n^{\text{dec}}_{\text{FIP}} \cdot \zeta_{\text{FIP}} \cdot e^{-t/\tau_{\text{FIP}}} \tag{3.1.20}$$

is the FIP's number density, and $P_{^4\text{He diss}}$ is the probability for a single meson to dissociate ^{4}He nuclei before decaying:

$$P_{\text{He diss}} = \frac{\langle\sigma^\pi_{\text{He diss}} v\rangle n_{\text{He}}}{\Gamma^\pi_{\text{decay}}} \simeq 8.3 \cdot 10^{-2} \cdot \frac{4 \cdot n_{\text{He}}}{n_B} \left(\frac{T}{1\text{ MeV}}\right)^3, \tag{3.1.21}$$

where we used the total cross-section of the dissociation processes (3.1.17), $\langle\sigma^\pi_{\text{He diss}} v\rangle \simeq 6.5 \cdot F_{\text{He}\pi^-}$ mb (a factor $F_{\text{He}\pi^-} \simeq 3.5$ accounts for the Coulomb attraction, Eq. (3.B.3)).

In our estimates, we use $T_{\text{BBN}} = 84$ keV, assuming that all free nucleons become bounded in ^{4}He nuclei at this temperature. We also do not take into account that after the dissociation of Helium the abundance of lighter elements will be also increased significantly.

Further, to make a conservative upper bound estimate, we will assume that the FIPs were in thermal equilibrium and then decoupled while being UR, such that their abundance is maximally possible. In this case, $\zeta_{\text{FIP}} \simeq 10^{-2} - 10^{-1}$ for FIPs with lifetimes $\tau_{\text{FIP}} \sim 10^2$ s, in dependence on the FIP mass.[4] Requiring $n_{\text{He}} \simeq n_B/4$ in (3.1.18), and using Eqs. (3.1.19), (3.1.21), we arrive at the upper bound on HNL lifetimes for which our analysis is applicable, $\tau_{\text{FIP}} \lesssim 50$ s.

For FIPs that decay hadronically and have lifetimes $\tau_N \lesssim 10^4$ s, the presence of mesons from their decays may lead to an increase of primordial ^{4}He abundance.

Indeed, nuclear reactions are efficient until temperatures $T \simeq$ few keV. Therefore, once mesons disappear from the plasma (and nuclear dissociation processes stop), neutrons and protons get bounded into ^{4}He. Since the meson-driven $p \leftrightarrow n$ conversion keeps n/p ratio at the level of $\mathcal{O}(1)$, the resulting abundance may be still larger than the SBBN value.

In order to derive corrections to the nuclear abundances, we need to estimate the impact of effects of long-lived FIPs on BBN, which is complicated to perform analytically. Namely, the BBN reaction chain is non-equilibrium. In addition, the impact of FIPs on η_B and $t(T)$ cannot be estimated as a perturbation, since FIP may be abundant non-relativistic particles which dominate the energy density of the Universe. Therefore, numeric solution of equations for nuclear abundances and Friedmann equations in presence of decaying FIPs is required. We do this for long-lived HNLs in Sec. 3.3.3.1.

3.2 CMB

As we have discussed in the Introduction (remind Sec. 1.3.2), the main impact of FIPs with lifetimes $\tau_{\text{FIP}} \ll t_{\text{recombination}}$ on CMB comes from their change of Y_p and N_{eff}.

The value of N_{eff} is given by

$$N_{\text{eff}} \equiv \frac{8}{7}\left(\frac{11}{4}\right)^{4/3}\left(\frac{\rho_{\text{rad}} - \rho_\gamma}{\rho_\gamma}\right) , \tag{3.2.1}$$

where ρ_{rad} and ρ_γ are the total radiation and photon energy densities respectively. We define the change in this quantity as $\Delta N_{\text{eff}} = N_{\text{eff}} - N_{\text{eff}}^{\text{SM}}$, where within the Standard Model $N_{\text{eff}}^{\text{SM}} \simeq 3.044$ [183–187]. Any deviation from the SM value is regulated by weak interactions between neutrinos and electromagnetic (EM) particles, which are efficient enough at temperatures $T \gg 1\,\text{MeV}$ to keep these species in equilibrium with each other. At lower temperatures, the interactions gradually go out of equilibrium and the energy exchange between the two sectors will stop. Decaying FIPs can affect this delicate process

[4] As the bound is sensitive logarithmically to this product, its precise value is not so important.

in different ways, depending on whether they inject most of their energy into EM particles or neutrinos.

The impact of FIPs predominantly decaying into EM particles has been extensively studied in the literature, see e.g. [64, 182, 188–191]. Such particles heat up the EM plasma and consequently decrease $N_{\rm eff}$, independently of whether the decay happens during or after neutrino decoupling.

For FIPs that mostly decay into neutrinos, we naively expect that $N_{\rm eff}$ would increase. This is indeed true for lifetimes $\tau_{\rm FIP} \gg t_\nu^{\rm dec} \sim 0.1 - 1\,{\rm s}$, where $t_\nu^{\rm dec}$ is the time of neutrino decoupling, see e.g. [192]. However, there are controversial results for the lifetimes $\tau_{\rm FIP} \sim t_\nu^{\rm dec}$.

Neutrinos are still in partial equilibrium and try to equilibrate with the injected neutrinos at such time scale. This scenario has been considered before in [193–195] that arrived at different conclusions about the impact on $N_{\rm eff}$. Namely, the work [193] studied a reheating scenario in which all the SM particles are absent before FIPs start decaying. In such a framework, all neutrinos have high energies, which means that they mainly thermalize via neutrino-EM interactions and $N_{\rm eff}$ naturally decreases. References [194, 195] considered HNLs with masses $m_N < m_\pi$ and lifetimes $\tau_N \lesssim 1\,{\rm s}$. Such HNLs are in thermal equilibrium in the early Universe, but decouple as the Universe expands and eventually decay mainly into high-energy neutrinos at MeV temperatures. These two works drew different conclusions about $N_{\rm eff}$: [194] reported $\Delta N_{\rm eff} > 0$ for the whole studied mass range, whereas [195] presented in their Fig. 3 that $\Delta N_{\rm eff} < 0$ for masses $60\,{\rm MeV} \lesssim m_N < m_\pi$ and lifetimes $\tau_N \ll 1\,{\rm s}$. The sign of $\Delta N_{\rm eff}$ is not emphasized in these two papers; [195] did not comment on the contradiction with [194] on this issue and no physical discussion of this phenomenon was provided[5].

In this section, we aim to clarify the behavior of $N_{\rm eff}$ in the presence of FIPs that decay mainly into neutrinos and have lifetimes $\tau_{\rm FIP} \sim t_\nu^{\rm dec}$. Here we will assume that a thermal bath of SM particles is already present in the primordial Universe. We will first construct a simple model in Sec. 3.2.1 that provides us with a qualitative understanding of how such particles impact $N_{\rm eff}$, the findings of which we then confirm by using the Boltzmann code `pyBBN` [64].

[5] A more recent work [196] considered *long-lived* (i.e., decaying after e^+e^- annihilation) HNLs that could decay both into EM particles and neutrinos. In this case, $N_{\rm eff}$ could both increase and decrease, as at such late times the injected energy densities from HNL decays can dominate over the SM densities of both the EM and neutrino sectors. Another recent paper [197], which appeared after our work was submitted, claims that $\Delta N_{\rm eff} \geq 0$ for all cases in which FIPs decay mostly into neutrinos. We comment on it in Appendix 3.D.

Our analysis shows that short-lived FIPs that inject most of their energy into neutrinos may decrease $N_{\rm eff}$. This is because during the equilibration process, the injected high-energy neutrinos redistribute their energy among the neutrino and EM plasma.

If the energy of the injected neutrinos is sufficiently large, the energy transfer to the EM sector occurs faster than the equilibration with the neutrino sector. This means that the EM plasma heats up more than the neutrino plasma, which eventually leads to $\Delta N_{\rm eff} < 0$. We will find that this mechanism is especially relevant for FIPs with masses larger than a few tens of MeV. We will then apply these general considerations to the well-motivated case of HNLs. Complementary details and simulation results are included in the appendix 3.2.1.2.

3.2.1 Impact of short-lived FIPs on $N_{\rm eff}$

We focus on FIPs with masses $\gg 1\,\mathrm{MeV}$ that decay when neutrinos are still in (partial) equilibrium. Such FIPs can decay into high-energy neutrinos with energies much higher than those in the primordial plasma, that then participate in interactions with thermal neutrinos and electrons/positrons.

We will find that even if most of the FIP energy is injected into neutrinos, these interactions may still cause a decrease in $N_{\rm eff}$. This feature appears since the injected high-energy neutrinos get quickly converted into electrons/positrons and drag thermal neutrinos residing in the plasma along with them. During this process, neutrino-neutrino interactions lead to the presence of residual non-thermal distortions in the distribution functions of neutrinos (neutrino spectral distortions) that keep the balance of $\nu \leftrightarrow \mathrm{EM}$ interactions shifted to the right till long after the injection (i.e., more energy is transferred from the neutrino plasma to the electromagnetic plasma than vice versa). The energy transfer from neutrinos to EM particles accumulated over time can then be sizeable enough, such that $\Delta N_{\rm eff}$ becomes negative. This effect diminishes with larger FIP lifetime, as neutrino-EM interactions go out of equilibrium and neutrinos can no longer be converted into electrons/positrons. Therefore, FIPs that decay into neutrinos after neutrino decoupling will lead to $\Delta N_{\rm eff} > 0$. In what follows, we will consider FIPs that can decay into both neutrinos and EM particles, and construct a simple model that provides a semi-analytic description of the aforementioned effect. At the end of this section, we will also highlight and further elaborate on the central role of neutrino spectral distortions in the dynamics of $N_{\rm eff}$.

3.2.1.1 Analytic considerations

We assume that the amount of injected non-equilibrium neutrinos is only a small fraction of the thermal neutrinos in the plasma. The evolution of the injected neutrinos is then mainly

governed by the following reactions:

$$\nu_{\text{non-eq}} + \nu_{\text{therm}} \rightarrow \nu_{\text{non-eq}} + \nu_{\text{non-eq}} \tag{3.2.2}$$

$$\nu_{\text{non-eq}} + \overline{\nu}_{\text{therm}} \rightarrow e^{+} + e^{-} \tag{3.2.3}$$

$$\nu_{\text{non-eq}} + e^{\pm} \rightarrow \nu_{\text{non-eq}} + e^{\pm} \, , \tag{3.2.4}$$

where 'non-eq' and 'therm' refer to neutrinos with non-equilibrium and thermal energies respectively.

Through the thermalization reactions (3.2.2)-(3.2.4), non-equilibrium neutrinos thermalize and quickly redistribute their energy among the neutrino and EM plasma.

The energy loss rate of these non-equilibrium neutrinos is higher than the interaction rates of thermal particles [49]:

$$\frac{\Gamma_{\text{non-eq}}}{\Gamma_{\text{therm}}} \sim \frac{G_F^2 T^4 E_\nu^{\text{inj}}}{G_F^2 T^5} = \frac{E_\nu^{\text{inj}}}{T} \gg 1 \, , \tag{3.2.5}$$

where E_ν^{inj} is the average energy of the injected non-equilibrium neutrinos. Note that reactions between thermal particles also exchange energy between the neutrino and EM sectors, but this energy exchange is subdominant as far as Eq. (3.2.5) holds.

The amount of energy that ends up in the EM plasma has three contributions: *1)* the direct decay of FIPs into EM particles, *2)* the energy transfer of non-equilibrium neutrinos to EM particles during thermalization and *3)* the energy transfer from thermal neutrinos to EM particles as a consequence of them being dragged by non-equilibrium neutrinos during thermalization (reactions (3.2.2) and (3.2.3)). The first process injects a fraction ξ_{EM} of the total FIP energy into the EM plasma, while the latter two increase this fraction to:

$$\xi_{\text{EM,eff}}(E_\nu^{\text{inj}}, T) = \xi_{\text{EM}} + \xi_\nu \times \epsilon(E_\nu^{\text{inj}}, T) \, , \tag{3.2.6}$$

where $\xi_\nu = 1 - \xi_{\text{EM}}$ is the energy fraction that FIPs directly inject into the neutrino sector and $\epsilon = \epsilon_{\text{non-eq}} + \epsilon_{\text{thermal}}$ is the effective fraction of ξ_ν that went to the EM plasma during the thermalization. The latter quantity can be split in a contribution from non-equilibrium neutrinos ($\epsilon_{\text{non-eq}} = E_\nu^{\text{non-eq}\rightarrow\text{EM}}/E_\nu^{\text{inj}}$) and an *effective* contribution from thermal neutrinos ($\epsilon_{\text{thermal}} = E_\nu^{\text{thermal}\rightarrow\text{EM}}/E_\nu^{\text{inj}}$).

Now, based on Eq. (3.2.6), if $\epsilon > 0.5$, then $\xi_{\text{EM,eff}} > 0.5$. This means that more than half of the FIP energy eventually ends up in the EM plasma (i.e., EM plasma heats up more than the neutrino plasma), which results in $\Delta N_{\text{eff}} < 0$ independently of the value of ξ_{EM}[6]. This simplified energy redistribution picture only holds if the non-equilibrium

[6]Note that it is not a requirement that ϵ must be larger than 0.5 in order for ΔN_{eff} to be negative. It only signifies the independence from ξ_{EM}.

neutrino energy is much larger than the average energy of thermal neutrinos. Once these two energies become similar in magnitude, backreactions cannot be neglected anymore and the evolution can only be accurately described with a system of Boltzmann equations.

Because of much faster thermalization rate of EM plasma than of neutrinos and growth of the interaction rate with neutrino energy, neutrinos may store a huge amount of their energy ϵ in EM plasma during the thermalization. We may estimate it analytically.

We can make a simple estimate of ϵ as a function of the injected neutrino energy $E_\nu^{\rm inj}$ and temperature T. We start with describing the thermalization process of a *single* injected neutrino, which causes a cascade of non-equilibrium neutrinos. Such a cascade can result after the injected neutrino participates in the processes (3.2.2)−(3.2.4). We assume that in the processes (3.2.2) and (3.2.4) each non-equilibrium neutrino in the final state carries half of the energy of the non-equilibrium neutrino in the initial state. Thus, roughly speaking, the thermalization occurs during $N_{\rm therm} \simeq \log_2(E_\nu^{\rm inj}/3.15T)$ interactions. In addition, the process (3.2.2) doubles the number of non-equilibrium neutrinos, while (3.2.3) makes neutrinos disappear and (3.2.4) leaves the number unchanged. Therefore, after the k-th step in the cascade, the average number of non-equilibrium neutrinos is given by:

$$N_\nu^{(k)} = N_\nu^{(k-1)} \left(2P_{\nu\nu\to\nu\nu} + P_{\nu e\to\nu e}\right) = N_\nu^{(0)} \left(2P_{\nu\nu\to\nu\nu} + P_{\nu e\to\nu e}\right)^k , \tag{3.2.7}$$

with $N_\nu^{(0)} = 1$, and the total non-equilibrium energy is:

$$E_\nu^{(k)} = E_\nu^{(k-1)} \left(P_{\nu\nu\to\nu\nu} + \frac{1}{2}P_{\nu e\to\nu e}\right) = E_\nu^{\rm inj} \left(P_{\nu\nu\to\nu\nu} + \frac{1}{2}P_{\nu e\to\nu e}\right)^k , \tag{3.2.8}$$

where $P_{\nu\nu\to\nu\nu}$, $P_{\nu\nu\to ee}$, and $P_{\nu e\to\nu e}$ are the average probabilities of the processes (3.2.2)-(3.2.4), respectively, and their sum equals unity. We define these probabilities as $P_i = \Gamma_i/\Gamma_\nu^{\rm tot}$, where Γ_i is the interaction rate of each process and $\Gamma_\nu^{\rm tot}$ is the total neutrino interaction rate. The relevant reactions and their corresponding matrix elements are summarized in appendix D of [64]. Assuming a Fermi-Dirac distribution for neutrinos and averaging over neutrino flavours, we find:

$$P_{\nu\nu\to\nu\nu} \approx 0.76, \quad P_{\nu\nu\to ee} \approx 0.05, \quad P_{\nu e\to\nu e} \approx 0.19\,. \tag{3.2.9}$$

Finally, the value of $\epsilon_{\text{non-eq}}$ that accounts for the energy transfer from non-equilibrium neutrinos to the EM plasma is given by:

$$\epsilon_{\text{non-eq}} = \frac{1}{E_\nu^{\rm inj}} \sum_{k=0}^{N_{\rm therm}} \left(\frac{P_{\nu e\to\nu e}}{2} + P_{\nu\nu\to ee}\right) E_\nu^{(k)} \,. \tag{3.2.10}$$

In addition to the transferred non-equilibrium energy, the non-equilibrium neutrinos catalyze the energy transfer from thermal neutrinos to the EM plasma via the processes (3.2.2) and (3.2.3). In other words, during the thermalization process non-equilibrium neutrinos drag thermal neutrinos along with them, which leads to part of the energy stored in the thermal neutrino sector to end up in the EM sector. We assume that each reaction (3.2.2) transfers an energy amount of $3.15T$ from the thermal neutrino sector to non-equilibrium neutrinos, which then via (3.2.3) ends up in the EM plasma. Moreover, each reaction (3.2.3) contributes to another energy transfer of $3.15T$ from thermal neutrinos to the EM plasma. The effective contribution coming from this transfer is therefore:

$$\epsilon_{\text{thermal}} = \frac{3.15T}{E_\nu^{\text{inj}}} N_\nu^{\text{therm}\to\text{EM}} = \frac{3.15T}{E_\nu^{\text{inj}}} P_{\nu\nu\to ee} \left(\sum_{k=0}^{N_{\text{therm}}} N_\nu^{(k)} + \left[P_{\nu\nu\to\nu\nu} + \sum_{k=1}^{N_{\text{therm}}} \left(2P_{\nu\nu\to\nu\nu}\right)^{(k)} \right] \right), \tag{3.2.11}$$

where the first term in the round brackets is the contribution from the process (3.2.3) and the terms in the square brackets are the contribution from the process (3.2.2). Note that the factor of 2 in the second sum accounts for the doubling of non-equilibrium neutrinos in the process (3.2.2). We find that $\epsilon_{\text{thermal}}$ is at least 5 times smaller than $\epsilon_{\text{non-eq}}$, which makes this a sub-dominant effect.

As the Universe expands and the temperature decreases, weak reaction rates start to compete with the Hubble rate H. The energy transfer from neutrinos to the EM plasma therefore becomes less and less efficient, and ϵ tends to zero. In order to incorporate this effect, we multiply the probabilities in (3.2.9) with a factor $\min[\Gamma_i/H, 1]$, where $\Gamma_i = \Gamma_i(E_\nu^{\text{inj}}/2^k)$ is the interaction rate of any of the processes (3.2.2)−(3.2.4). The resulting energy fraction of neutrinos that is transferred to the EM plasma $\epsilon = \epsilon_{\text{non-eq}} + \epsilon_{\text{thermal}}$ is shown in Fig. 3 for a number of injected neutrino energies E_ν^{inj}.

The analytic model tells us that ϵ can exceed 0.5 for $E_\nu^{\text{inj}} \gtrsim 60\,\text{MeV}$. This means that when FIPs decay into neutrinos with such energies at temperatures of a few MeV, the majority of the injected neutrino energy will end up in the EM plasma during the thermalization. This then leads to a decrease of N_{eff}, independently of how much energy the FIPs inject into the EM sector.

Now that we are able to estimate ϵ, we can compute the correction to N_{eff} for some benchmark FIP scenario. It is worth noting here again that ϵ only depends on the energy of the injected neutrino and the temperature at which the injection happens. This means that ϵ is an independent quantity of the FIP model considered, in contrast to ξ_{EM} and ξ_ν, which do depend on the choice of the model. As an illustrative example, we assume that $\xi_{\text{EM}} = 0$, i.e., the FIP injects all of its energy into neutrinos ($\xi_\nu = 1$). Given that in our simple model neutrinos thermalize very quickly, we assume that they have a thermal-like distribution with a temperature T_ν and follow the approach in [184, 198] to obtain the time evolution of T_ν

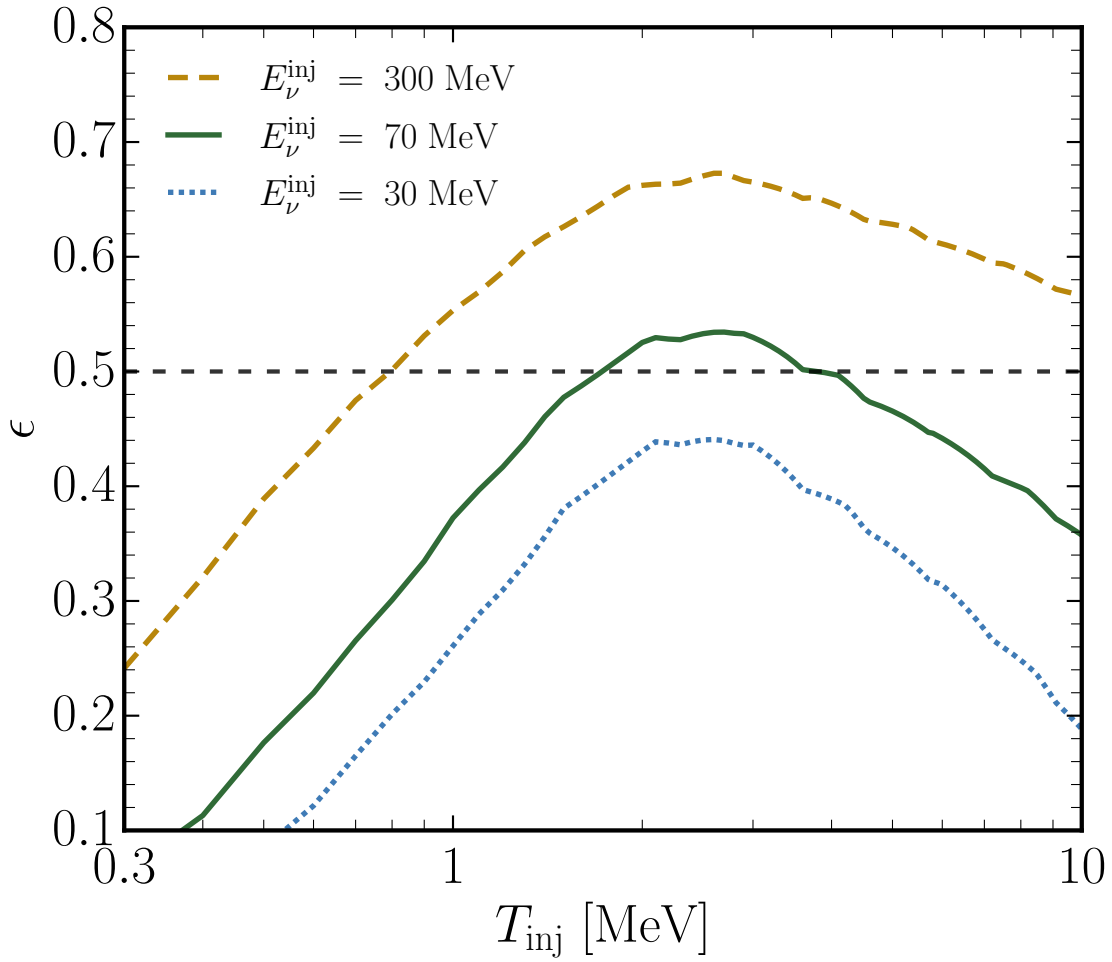

Figure 3: Estimate of the fraction of injected neutrino energy ϵ (both thermal and non-equilibrium) that gets transferred to the EM plasma during thermalization (see text for details). The three curves indicate the value of ϵ when a neutrino of energy $E_\nu^{\rm inj}$ is injected at a temperature $T_{\rm inj}$. At high temperatures of order of $T_{\rm inj} \simeq E_\nu^{\rm inj}$, the injected neutrinos are thermal-like, and hence ϵ is small. Once the temperature decreases, we enter the regime $E_\nu^{\rm inj} \gg 3.15 T_{\rm inj}$ and neutrinos transfer a significant amount of their energy to the EM plasma while thermalising. With further decrease of $T_{\rm inj}$, weak reactions go out of equilibrium and the energy transfer becomes less and less efficient, which results in a quick drop-off of ϵ.

and $T_{\rm EM}$ in the presence of decaying FIPs (see Appendix 3.C, where we provide the relevant equations). In this benchmark example, we consider a generic FIP of mass 500 MeV that can decay only into three neutrinos and show $\Delta N_{\rm eff}$ as a function of its lifetime in Fig. 5. In order to compare the accuracy of our simple model, we also include in this figure the evolution of $\Delta N_{\rm eff}$ as obtained from the publicly available Boltzmann code `pyBBN`[7] [64]. The grey band in this figure indicates the current sensitivity of $N_{\rm eff}$ by Planck, which at 2σ

[7] https://github.com/ckald/pyBBN

reads[8] $N_{\rm eff}^{\rm CMB} = 2.89 \pm 0.62$ [199, 200]. We see that $N_{\rm eff}$ can significantly decrease as a result of the thermalization of the injected neutrinos. This decrease of $N_{\rm eff}$ would only be further amplified if the FIPs were also to inject some of their energy into the EM plasma.

3.2.1.2 Effect of Residual Non-equilibrium Neutrino Distortions

> The simple model described in Sec. 3.2.1 relies on the assumption that the remaining fraction $1-\epsilon$ of the injected neutrino energy is perfectly thermal. In reality, this may not be the case and the full thermalization would occur during a much larger number of interactions than $N_{\rm therm} \simeq \log_2(E_\nu^{\rm inj}/3.15T)$.

Therefore, this simple model underestimates the energy fraction that goes into the EM plasma[9]. The remaining non-equilibrium neutrinos will manifest themselves as residual non-thermal spectral distortions in the distribution function of neutrinos. These spectral distortions keep the energy exchange balance of $\nu \leftrightarrow \text{EM}$ reactions shifted to the right till long after FIP decay. As a result, more neutrino energy will be transferred to the EM plasma and $N_{\rm eff}$ can further decrease. There is a subtlety here that the remaining $1-\epsilon$ non-equilibrium neutrinos are only slightly hotter than the thermal neutrinos, and we cannot describe their thermalization as an instant process: The corresponding rate is comparable to the thermal energy exchange rate. As such, the energy transfer process is extended in time, and a proper study of this effect requires solving the Boltzmann equation for the neutrino distribution function.

To study the impact of neutrino spectral distortions on the $\nu \to \text{EM}$ energy balance shift, we consider a simple scenario where high-energy neutrinos are instantly injected into the primordial plasma. We make use of the publicly available Boltzmann code `pyBBN`[10] [64] to simulate this process and to track the evolution of the neutrino distribution functions. Within this setup, neutrinos with energy $E_\nu^{\rm inj} = 70\,\text{MeV}$ are instantly injected at $T = 3\,\text{MeV}$. They amount for a fixed percentage of the total neutrino energy density and are equally distributed over the three neutrino flavours. All Standard Model interactions as specified in [64] are included, but with neutrino oscillations turned off (without any loss of generality). In order to highlight the importance of neutrino spectral distortions, we perform this procedure a second time, but with neutrino spectral distortions turned off. In that case, the neutrino distribution function is given by a Fermi-Dirac distribution with temperature $T_{\nu_\alpha} = \left(\frac{240\rho_{\nu_\alpha}}{7\pi^2 g_{\nu_\alpha}}\right)^{1/4}$, where ρ_{ν_α} and $g_{\nu_\alpha} = 2$ are the energy density (of both neutrinos and anti-neutrinos) and number of degrees of freedom of neutrino flavour α respectively.

[8] This value is obtained from the Planck 2018 baseline TTTEEE+lowE analysis, where $N_{\rm eff}$, $Y_{\rm P}$ and the six base parameters in ΛCDM are varied.

[9] Once the energy of the non-equilibrium neutrinos is close to the average thermal energy of $3.15T$, they lose roughly $\Delta E_\nu = (E_\nu - 3.15T)/2$ of energy per scattering. Therefore, the number of scatterings required to diminish E_ν down to $3.15T$ is larger.

[10] https://github.com/ckald/pyBBN

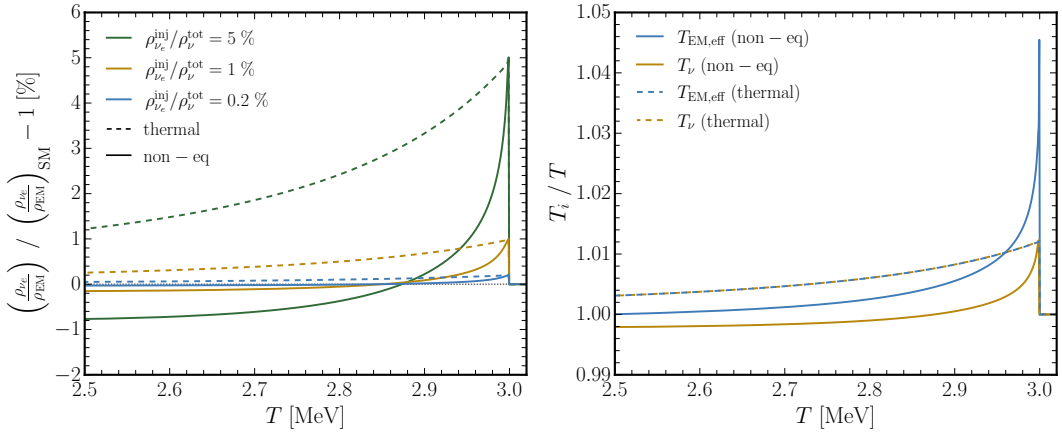

Figure 4: Evolution of the neutrino and EM plasma after the instant injection of neutrinos with energy $E_\nu^{\rm inj} = 70\,{\rm MeV}$ at $T = 3\,{\rm MeV}$. *Left panel*: The ratio of electron neutrino energy density to electromagnetic energy density, relative to the SM prediction. Three fractions of the injected energy density are considered: $\rho^{\rm inj}_{\nu_e}/\rho^{\rm tot}_{\nu} = \{0.2\%, 1\%, 5\%\}$. The solid lines are obtained by taking into account the full non-equilibrium spectrum of neutrinos, whereas the dashed lines correspond to the evolution assuming that neutrinos always have a thermal-like spectrum with temperature $T_\nu \propto \rho_\nu^{1/4}$. *Right panel*: Evolution of the neutrino temperature (dashed) and effective EM plasma temperature (solid) for which the energy transfer rate in Eq. (3.2.13) vanishes. An injected fraction of $\rho^{\rm inj}_{\nu_e}/\rho^{\rm tot}_{\nu} = 5\%$ is considered here. The solid and dashed lines indicate when non-equilibrium and thermal-like neutrino distributions are used respectively.

The evolution of the ratio $\rho_{\nu_e}/\rho_{\rm EM}$ (relative to the one in the SM) is shown in the left panel of Fig. 4 for different amounts of injected neutrino energy. In agreement with the story in Sec. 3.2.1, we observe a fast drop-off in the ratio right after the injection, which signifies the quick transfer of energy from the neutrino plasma to the EM plasma. After reaching the SM value (which naively corresponds to an equilibrium state), the ratio continues decreasing. This is the effect of the extended thermalization due to neutrino spectral distortions, as caused by the remaining fraction $1 - \epsilon$ of non-equilibrium neutrinos. Eventually, the ratio will be smaller than the SM value and $\Delta N_{\rm eff}$ becomes negative. In this plot, the dashed lines correspond to the same simulations but with a thermal-like distribution for the neutrinos. It is clear that without spectral distortions, the energy transfer from the neutrino sector to the EM sector is much less efficient.

Another way to look at this shift in the energy transfer balance from the neutrino plasma to the EM plasma is to ask the question: Which temperature $T_{\rm EM,eff}$ is the EM plasma *trying* to reach after the injection? As we will see, depending on whether neutrinos have a non-equilibrium or a thermal-like distribution, this temperature can be either larger than or equal to the neutrino temperature[11]. In the former case, it means that the EM plasma

[11] In all cases, with 'neutrino temperature' we refer to the quantity $T_\nu = \left(\frac{240\rho_\nu}{7\pi^2 g_\nu}\right)^{1/4}$, where $g_\nu = 2$ and ρ_ν is the energy density of both neutrinos and anti-neutrinos.

temperature can exceed the neutrino temperature (and thus $\Delta N_{\rm eff}$ can be negative), while in the latter case $\Delta N_{\rm eff}$ *cannot* be negative.

In more technical terms, the exchange of energy between neutrinos and EM particles is regulated by the Boltzmann collision integral $I_{\rm coll}$, which encodes all interactions between the species. For neutrinos that participate in reactions of the form $\nu + 2 \leftrightarrow 3 + 4$, the collision integral is given by [201]:

$$\begin{aligned} I_\nu =& \frac{1}{2 g_\nu E_\nu} \sum_{\text{reactions}} \int \prod_{i=2}^{4} \left(\frac{\mathrm{d}^3 p_i}{(2\pi)^3 2E_i} \right) |\mathcal{M}|^2 \times \\ & \times \left[(1-f_\nu)(1-f_2) f_3 f_4 - f_\nu f_2 (1-f_3)(1-f_4) \right] \times \\ & \times (2\pi)^4 \delta^4 (P_\nu + P_2 - P_3 - P_4) \, , \end{aligned} \tag{3.2.12}$$

where f_i and P_i are the distribution function and four momentum of species i respectively, and $|\mathcal{M}|^2$ is the unaveraged squared matrix element summed over degrees of freedom of initial and final states. The energy transfer rate between the neutrino and EM plasma can be written as:

$$\Gamma(T_{\rm EM}) = \int \frac{\mathrm{d}^3 p_\nu}{(2\pi)^3} I_{\rm coll}(T_{\rm EM}) E_\nu \, , \tag{3.2.13}$$

where we consider $I_{\rm coll}$ to be a function of the EM plasma temperature $T_{\rm EM}$. There exists a temperature $T_{\rm EM,eff}$ for which this rate is equal to 0. This corresponds to the temperature the EM plasma *tends* to during thermalization, since then the system would be in equilibrium. In the case where neutrinos would have a thermal-like spectrum with temperature T_ν, the rate vanishes when $T_{\rm EM,eff} = T_\nu$. On the other hand, when a non-equilibrium neutrino spectrum is considered, we find that $T_{\rm EM,eff} > T_\nu$ when $\Gamma = 0$. In the former case $N_{\rm eff}$ *cannot* decrease, while in the latter case the EM plasma temperature can exceed T_ν and $N_{\rm eff}$ can thus decrease. We show the evolution of $T_{\rm EM,eff}$ and T_ν as obtained from the instant neutrino injection simulations in the right panel of Fig. 4.

The conclusion here is that neutrino spectral distortions play a central role in transferring energy from the neutrino sector to the EM sector. When considering short-lived FIPs that can decay into neutrinos, the impact of these distortions on the evolution of $N_{\rm eff}$ should not be neglected.

Comparing the analytic model with the numeric simulations, we find that when using the Boltzmann equation $N_{\rm eff}$ decreases more than predicted by our semi-analytic model.

> The remaining fraction $1 - \epsilon$ of the injected neutrinos is not perfectly thermal, manifesting themselves as residual spectral distortions in the distribution function of neutrinos that further lead to a transfer of energy from the neutrino sector to the EM sector.

We see that in some cases the inclusion of this effect can make the difference between

being excluded by current data or not. We elaborate more on the effect of spectral distortions in Appendix 3.2.1.2. In short, the semi-analytic model is useful in providing a qualitative understanding of the behavior of $N_{\rm eff}$ in the presence of decaying FIPs. On the other hand, if the aim is to obtain accurate predictions for $N_{\rm eff}$ (relevant for setting bounds and forecasting), it is crucial to use the Boltzmann equation to track the evolution of the neutrino distribution functions. As such, we will use `pyBBN` in the remainder of this paper to simulate the impact of FIPs on $N_{\rm eff}$.

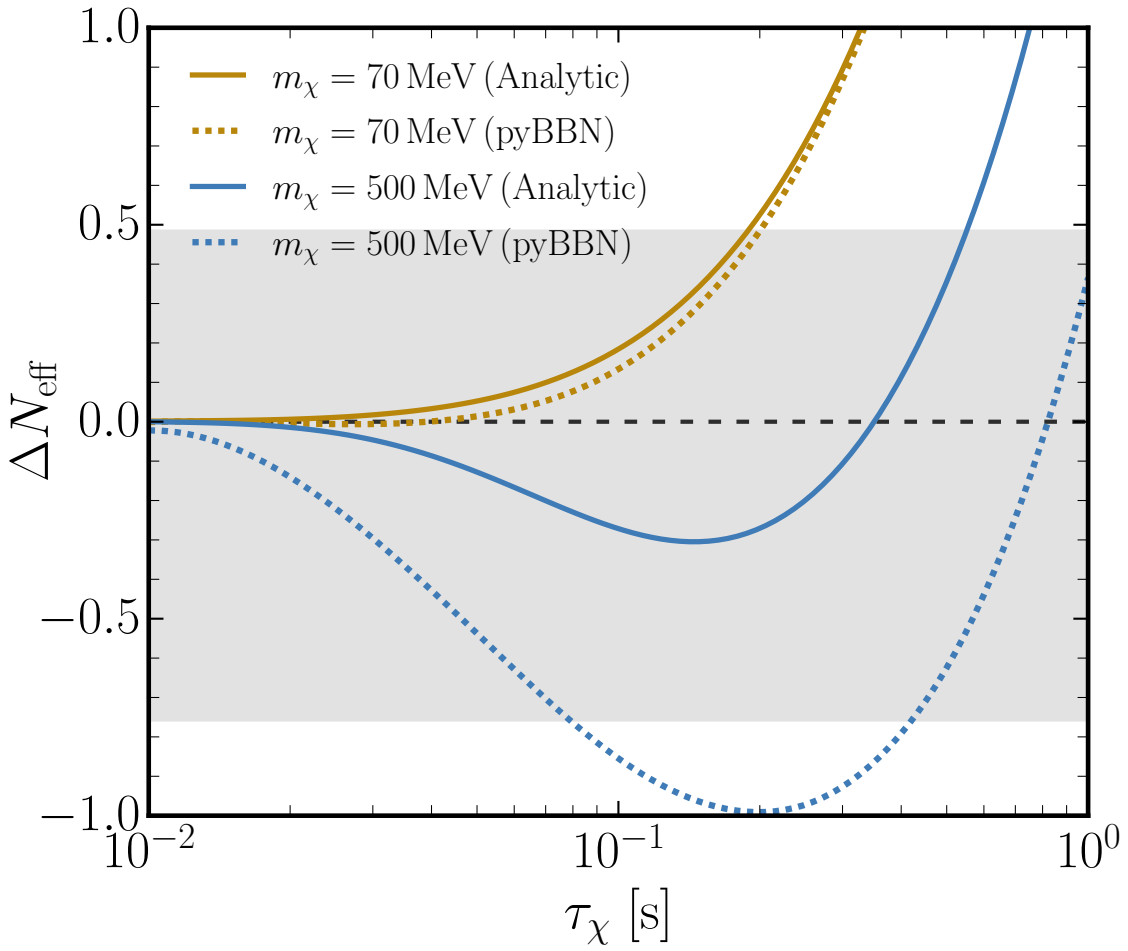

Figure 5: $\Delta N_{\rm eff}$ as a function of the lifetime of a FIP χ that can only decay into neutrinos through $\chi \to \nu_e + \nu_\mu + \overline{\nu}_\mu$. The initial FIP abundance is assumed to be $n_\chi/s = 0.01$ at $T = 1\,\text{GeV}$, where s is the total entropy density of a universe consisting of photons, neutrinos and electrons/positrons. The solid lines are the result of our semi-analytic model, while the dotted lines are obtained with the Boltzmann code `pyBBN`. The grey band is the current sensitivity by Planck (see text for details). The golden curves roughly indicate the lowest FIP mass for which $N_{\rm eff}$ can decrease due to the thermalization of the injected neutrinos. The stronger decrease of the blue, dotted curve as compared to the solid curve highlights the significance of residual neutrino spectral distortions in the evolution of $N_{\rm eff}$ (see Appendix 3.2.1.2 for more details).

As a final point, we can make a rough model-independent estimate for which neutrino energies the decrease of $N_{\rm eff}$ happens. In the particular FIP scenario considered here, we find that this effect occurs for masses higher than $\sim$70 MeV (see Fig. 5). Given that in this case the neutrinos are created via 3-body decays, this would correspond to an average injected neutrino energy of roughly $E_\nu^{\rm inj} \sim m_{\rm FIP}/3 \sim 25\,{\rm MeV}$.

> As long as a FIP injects most of its energy into neutrinos around neutrino decoupling, $N_{\rm eff}$ could decrease if neutrinos with energies of at least $E_\nu^{\rm inj} \sim 25\,{\rm MeV}$ are produced.

3.2.2 Summary

In this work, we have studied how heavy, unstable FIMPs that can decay into neutrinos impact the number of relativistic species $N_{\rm eff}$ in the Early Universe. A particularly interesting effect that could occur with these particles, is when they inject most of their energy into neutrinos but still decrease $N_{\rm eff}$. This could happen if FIMPs decay when neutrinos are still in (partial) equilibrium ($\tau_{\rm FIMP} \sim \mathcal{O}(0.1)\,{\rm s}$) and is a direct consequence of the thermalization process of the injected high-energy neutrinos (see Sec. 3.2.1 for a semi-analytical treatment of this effect). Here we identify neutrino spectral distortions as the driving power behind this effect, since they lead to an efficient transfer of energy from the neutrino plasma to the electromagnetic plasma (see Figs. 5 and 4). Some of the injected neutrino energy gets quickly transferred to the EM plasma, while the remaining will stay as residual spectral distortions in the neutrino distribution functions. These spectral distortions keep the energy transfer balance of $\nu \leftrightarrow {\rm EM}$ reactions shifted to the right till long after FIMP decay. In order to accurately account for this effect, it is therefore important to solve the Boltzmann equation and track the evolution of the neutrino distribution functions. Using a thermal-like distribution for neutrinos as an approximation can lead to incorrect results, e.g., that $N_{\rm eff}$ can never decrease when FIMPs inject most of their energy into neutrinos.

From our simulations, done with the publicly available Boltzmann code `pyBBN` [64], we find that this mechanism is especially relevant for FIMPs that can decay into neutrinos with average energies $E_\nu^{\rm inj} \gtrsim 25\,{\rm MeV}$. In case such neutrinos are created via 2- or 3-body decays, this roughly corresponds to FIMP masses $m_{\rm FIMP}^{\text{2-body}} \gtrsim 50\,{\rm MeV}$ and $m_{\rm FIMP}^{\text{3-body}} \gtrsim 70\,{\rm MeV}$ respectively. This is in agreement with the results presented in [193]. As such, this effect may be relevant for many classes of FIMPs[12], including Higgs-like dark scalars [38], dark photons [202], neutralinos in supersymmetric models with broken R-parity [203], vector portals coupled to anomaly-free currents [151] and short-lived neutrinophilic scalars [204].

[12] While `pyBBN` is mainly built to simulate the cosmological history in the presence of Heavy Neutral Leptons, it can in principle be modified to include many other classes of FIMPs.

3.3 Case study: HNLs

In this section, we consider applications of the findings of previous sections to the case of HNLs. In what follows, we will consider two quasi-degenerate HNLs [205, 206], as motivated by the Neutrino Minimal Standard Model (or νMSM) [see e.g. 207–209])

HNLs alter the cosmological history through their contribution to the total energy density of the Universe and their decay into SM particles. HNLs that decay well before the decoupling of active neutrinos, i.e. at temperatures $T \gg 1$ MeV, will leave no traceable impact. On the other hand, if HNLs live long enough, they could alter several physical quantities, such as N_{eff} and the primordial abundances of light elements [192, 194, 195, 210–212]. Indeed, strong limits have been set on their mass and lifetime by considering their impact on Big Bang Nucleosynthesis and the Cosmic Microwave Background, see e.g. [37, 64, 196] for recent works on this subject.

The influence of HNLs on BBN and CMB depends on their abundance, and we will first discuss how HNLs are produced in the primordial plasma (Sec. 3.3.1). In Sec. 3.3.3.1, we derive the bounds from BBN, while in Sec. 3.3.4 we consider the impact of HNLs on CMB.

3.3.1 Thermal history of HNLs

At large temperatures, the interaction rate of HNLs with SM particles is temperature-suppressed, although the particle densities are high.

Indeed, in the plasma without lepton asymmetry at temperatures $T \gtrsim 1$ GeV the effective mixing angle is given by [213, 214]

$$U_{\text{m}}^2(T) \approx \frac{U^2}{\left[1 + 9.6 \cdot 10^{-24} \left(\frac{T}{1\text{ MeV}}\right)^6 \left(\frac{m_N}{150\text{ MeV}}\right)^{-2}\right]^2}, \tag{3.3.1}$$

see Appendix 3.A. As a result, the interaction rate of HNLs with SM particles $\Gamma_N^{\text{int}} \propto G_F^2 T^5 U_{\text{m}}^2$ is suppressed at both high and low temperatures and reaches its maximum at the temperature

$$T_{\max} \approx 12 (m_N/1\text{ GeV})^{1/3}\text{ GeV} \tag{3.3.2}$$

(see Fig. 6).

The HNLs were in thermal equilibrium if during some period $T_- < T < T_+$ the interaction rate $\Gamma_N^{\text{int}}(T)$ exceeded the Hubble expansion rate. For heavy HNLs with

masses $m_N \gtrsim 50$ MeV, this happens for mixing angles larger than

$$U^2 \gtrsim U^2_{\text{min}} \approx 3 \cdot 10^{-12} \left(\frac{1 \text{ GeV}}{m_N} \right) \tag{3.3.3}$$

Namely, using the condition $\Gamma_N^{\text{int}}(T_{\max}) = 3H(T_{\max})$, and approximating the interaction rate as $\Gamma_N^{\text{int}} \approx 10 U_m^2 G_F^2 T^5$, we find the minimal value on the mixing angle at which HNLs may enter the equilibrium, Eq. (3.3.3).

Notice that if HNLs are responsible for the generation of neutrino masses, there exists another lower bound on the mixing angle – the seesaw bound. At least one HNL with mass m_N should have mixing angle above this bound to be responsible for the generation of the atmospheric neutrino mass difference, *c.f.* [82]. The bound depends on details of the given HNL model – mixing pattern and neutrino mass hierarchy (see, e.g., [77, 208]). For simplicity, as the scale of the see-saw bound we will use the toy-model estimate

$$U^2 \gtrsim U^2_{\text{see-saw}} \simeq 5 \cdot 10^{-11} \left(\frac{1 \text{ GeV}}{m_N} \right) \tag{3.3.4}$$

The true see-saw bound may differ from the toy model estimate by within an order of magnitude.

The resulting parameter space of HNLs is shown in Fig. 6.

It is convenient to parametrize the population of HNLs in terms of the abundance, defined by

$$Y_N = \left(\frac{n_N}{s} \right)_{T=T_-}, \tag{3.3.5}$$

where n_N is the number density of HNLs and $s = g_* \frac{2\pi^2}{45} T^3$ is the entropy density.

3.3.1.1 HNLs with mixing angles below U_{min}

Let us now calculate the abundance of HNLs that never enter thermal equilibrium, i.e. of those with $U^2 \lesssim U^2_{\text{min}}$. The temperature evolution of the HNL abundance, $\mathcal{Y}_N$, may be found with the help of a simple equation

$$\frac{d\mathcal{Y}_N}{dt} = -\Gamma_{N,\text{int}} (\mathcal{Y}_N - \mathcal{Y}_{N,\text{eq}}), \tag{3.3.6}$$

where $\mathcal{Y}_{N,\text{eq}}(T)$ is the abundance of HNLs at equilibrium, Y_N is defined as the value of $\mathcal{Y}_N(T \ll T_{\text{max}})$, and $\Gamma_{N,\text{int}}$ is the total rate of processes $A + N \to X$. At temperatures $T \gg m_N$, we may approximate the rate $\Gamma_{N,\text{int}}$ by an expression

$$\Gamma_{N,\text{int}} \approx b G_F^2 T^5 \cdot U^2_{\text{m}}(T), \tag{3.3.7}$$

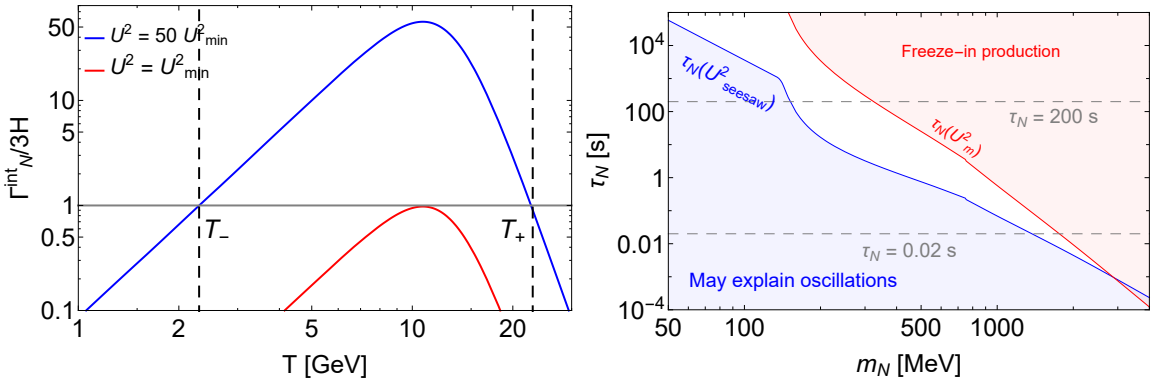

Figure 6: *Left panel*: The reaction rate of the HNL with SM particles, $\Gamma_N^{\rm int}$, compared to the Hubble rate, $H(T)$. T_+ and T_- are the temperatures at which HNLs enter and exit the thermal equilibrium. For illustration, we used HNL mass $m_N = 1$ GeV, and mixing angles $U^2 = U^2_{\rm min}$ and $U^2 = 50U^2_{\rm min}$, see Eq. (3.3.3). *Right panel*: the parameter space of HNLs that mix purely with ν_e. The blue domain roughly denotes the parameter space of HNLs that may explain neutrino oscillations, see Eq. (3.3.4). The red domain defines the parameter space for which HNLs never enter thermal equilibrium, see Eq. (3.3.3). The dashed scale $\tau_N = 0.02$ s denotes the shortest lifetime that may be constrained by BBN (the effect of the meson-driven $p \leftrightarrow n$ conversion as discussed in Sec. 3.1), while the scale $\tau_N = 200$ s defines the onset of nuclear reactions

where $b(T)$ is a factor depending on the number of SM species present in the primordial plasma. Also, if $T \gg \Lambda_{\rm QCD}$, we may use $g_* \approx 86.25$ [215], and the equilibrium abundance is $\mathcal{Y}_{N,\rm eq} \approx 0.01$ for Dirac HNLs and 0.005 for Majorana HNLs.

Using matrix elements for processes $N + A \to B + C$ from [64], we find $b \approx 10$ for Dirac HNLs (correspondingly, $b \approx 20$ for Majorana HNLs) at $T \gtrsim 1$ GeV. The value of b for Majorana HNLs is a factor 5 larger than that is used in [196], $b \approx 3.6$ (Majorana neutrinos are considered). A reason is that [196] uses rates from [216], where temperatures below 20 MeV are considered (see Eq. (6.8) from [216]), and hence A, B, C may be $e^\pm, \nu$ only, which is a huge underestimate.

Using (3.3.1), Eq. (3.3.12) may be integrated to obtain the final abundance of HNLs Y_N:

$$\int\limits_0^{Y_N} \frac{d\mathcal{Y}_N}{\mathcal{Y}_N - \mathcal{Y}_{N,\rm eq}} = -\int\limits_0^{\infty} dT \frac{\Gamma_{N,\rm int}(E_N)}{TH(T)} \approx -2.8\cdot 10^7 \frac{b}{\sqrt{a}} \frac{m_N}{1\ \text{GeV}} U^2 \Rightarrow \tag{3.3.8}$$

$$Y_N = \mathcal{Y}_{N,\rm eq}\left(1 - e^{-6\cdot 10^{11}\frac{m_N}{1\ \text{GeV}}U^2}\right) \tag{3.3.9}$$

Using $U^2 \ll U^2_{\rm min}$, we find that the abundance of HNLs that never entered thermal equilibrium is given by

$$Y_N \approx 2.8 \cdot 10^7 \cdot \mathcal{Y}_{N,\rm eq} \cdot 6 \cdot 10^{11} \frac{m_N}{1\ \rm GeV} U^2 \approx 5.7 \cdot 10^{10} \frac{m_N}{1\ \rm GeV} U^2 \tag{3.3.10}$$

3.3.1.2 HNLs with mixing angles above $U_{\rm min}$

Let us now consider HNLs with mixing angles above $U_{\rm min}$. It is important *(i)* whether HNLs froze out while being ultra-relativistic (UR regime, $m_N \ll T_-$, no exponential Boltzmann suppression for the number density) or non-relativistic (NR regime, $m_N \gg T_-$) and *(ii)* the value of g_* at the moment of the decoupling (depending on decoupling temperature it can change rapidly - see left Fig. 7). Using $n_N = \frac{3}{4} 2 \frac{\zeta(3)}{\pi^2} T^2$ for the UR regime or $n_N \sim \left(\frac{m_N}{T_-}\right)^{3/2} e^{-m_N/T_-}$ for the NR regime, we get the abundance in these two limits:

$$Y_N \simeq \begin{cases} \dfrac{0.6}{g_*(T_-)}, & \text{UR regime} \\ \alpha(m_N, \tau_N) \left(\dfrac{m_N}{T_-}\right)^{3/2} e^{-m_N/T_-}, & \text{NR regime} \end{cases} \tag{3.3.11}$$

The coefficient $\alpha(m_N, \tau_N)$ in Eq. (3.3.11) appears since the decoupling is not an instantaneous process; in dependence on the mass and lifetime it can vary by a factor of $\mathcal{O}(10)$.

To improve these estimates, we find the abundance numerically. We assume the Boltzmann approximation for the distribution function of the plasma particles and the equilibrium shape of the energy distribution of HNLs (such that, in particular, $\langle E_N \rangle$ is $3.15T$ for $T \gg m_N$). In this case, the equation for the evolution of the abundance of HNLs has the form (see, e.g., [217])

$$\frac{d\mathcal{Y}_N}{dt} = -\Gamma_{N,\rm int}(\mathcal{Y}_N - \mathcal{Y}_{N,\rm eq}), \tag{3.3.12}$$

where $\Gamma_{N,\rm int}$ is given by

$$\Gamma_{N,\rm int} = \sum_{A,B,C} \frac{g_N g_A}{8\pi^4} \int\limits_{s_{\rm min}}^{\infty} p^2_{AN} \sqrt{s} \sigma_{N+A\to B+C} K_1\left(\frac{\sqrt{s}}{T}\right) ds \tag{3.3.13}$$

In $\Gamma_{N,\rm int}$, the threshold invariant mass is $s_{\rm min} = \min[(m_N + m_A)^2, (m_B + m_C)^2]$, and

$$p^2_{AN} = \frac{s}{4}\left(1 - \frac{(m_N - m_A)^2}{s}\right)\left(1 - \frac{(m_N + m_A)^2}{s}\right) \tag{3.3.14}$$

The values of abundances Y_N for particular lifetimes are shown in Fig. 8 (right panel).

For masses 0.2 GeV $\lesssim m_N \lesssim 2$ GeV and lifetimes above 0.001 s, HNLs decouple at temperatures $T \gg m_N$, i.e. being ultra-relativistic and above the QCD transition. Their abundance Y_N is therefore universal and almost constant, owing to the temperature dependence of g_* (left panel) for $T \gtrsim 200$ MeV. For masses $\mathcal{O}(100)$ MeV and for large lifetimes $\tau_N \gtrsim 0.1$ s, HNLs still decouple while being ultra-relativistic. With the decrease of τ_N, Y_N first grows by a factor of few (due to rapid decrease of g_*), and with further decrease it becomes strongly suppressed.

Eq. (3.3.11) means that for UR regime later decoupling (*i.e.* larger mixing angles) leads to larger HNL abundance. In Fig. 6 (the right panel), we summarize the HNL

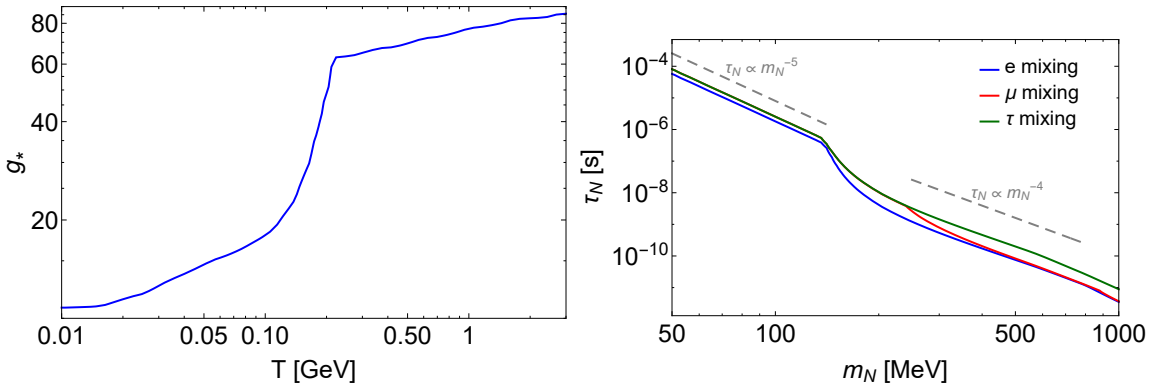

Figure 7: *Left panel*: temperature dependence of g_* in SM (reproduced from [215]). The drop around $T = 200$ MeV is caused by the entropy dilution at the QCD transition. *Right panel*: HNL lifetime as a function of mass for mixing with different flavors. The dashed gray lines show the scaling of the lifetime with mass. The lifetimes is shown for $U^2 = 1$ and scales as U^{-2}.

parameter space explored by the current study. It shows the domain in which HNLs never entered thermal equilibrium as well as the regime in which HNLs decouple while being non-relativistic. We see that these two regimes are separated by the broad parameter space for which HNLs enter thermal equilibrium and decouple while being UR. A dashed line in the middle of this region is the seesaw bound (3.3.4).

The temperature of freeze-out (T_-) is roughly defined via

$$\Gamma_N^{\rm int}(T_-) \simeq 3H(T_-), \tag{3.3.15}$$

see left panel in Fig. 6 (this equation has two solutions, the larger one defines T_+). The

factor of two estimate for T_- reads

$$T_- \simeq T_{\nu,\rm dec} \times \begin{cases} \dfrac{1}{U^{2/3}} \dfrac{1}{n_{\rm int}^{1/3}} \left(\dfrac{g_*(T_-)}{10.75}\right)^{1/6}, & \text{UR regime} \\ \dfrac{1}{U^2} \dfrac{1}{n_{\rm int}} \left(\dfrac{100\ \text{MeV}}{m_N}\right)^2 \left(\dfrac{g_*(T_-)}{10.75}\right)^{1/2} & \text{NR regime,} \end{cases} \tag{3.3.16}$$

where $T_{\nu,\rm dec} \approx 1.4$ MeV is the decoupling temperature of active neutrinos, $n_{\rm int} = \Gamma_{N,\rm int}/G_F^2 T^5$ is a factor that varies from $\simeq 2$ at $T \simeq \mathcal{O}(1\ \text{MeV})$ to $\simeq 9$ at $\mathcal{O}(1\ \text{GeV})$ temperatures. The different dependence on U^2 and on m_N in two regimes is due to the change of centre-of-mass energy ($E_{\rm cm} \sim T$ for UR and $E_{\rm cm} \approx m_N$ in the NR regimes). The values of T_- for different masses are shown in Fig. 8 (left). Instead of the mixing angles we use the lifetime $\tau_N \propto U^{-2}$ (Fig. 7, left) that is more intuitive when studying the influence on BBN.

For masses around $m_N \simeq 200$ MeV and lifetimes $\tau_N \sim 0.1$ s the HNL freeze-out occurs around the hadronization epoch. During this epoch, g_* drops by a factor ~ 3 [215] while $T_- \simeq m_N$, and therefore the abundance of HNLs can be *higher than for relativistic decoupling*.

For smaller masses, the decoupling temperature rapidly drops (see Fig. 8, left panel), which results in the Boltzmann suppression of the abundance. This effect is translated into a factor of few "kink" below the mass $m_N \simeq 200$ MeV in the final plots (Figs. 12).

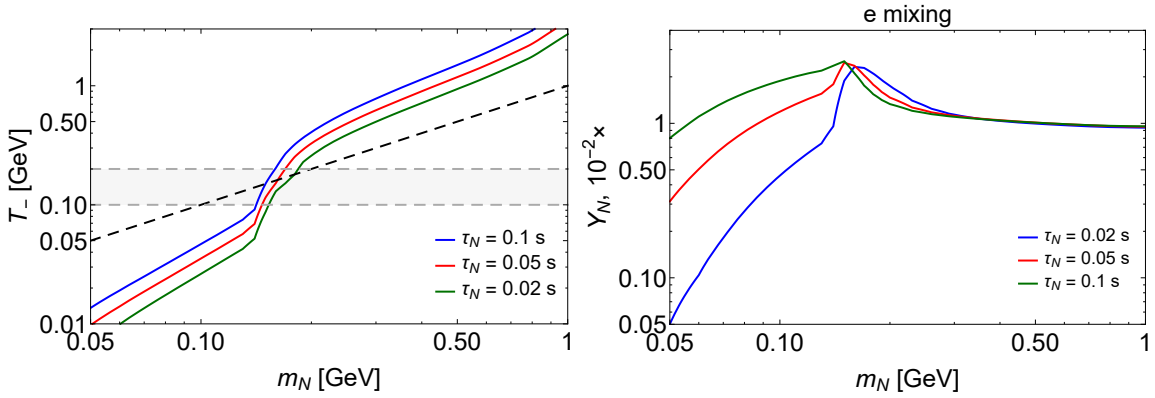

Figure 8: *Left panel*: the behavior of the decoupling temperature T_-, defined via $\Gamma_{N,\rm int}(T_-) = 3H(T_-)$, versus the HNL mass for particular lifetimes. The black dashed line defines the parameter space $T_- = m_N$, which roughly indicates the transition from relativistic to non-relativistic regime of HNL decoupling. The gray horizontal band shows a temperature when the hadronization of quarks takes place, and therefore the effective number of relativistic degrees of freedom, g_*, drops sharply (remind Fig. 7). *Right panel*: HNL abundances versus the HNL mass for particular values of the lifetime.

3.3.1.3 Evolution after decoupling

After the freeze-out, the comoving number density of HNLs changes only due to HNL decays. The physical number density thus evolves as

$$n_N(T) = n_N(T_{\text{dec}}) \cdot \left(\frac{a(T_{\text{dec}})}{a(T)}\right)^3 \cdot e^{-t/\tau_N} \tag{3.3.17}$$

Decays of HNLs inject energy into the primordial plasma. This effect changes the time-temperature relation and the scale factor evolution as compared to SBBN. The HNL decays provide additional dilution of any decoupled relics (including themselves) in comparison to the SBBN case:

$$\zeta = \left(\frac{a_{\text{SBBN}}}{a_{\text{SBBN + N}}}\right)^3 < 1, \tag{3.3.18}$$

where $a^{-1}_{\text{SBBN}}(T) \propto g_*^{1/3} T$ is the scale factor in SBBN, and the scale factors are evaluated at times $t \gg \tau_N$. To calculate ζ, we solve the Friedmann equation under an assumption that neutrinos are in perfect equilibrium and neglecting the mass of electrons:

$$\begin{aligned} H^2(t) =& \frac{1}{M_{\text{Pl}}^2}\frac{8\pi}{3}\Big[\rho_{\text{rad}} + m_N \cdot n_N(T)\Big], \\ 4\frac{\rho_{\text{rad}}}{T}\frac{dT}{dt} =& \frac{m_N n_N(t)}{\tau_N} - 4H(t)\cdot\rho_{\text{rad}}, \end{aligned} \tag{3.3.19}$$

where the number density of HNLs is given by Eq. (3.3.17). This is a reasonable assumption, since most of the HNLs with lifetimes $\tau_N \ll 0.1$ s decay much earlier than neutrinos decouple.

Effects of meson-driven conversion force us to trace the number density of HNLs even at times $t \gg \tau_N$:

$$n_N(t \gg \tau_N) = n_N(T_-) \cdot \left(\frac{a_{\text{SBBN}}(T_-)}{a_{\text{SBBN}}(T)}\right)^3 \cdot \zeta \cdot e^{-t(T)/\tau_N} \approx 0.4 Y_N \cdot g_{*,\text{SBBN}}\, T^3 \cdot \zeta \cdot e^{-t(T)/\tau_N}, \tag{3.3.20}$$

where $t(T)$ is the same as in SBBN.[13] Because of the suppression, the effect of this population on the expansion of the Universe may be neglected. However, this exponential tail still may produce mesons in amounts sufficient to change the dynamics of the n/p ratio.

The values of the HNL abundance and the dilution factor versus its mass and lifetime are given in Fig. 9.

[13] At times $t \gg \tau_N$, the time-temperature relation differs from SBBN only by the value of N_{eff}. However, the latter may change only if neutrinos are not in perfect equilibrium, and hence $t(T)$ is the same as in SBBN for lifetimes $\tau_N \ll 0.1$ s.

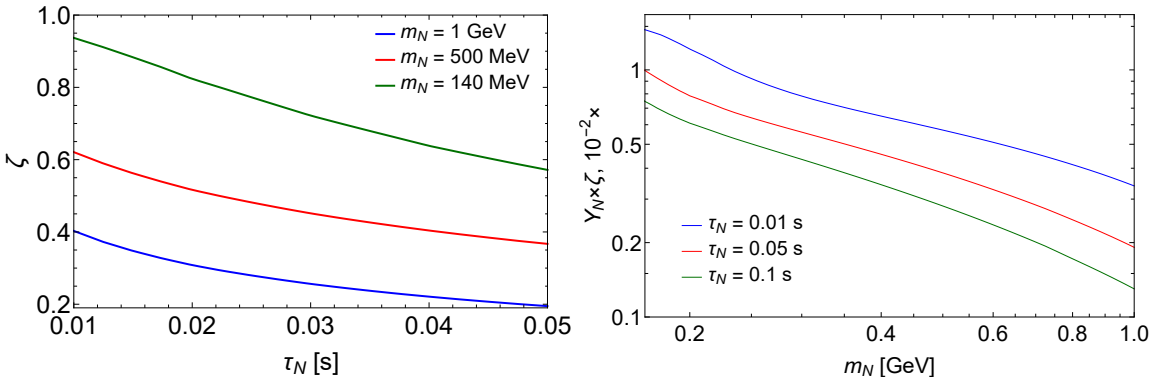

Figure 9: *Left panel*: Dilution factor (3.3.18) for short-lived HNLs mixing with ν_e. *Right panel*: HNL abundance times dilution factor as a function of mass for particular values of the lifetime. Details of the calculation of the abundances and ζ are given in [218]. Dilution factor is calculated, when most of HNLs has decayed and do not contribute to entropy density. Note, that we define abundance at the moment of decoupling, hence it does not change with decays.

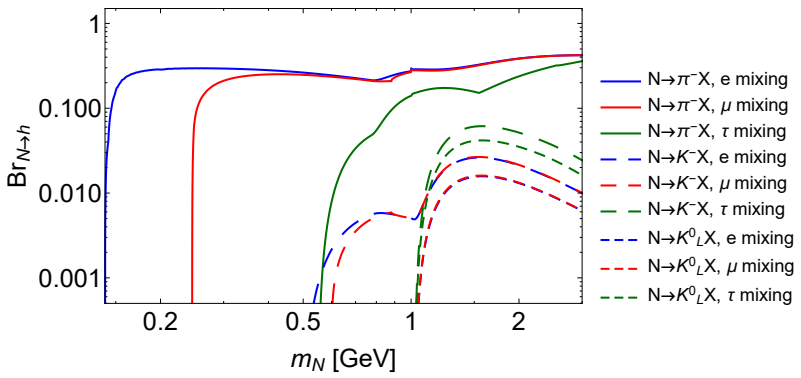

Figure 10: Branching ratios of HNL decays into mesons $h = \pi^-, K^-, K^0_L$. Secondary decays are also included (see text for details).

3.3.2 Hadronic decays of HNLs

In this work, we consider a pair of HNLs, degenerate in mass and having similar mixing angles. Two such HNLs form a single quasi-Dirac fermion [205, 219]. The abundance of a meson h produced from such HNLs is proportional to the quantity $Y_N \cdot \text{Br}_{N\to h}$. The mass dependence of $\text{Br}_{N\to h}$ for different mesons h and mixing patterns is shown in Fig. 10. We are interested only in the abundances of light mesons (pions and kaons) and for HNL masses well above pion/kaon thresholds we should account for "secondary mesons". This is discussed below, mainly following [39]. **Decays into pions.** In the case of the pure e/μ

mixings, the charged pion production threshold corresponds to $m_N = m_\pi + m_l$, where $l = e/\mu$. For τ mixing, the similar charged current-mediated channel opens up only at $m_N = m_\tau + m_\pi \simeq 1.9$ GeV. However, for all types of mixings charged pions may appear

as secondary particles in decays of neutral mesons,

$$N \to h^0 + \nu_\alpha, \quad h^0 \to \pi^\pm + X, \quad \text{where} \quad h^0 = \rho^0, \eta^0, \eta', \omega^0, \phi \tag{3.3.21}$$

Therefore, for τ mixing charged pions may appear at masses $m_N \geq m_{\eta^0}$. We use the branching ratios $\text{Br}_{\eta^0 \to \pi^\pm X} \approx 0.27$, $\text{Br}_{\rho^{0,\pm} \to \pi^\pm X} \approx 1$ [52].

Above $m_N \simeq 1$ GeV, decays of HNLs into pions cannot be approximated by single meson decays. Indeed, decays of GeV mass range HNLs are similar to decays of τ lepton [39], whereas for the latter hadronic decays are dominated by multi-pion channels [52]. We estimate the width of multi-pion decays as the difference between the total width into quarks and the width into single mesons:

$$\Gamma_{N \to n\pi} = \Gamma_{N \to \text{quarks}} - \sum_{h=\pi,K,\rho,\ldots} \Gamma_{N \to hX} \tag{3.3.22}$$

For multiplicities $\mathcal{N}$ of decays of HNLs into charged pions $N \to \pi^\pm$ (i.e., the amount of $\pi^\pm$ per multi-hadronic decay of HNLs), we will use multiplicities for multihadronic decays of τ leptons. Namely, $\mathcal{N}_{N \to \pi^+} = \mathcal{N}_{\tau^+ \to \pi^+} \approx 1.35$, $\mathcal{N}_{N \to \pi^-} = \mathcal{N}_{\tau^+ \to \pi^-} \approx 0.34$. The effective branching into π^- from multi-pion decays is

$$\text{Br}^{\text{multi-pion}}_{N \to \pi^-} = \mathcal{N}_{N \to \pi^-} \cdot \frac{\Gamma_{N \to n\pi}}{\Gamma_N}, \qquad \text{Br}^{\text{multi-pion}}_{\bar{N} \to \pi^-} = \mathcal{N}_{\bar{N} \to \pi^-} \cdot \frac{\Gamma_{N \to n\pi}}{\Gamma_N} \tag{3.3.23}$$

Since the bound on the meson driven $p \leftrightarrow n$ conversion is only logarithmically sensitive to the value of $\text{Br}_{N \to \pi^\pm}$, our results depend on these assumptions weakly.

Decays into kaons. Below $m_N = m_\phi$, charged kaons may appear only through the mixing with e/μ in the process $N \to K^- l$. This decay is Cabibbo suppressed [39] and almost two orders of magnitude smaller than into pions. Neutral kaons appear only in the final states with three or more particles (such as $N \to K^0 + \bar{K}^0 + \nu_\alpha$ and $N \to K^+ + \bar{K}^0 + \ell^-$, etc).

HNLs heavier than ϕ meson may produce both charged and neutral kaons via decays $N \to \phi\nu, \phi \to KK$. We assume that K^0 contains equal admixtures of K^0_L and K^0_S, i.e. $\text{Br}_{N \to K^0_L} = \text{Br}_{N \to K^0}/2$. We use the branching ratios $\text{Br}_{\phi \to K^-} \approx 0.5$, $\text{Br}_{\phi \to K^0_L} \approx 0.34$ [52].

3.3.3 Bounds from BBN and CMB

3.3.3.1 BBN

Let us first consider the bound from BBN. From the previous section we conclude that in dependence on the mixing patter, decays into charged pions become possible for HNLs with masses from $m_N = m_\pi + m_e$ (for the pure e mixing) to $m_N = m_\eta \approx 547$ MeV (for the pure τ mixing), see [39, 220] or Sec. 3.3.2.

HNLs with minimal lifetimes that may be constrained by BBN, $\tau_N \simeq \mathcal{O}(0.02\text{ s})$, are produced thermally, remind Fig. 6. For such HNLs, in Eq. (3.1.13) $n_{N,\text{dec}}/n_\gamma(T_{\text{dec}}) \approx 3/2$.

The dilution factor $(a_{\text{dec}}T_{\text{dec}}/a_0T_0)^3$ is of $0.1 - 0.6$ for HNL masses $m_\pi \lesssim m_N \lesssim 3$ GeV (we will use $\frac{1}{3}$ for normalization below), see Appendix 3.3.1.

Using values of $\text{Br}_{N\to h}$, P_{conv} and the scale factors ratio (which may be found in Appendix 3.B and sections 3.3.2 3.3.1), we conclude that the logarithm term in (3.1.13) is $\mathcal{O}(1)$ for HNLs in the mass range $m_N = \mathcal{O}(1\text{ GeV})$ and affects the overall bound very weakly. Therefore, the bound depends only on $T_0^{\min}$.

The maximal admissible correction (3.1.14) is reached for $T_0^{\min} = 1.50$ MeV, almost independently on HNL mass (see Fig. 2 and Appendix 3.B.1).

Plugging $T_0^{\min} = 1.50$ MeV into (3.1.13), we obtain our final limit from the analytic estimates

$$\tau_N \lesssim 0.023 \text{ s}. \tag{3.3.24}$$

Numeric calculations from Appendix 3.B.1 confirm this result, predicting constraints at the level of $0.019 - 0.021$ s.

Let us now comment on the maximal lifetimes for which our bounds are applicable. In Sec. 3.1.3, we restricted the applicability of our bounds by lifetimes $\tau_{\text{FIP}} \simeq \mathcal{O}(50\text{ s})$ – which is an estimate derived under the assumption of absence of effects of FIPs during the nuclear reaction chain. To push the bound to larger lifetimes, we solve numerically the system of equations for abundances of $d, t,^3$He,^{4}He,^{7}Li,^{7}Be in presence of mesons from decaying HNLs. We incorporate the change of the time-temperature relation and the dynamics of η_B via the Friedmann equations with HNLs. We use the nuclear rates and reactions chain from [48]. Further description of our numeric approach may be found in Appendix 3.B.2. The example of the temperature behavior of the nuclear abundances in presence of HNLs is shown in Fig. 11.

Using the numeric approach, we conclude that long-lived HNLs with lifetimes $\tau_N \lesssim 10^4$ s increase nuclear abundances of all elements. The behavior of abundances with HNL lifetime for a particular mass $m_N = 200$ MeV is shown in Fig. 11.

For larger lifetimes, we need to include the effect of photo-dissociation. Therefore, using numeric approach, we have extended the domain of applicability of BBN bounds on HNLs to lifetimes $\tau_N = 10^4$ s.

3.3.3.2 Results

We demonstrated that HNLs with semi-leptonic decay channels significantly affect the primordial ^{4}He abundance, as mesons from their decays drive the $p \leftrightarrow n$ conversion rates away from their SBBN values (*c.f.* [178, 181, 182]). In order to avoid ^{4}He overproduction, mesons should disappear from the primordial plasma by $T = T_0^{\min} \simeq 1.50\,\text{MeV}$. The

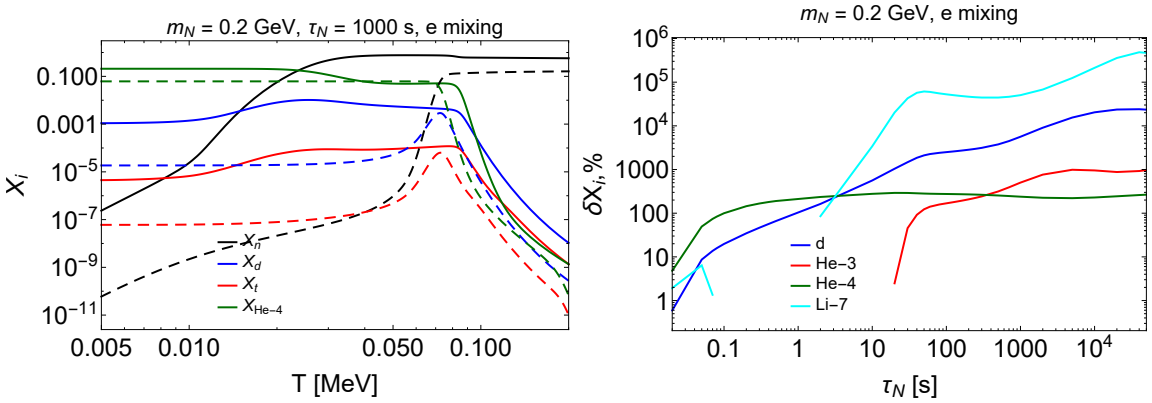

Figure 11: *Left panel*: the temperature behavior of the nuclear abundances in presence of HNLs (the solid lines), as well as their behavior in SBBN (the dashed lines). *Right panel*: the behavior of the change of nuclear abundances $\delta X_i \equiv (X_i - X_i^{(\mathrm{SBBN})})/X_i^{(\mathrm{SBBN})}$ with HNL lifetime. In both figures, an HNL with mass $m_N = 200$ MeV and pure e mixing is considered.

neutron abundance will then have enough time to relax down to its SBBN value before the onset of deuteron formation. These requirements severely constrain the parameter space of the HNLs with $0.023\ \mathrm{s} \leq \tau_N \leq 10^4\ \mathrm{s}$ for masses $m_N > 140\,\mathrm{MeV}$.

The final bounds for different mixing patterns are shown in Figs. 13 and 12. Our constraints can be generalized to other HNL models, see e.g. [221].

Confronted with the bounds from accelerator searches, we ruled out HNLs with mass below 500 MeV (for electron mixing) and 350 MeV (for muon mixing). Moreover, tighter bound means that future searches at Intensity Frontier (specifically, SHiP experiment [80]) can reach the BBN bottom line and completely rule out HNLs with the masses up to 750 MeV, which was not the case before [see e.g. 67, 222].

The comparison with the previous results [64, 194, 195] is shown in Fig. 13 (right panel). Our bound (3.3.24) is a factor of ~ 5 stronger than the previous result [194]. The recent reanalysis [64] did not take into account the effects of mesons, therefore their results are a factor $2-3$ less conservative.

The clear qualitative effect discussed in this paper not only leads to a tighter bound on HNL lifetime and provides an reachable goal for experimental searches, but also allows for an analytic description, unusual in the realm of BBN predictions driven by sophisticated numerical codes.

3.3.4 Bound from CMB

Let us now discuss the influence of HNLs on the physics at the CMB epoch. We do not consider masses higher than $m_N \simeq 1$ GeV, since there is no adequate description of HNL decay widths due to theoretical uncertainties [39], while it is crucial to know them for the calculation of N_{eff}. Indeed, this makes it complicated to compute ξ_ν (and thus ΔN_{eff}), as it depends on the branching ratios of the different multi-meson decay channels. For instance,

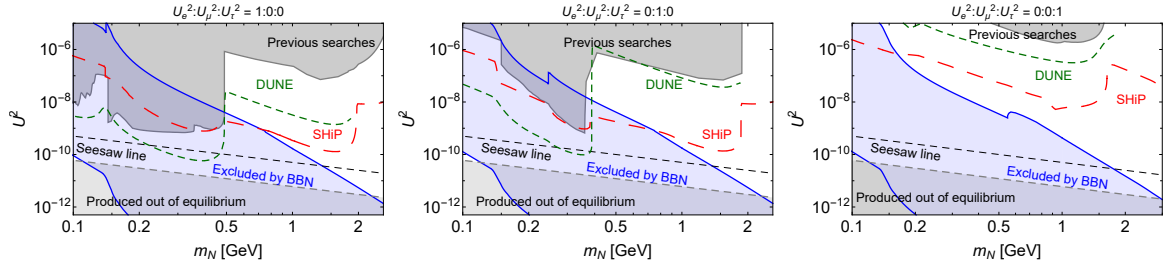

Figure 12: Bounds for HNLs mixed with a particular flavor. The blue area is excluded by our present analysis combined with [218] (for HNL masses below the charged pion production threshold). The dark gray area denotes the excluded HNL parameter space from previous searches [82], including the latest NA62 search [223]. The red and greed dashed lines show the sensitivity of several future intensity frontier experiments with the highest sensitivity in the regions of interest – SHiP [80, 224] and DUNE [225–227] (see [67]). Finally, the black dashed line denotes the seesaw bound applicable if two degenerate in mass HNLs are responsible for neutrino oscillations (as in the νMSM) [77, 82]. Our bounds are applicable up to lifetimes $\tau_N = 10^4$ s, from which EM decay products of HNLs may lead to nuclear photodissociation, see text for details.

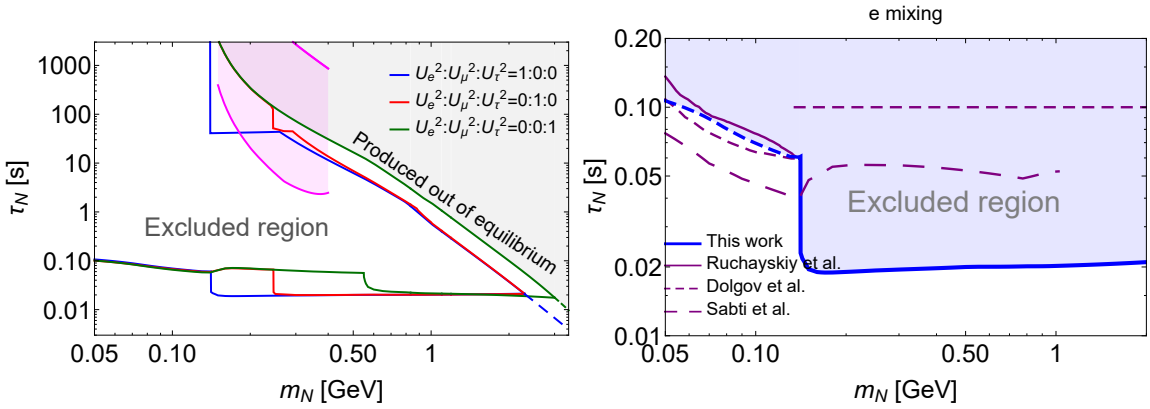

Figure 13: *Left panel*: BBN bounds on HNL lifetime for different mixing patterns. The gray region is excluded as a result of this work (for masses below pion threshold we use the results of [218]). The magenta shaded region corresponds to the domain excluded in [196]. *Right panel*: comparison of the results of this work (thick blue line) with the results of the previous works [64, 194, 195] (purple lines) assuming mixing with electron flavor only. Notice that other works have adopted different values for the maximally admissible ^{4}He abundance when deriving their bounds: $Y_{p,\max} = 0.2696$ in [194, 195] and $Y_{p,\max} = 0.253$ in [64] as compared to $Y_{p,\max} = 0.2573$ in this work (see text for details).

the decay $N \to 3\pi^0 + \nu$ injects more energy into the EM plasma and diminishes ξ_ν, while $N \to 3\pi^\pm + \ell^\mp$ may inject more energy into neutrinos and compensate for this decrease. Therefore, both such channels should be accounted for.

We make use of `pyBBN` [64] to simulate their impact on the cosmological history, in

particular on N_{eff}. We examine the region of parameter space in which HNLs inject most of their energy into neutrinos, but where ΔN_{eff} is negative, illustrating the effect described in the previous section. Finally, we derive bounds from the CMB and comment on the possible role of HNLs in alleviating the Hubble tension.

3.3.4.1 Behavior of N_{eff}

HNLs inject (eventually) all of their energy either into the neutrino or electromagnetic plasma. The fraction of the HNL energy that is injected into each of these two sectors is mass-dependent and shows a significant shift to the EM plasma once HNLs can decay into neutral pions ($\sim$135 MeV), see Fig. 16.

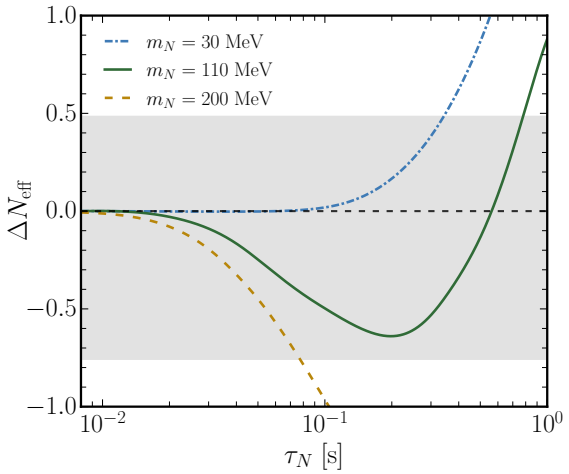

Figure 14: ΔN_{eff} as a function of HNL lifetime for a number of benchmark masses. Mixing with electron neutrinos only is considered here. The curves illustrate three cases of how HNLs can affect N_{eff}: *1)* they can decay mostly into neutrinos and simply increase N_{eff} (30 MeV curve), *2)* they can decay mostly into neutrinos and either decrease or increase N_{eff} depending on their lifetime (110 MeV curve), and *3)* they can decay mostly into EM particles and simply decrease N_{eff} (200 MeV curve). HNLs with masses $m_N \gtrsim 70$ MeV that decay mainly into neutrinos around neutrino decoupling, show an initial decrease of ΔN_{eff} as a result of the thermalization of the injected high-energy neutrinos. The grey band is the current sensitivity by Planck.

This plot shows that HNLs below the pion mass decay mainly into neutrinos and, therefore,

One would naively expect that in the mass range $m_N < m_\pi$ N_{eff} increases. However, we find that HNLs are able to decrease N_{eff} for masses already above $\sim$70 MeV, while for smaller masses an increase of N_{eff} is observed.

The origin of this sign change in ΔN_{eff} at $m_N \gtrsim 70\,\text{MeV}$ (rather than $m_N > m_\pi$ as one would guess from Fig. 16) lies in the energy transfer from the neutrino plasma to the electromagnetic plasma that is induced by the injected non-equilibrium neutrinos, as discussed earlier in Sec. 3.2.1. We run `pyBBN` simulations to examine in which region of parameter space this sign change happens[14]. We show ΔN_{eff} as a function of the HNL lifetime in Fig. 14 for a number of benchmark masses. The grey band in this figure indicates the current sensitivity by Planck. Included in this figure is an HNL of mass $110\,\text{MeV}$, which decreases N_{eff} for lifetimes below $\tau_N \lesssim 0.6\,\text{s}$ and increases N_{eff} for longer lifetimes. Such a lifetime ($\tau_N \sim 0.6\,\text{s}$) roughly corresponds to the time of neutrino decoupling, beyond which thermalization between the neutrino and EM plasma is not efficient anymore and the injected neutrinos remain in the neutrino sector. This exemplifies the ability of HNLs below the pion mass to diminish N_{eff}, even when neutrinos are on the verge of being completely decoupled. With the current sensitivity of Planck, however, this initial decrease of ΔN_{eff} for this mass falls within the error range and is thus not observable. Nevertheless, a number of upcoming and proposed CMB missions, such as the Simons Observatory [228] and CMB-S4 [229], could provide a determination of N_{eff} around the percent-level and probe this effect.

We depict the region of HNL parameter space where ΔN_{eff} changes sign in the top panel of Fig. 15. This is shown for the case of pure mixing with tau neutrinos only, as the parameter space where HNLs mix purely with electron and muon neutrinos is excluded in the lower mass range (where ΔN_{eff} can be positive) by BBN, the CMB and experimental searches [37, 64, 221]. In these latter two cases, ΔN_{eff} can only be negative in the unconstrained parameter space. This top panel shows that there is a large region of HNL parameter space, where these particles inject most of their energy into neutrinos and still decrease N_{eff}. The behavior of negative ΔN_{eff} continues for short-lived HNLs with masses $m_N > 1\,\text{GeV}$, since the neutrino energy increases with the HNL mass. On the other hand, for HNLs with lifetimes $\tau_N \gg 1\,\text{s}$, it depends on how much energy they inject into the neutrino plasma. Indeed, such HNLs decay long after neutrino decoupling, when non-equilibrium effects are not important anymore and the injected neutrinos remain in the neutrino sector. This means that the sign of ΔN_{eff} is simply determined by the value of ξ_ν. As a result, for masses where $\xi_\nu > 0.5$ (see Fig. 16) this would mean that eventually $\Delta N_{\text{eff}} > 0$ and vice versa (see Fig. 17 for an illustration).

We summarize `pyBBN` predictions for N_{eff} in the form of fitting functions for the three pure HNL mixing cases. This may provide a quick way to predict the impact of HNLs on several cosmological probes through the change in N_{eff}. They read:

[14] We note that `pyBBN` predicts a SM value for N_{eff} of 3.026, rather than 3.044. This is because the code does not include higher-order QED corrections that account for a $\Delta N_{\text{eff}} = 0.01$ increase [183–187], while the remaining is due to numerical inaccuracy. This, however, is only a minor difference and does not change any of the results presented in this work.

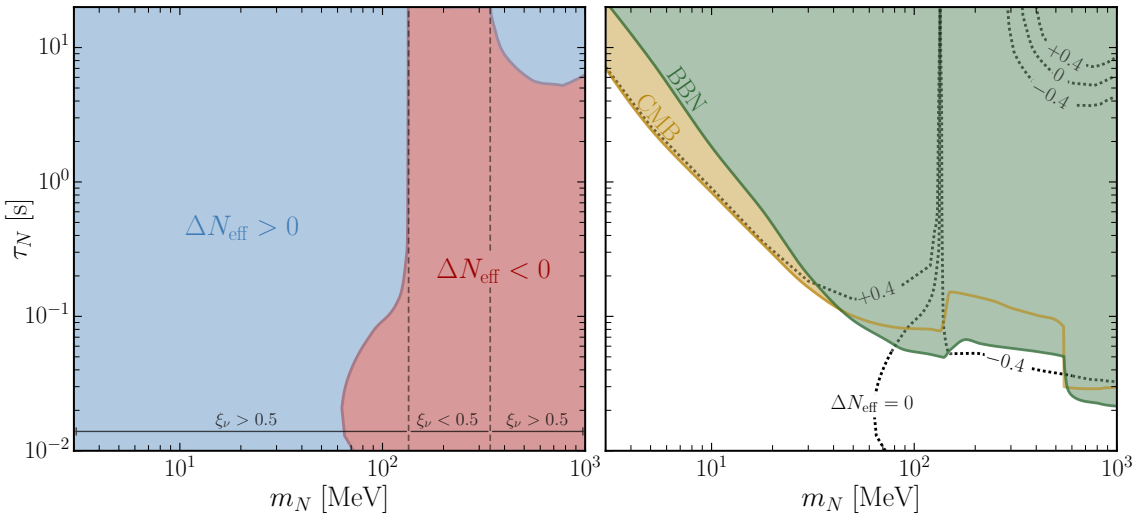

Figure 15: How HNLs change $\Delta N_{\rm eff}$ as a function of their mass and lifetime. Mixing with tau neutrinos only is considered here. *Left panel*: Regions of the HNL parameter space that predict an increase (blue) or decrease (red) of $N_{\rm eff}$ with respect to the SM value. The horizontal lines at the bottom of the plot indicate the mass ranges where HNLs inject most of their energy into neutrinos ($\xi_\nu > 0.5$) or the EM plasma ($\xi_\nu < 0.5$). In the former case, HNLs can still decrease $N_{\rm eff}$ as a result of the efficient transfer of energy from neutrinos to EM particles. *Right panel*: Regions of the HNL parameter space that are excluded by BBN abundance measurements (green) and CMB observations (yellow). The $\Delta N_{\rm eff} = \{0, \pm 0.4\}$ contours give an indication of by how much HNLs can change $N_{\rm eff}$ at the most in the unconstrained region. The BBN bound is from [64] and uses a central value for the primordial helium abundance of $Y_{\rm P} = 0.245$ [230] with an error of 4.35% (see [37] for a discussion on how this error is obtained). For masses higher than the eta-meson mass ($\sim$550 MeV), the meson effect from [37] is included in the analysis. The CMB constraint is obtained using the approach as detailed in [64] (see Sec. 3.3.5 for more details on the CMB bound). This panel also shows that there is only a relatively small unconstrained region of parameter space left that can increase $N_{\rm eff}$ and where HNLs could play a role in alleviating the Hubble tension.

$$\Delta N_{\rm eff}^{\rm Fit}\Big|_{\rm e-mixing} = -\frac{9.78\tau_N e^{5.72\tau_N}}{1+\frac{1.28\cdot 10^5}{m_N^{2.42}}} \tag{3.3.25}$$

$$\Delta N_{\rm eff}^{\rm Fit}\Big|_{\mu-\rm mixing} = -\frac{7.49\tau_N e^{12.1\tau_N}}{1+\frac{2.41\cdot 10^6}{m_N^{2.87}}} \tag{3.3.26}$$

$$\Delta N_{\rm eff}^{\rm Fit}\Big|_{\tau-\rm mixing} = -\frac{8.72\tau_N e^{13.9\tau_N}}{1+\frac{3.49\cdot 10^3}{m_N^{1.51}}}\,, \tag{3.3.27}$$

where m_N is the HNL mass in MeV and τ_N is the HNL lifetime in seconds. The

change in $N_{\rm eff}$ is with respect to the SM value of $N_{\rm eff}^{\rm SM} = 3.026$. The fitting functions are tested for masses $100\,\text{MeV} \leq m_N \leq 1\,\text{GeV}$ and lifetimes $0.02\,\text{s} \leq \tau_N \leq 0.05\,\text{s}$, and have a maximum deviation from the simulated data of roughly $\sim 3\%$.

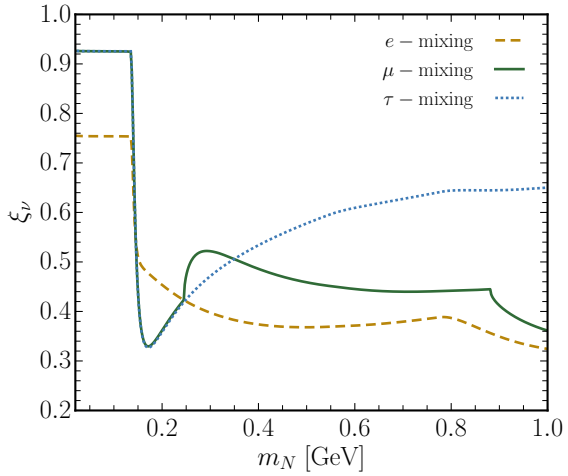

Figure 16: The fraction of HNL mass that is injected into the neutrino plasma. Contributions to this fraction from unstable HNL decay products (mesons and muons) are included and we assume that the kinetic energy of all created charged particles goes into the EM plasma. For $m_N \gtrsim 135\,\text{MeV}$, HNLs can decay into neutral pions, which in their turn decay into two photons. This causes the sudden decrease of ξ_ν around that mass. At higher masses, ξ_ν keeps increasing in the case of τ-mixing, which is due to the absence of HNL decays into charged mesons (such decays are possible in the other two mixing cases).

3.3.5 Bounds from CMB

The CMB anisotropies are mainly sensitive to $N_{\rm eff}$ through its impact on the damping tail [54, 230–232]. For example, a larger number of relativistic degrees of freedom causes a stronger suppression of the power spectrum at high multipoles, as temperature anisotropies below the scale of the photon diffusion length are more damped by the increased expansion rate. This effect is, however, degenerate if the primordial helium abundance $Y_{\rm P}$ is also considered as a free parameter [54]. $Y_{\rm P}$ is related to the number density of free electrons[15], $n_e \propto (1 - Y_{\rm P})$, which in its turn enters in the CMB damping scale. A larger $Y_{\rm P}$ leads to a lower electron density, a larger electron-photon interaction rate, a larger photon diffusion length and thus a stronger damping.

We extend the CMB constraint on HNLs for masses up to $1\,\text{GeV}$ using the same approach as detailed in [64, 233] and show the result in the bottom panel of Fig. 15. Also included in this panel are the contours where $\Delta N_{\rm eff} = \pm 0.4$, which give an indication of by how much HNLs can change $N_{\rm eff}$ at the most, given the current constraints imposed

[15]This relation between n_e and $Y_{\rm P}$ is obtained by imposing charge neutrality on the primordial plasma. Therefore, $Y_{\rm P}$ is allowed to change even if the total baryon density is fixed.

by BBN and the CMB. The CMB bound is only stronger than the BBN bound in the lower mass range, as this is where N_{eff} strongly increases. HNLs with short lifetimes and masses around $\mathcal{O}(10)\,\text{MeV}$ decouple while being non-relativistic and thus have a suppressed number density. They can therefore survive beyond the decoupling of SM weak reactions, without significantly affecting the primordial abundances. However, since the HNL energy density here falls off as $(\text{scale factor})^{-3}$, the HNLs could eventually dominate the total energy density of the Universe. As can be seen in Fig. 16, HNLs in this lower mass range inject most of their energy into neutrinos, which remains in the neutrino sector after neutrino decoupling. The result is then a significant increase in N_{eff}, which can be constrained with the CMB. On the other hand, for masses higher than $\sim$70 MeV, N_{eff} starts decreasing. This decrease is relatively small in magnitude, especially in the region that is not constrained by BBN, where $N_{\text{eff}} - N_{\text{eff}}^{\text{CMB}} \lesssim 0.4$. In addition, the error in the determination of Y_{P} by the CMB is larger than the one by BBN [199, 233]. These two properties make the CMB a weaker probe of HNLs in the higher mass range.

Currently, CMB is a weaker probe of HNLs than BBN in the mass range $m_N \gtrsim 40$ MeV. However, as mentioned before, future CMB experiments could improve upon this result.

3.3.6 Implications for the Hubble Parameter

An increase or decrease of N_{eff} subsequently also changes the Hubble parameter. As such, HNLs could play a role in alleviating the longstanding tension between local determinations of the current day Hubble rate H_0 and the one as inferred from the CMB[16] [235, 236]. The usual approach involves increasing N_{eff}, while keeping the angular scale of the sound horizon $\theta_{\text{s}} = r_{\text{s}}/D_{\text{A}}$ fixed, see e.g. [237–239]. Here, r_{s} is the comoving sound horizon and D_{A} is the comoving angular diameter distance to the surface of last scattering. Both of these quantities depend on the Hubble parameter:

$$r_{\text{s}}(z_*) = \int_{z_*}^{\infty} \frac{c_s(z)\mathrm{d}z}{H(z)} \tag{3.3.28}$$

$$D_{\text{A}}(z_*) = \int_{0}^{z_*} \frac{\mathrm{d}z}{H(z)}\,, \tag{3.3.29}$$

where z_* is the redshift of the last-scattering surface and $c_{\text{s}}(z)$ is the speed of sound of the baryon-photon fluid in the early Universe. The Hubble rate in Eq. (3.3.28) depends mainly on the radiation (photons *and* neutrinos) and matter energy densities, while the one in Eq. (3.3.29) is the late-time Hubble rate and depends mostly on the dark energy and

[16] This question has been considered before in [234]. Importantly, this study used the results of [194], where the assumption was made that any change in the primordial helium abundance is due to ΔN_{eff}. In contrast, here we find that neutrino spectral distortions are the driving power behind ΔN_{eff} for short-lived HNLs. As a consequence, the results presented in our work and in [234] are rather different.

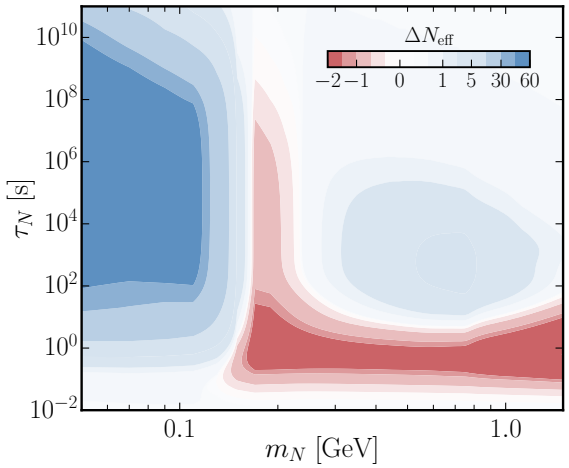

Figure 17: Semi-analytic estimate of $\Delta N_{\rm eff}$ as a function of HNL mass and lifetime in the case of pure tau mixing. This plot is obtained using the method described in Sec. 3.2.1 (and is therefore only accurate up to a factor $3-4$ for short lifetimes, when neutrinos are still in partial equilibrium). Nevertheless, it allows for a qualitative understanding of the behavior of $\Delta N_{\rm eff}$ at lifetimes larger than considered in the main analysis (Fig. 15). Importantly, for lifetimes well beyond the time of neutrino decoupling ($\mathcal{O}(1)$ s), non-equilibrium effects are absent and the sign of $\Delta N_{\rm eff}$ is thus completely determined by the fraction of HNL energy ξ_ν that is injected into the neutrino plasma, see Fig. 16. We see that HNLs with low masses and long lifetimes can still considerably affect $N_{\rm eff}$, while in the higher mass range $\Delta N_{\rm eff}$ tends to 0. This is because low-mass HNLs are more abundantly produced in this region of parameter space [196], where their mixing angles are relatively large.

matter energy densities. This means that increasing $N_{\rm eff}$ only results in a larger early-time Hubble rate and a smaller $r_{\rm s}$. In order to keep $\theta_{\rm s}$ fixed, the comoving angular diameter distance must satisfy $D_{\rm A} = r_{\rm s}/\theta_{\rm s}$, which then also decreases if $r_{\rm s}$ decreases. Looking at Eq. (3.3.29), such a decrease can be accomplished by increasing the dark energy density ω_Λ, or equivalently, H_0 (as $\Omega_\Lambda = 1 - \Omega_{\rm m}$). Since local measurements find a larger value of H_0 than the one inferred from the CMB within the Standard Model, this approach provides a way to reduce the Hubble tension.

This method, however, does not take into account the increased Silk damping induced by a larger $N_{\rm eff}$ [54, 230–232]. Therefore, a price must be paid when alleviating the Hubble tension in this way: An increase of $N_{\rm eff}$ leads to a larger disagreement with the CMB itself. Given our CMB constraint in Fig. 15, we see that HNLs can increase $N_{\rm eff}$ by at most $\Delta N_{\rm eff} \approx 0.4$. This gives us an indication of the extent to which unconstrained HNLs could increase H_0 and ameliorate the Hubble tension. We estimate the corresponding H_0 by running `Monte Python` [240, 241] with the Planck 2018 baseline TTTEEE+lowE analysis. Fixing the primordial helium abundance to[17] $Y_{\rm P} = 0.25$, we obtain[18] $H_0 =$

[17] This is approximately the value of $Y_{\rm P}$ along the $\Delta N_{\rm eff} = +0.4$ curve on the left in the bottom panel of Fig. 15.

[18] All errors in H_0 reported here are at 68% CL.

$70.5 \pm 0.7\,\mathrm{km\,s^{-1}Mpc^{-1}}$. This value can be compared to the one as obtained from, e.g., a distance ladder approach, which gives $H_0^{\mathrm{local}} = 73.0 \pm 1.4\,\mathrm{km\,s^{-1}Mpc^{-1}}$ [242]. Given the Hubble rate obtained within ΛCDM ($H_0 = 67.3 \pm 0.6\,\mathrm{km\,s^{-1}Mpc^{-1}}$ [199]), we see that HNLs which are not excluded by BBN, the CMB and terrestrial experiments can moderately alleviate the Hubble tension.

Appendix

3.A Thermal dependence of mixing angle of HNLs

3.A.1 Neutrino self-energy

Consider hot dense plasma with 4-velocity $u^\mu, u^2 = 1$, temperature $T \gg m_e$ and zero lepton asymmetry. Neutrinos in this plasma may interact elastically with electron-positron pairs:

$$\nu + e^\pm \to \nu + e^\pm \tag{3.A.1}$$

If the 4-momentum of the neutrino does not change in these processes, they contribute to the self-energy of neutrinos Σ, see Fig. 18. In explicit form, the self-energy is

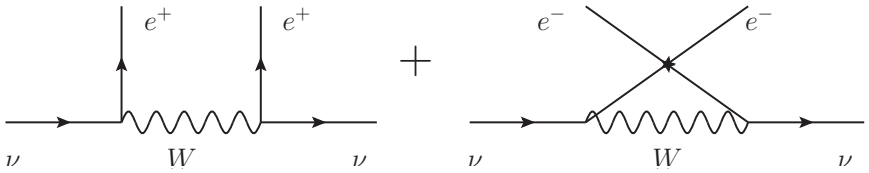

Figure 18: Diagrams of the contribution of the processes (3.A.1) to the self-energy of neutrinos.

$$\Sigma \sim 2 \int \frac{d^3\mathbf{k}}{(2\pi)^3} f_{\text{FD}}(k) \Sigma_k, \tag{3.A.2}$$

where

$$f_{\text{FD}}(k) = \frac{1}{\exp{(k \cdot u)} + 1} \tag{3.A.3}$$

is the distribution function of electrons and positrons (with u being the 4-velocity of the plasma), while Σ_k is (σ denotes the equality up to sign)

$$\Sigma_k \sim \frac{1}{2k} \frac{\sqrt{2} G_F}{2} \Big[\gamma^\mu (1 - \gamma_5) u(k) D_{\mu\nu}(p - k) \bar{u}(k) \gamma^\nu (1 - \gamma_5) - \\ - \gamma^\mu (1 - \gamma_5) v(k) D_{\mu\nu}(p + k) \bar{v}(k) \gamma^\nu (1 - \gamma_5) \Big], \tag{3.A.4}$$

with $D_{\mu\nu}$ being the W boson propagator

$$D_{\mu\nu}(p \pm k) = -\frac{g_{\mu\nu} - \frac{(p\pm k)_\mu (p\pm k)_\nu}{m_W^2}}{(p \pm k)^2 - m_W^2} \tag{3.A.5}$$

The minus sign in (3.A.4) appears because of the Pauli principle – the processes (3.A.1) differ only by exchanged in- and out- charged fermions lines.

Let us simplify estimates:

1. The self-energy (3.A.4) vanishes in the leading order on $G_F E^2$ (i.e., approximating the propagator by $-g_{\mu\nu}/m_W^{-2}$). We need to keep the next order corrections:

$$D_{\mu\nu}(p \pm k) = \frac{g_{\mu\nu}}{m_W^2} - \frac{1}{m_W^4}\left[g_{\mu\nu}(p \pm k)^2 + (p \pm k)_\mu (p \pm k)_\nu\right] + \dots \tag{3.A.6}$$

   The $(p \pm k)_\mu (p \pm k)_\nu / m_W^2$ terms in the numerator either are $\mathcal{O}(m_e/m_W)$ suppressed (this can be shown by acting them on electron-positron spinors $u(k)/v(k)$), or cancel when plugging in Eq. (3.A.4), so only the term $g_{\mu\nu}(p \pm k)^2/m_W^4$ matters. Averaging over incoming electron/positron polarizations, $v\bar{v}, u\bar{u} \to \not{k}/2$, we get

$$\begin{aligned}\Sigma_k &\sim \frac{\sqrt{2}G_F}{8km_W^2}[(p+k)^2 - (p-k)^2]\gamma^\mu(1-\gamma_5)\not{k}\gamma_\mu(1-\gamma_5) = \\ &= \frac{\sqrt{2}G_F(p\cdot k)}{2km_W^2}\gamma^\mu(1-\gamma_5)\not{k}\gamma_\mu(1-\gamma_5) \sim \frac{2\sqrt{2}G_F(p\cdot k)}{km_W^2}\not{k}(1-\gamma_5),\end{aligned} \tag{3.A.7}$$

   where in the last step we used the identity $\gamma^\mu(1-\gamma_5)\not{k}\gamma^\mu(1-\gamma_5) = -4\not{k}(1-\gamma_5)$.

2. Let us integrate Σ_k over the momenta of $e^\pm$. We have

$$\frac{1}{n_e}\int \frac{d^3\mathbf{k}}{(2\pi)^3}(p\cdot k)\frac{\not{k}}{k} f_{\text{FD}}(k) = \frac{p^\alpha \gamma^\beta}{n_e}\int \frac{d^3\mathbf{k}}{(2\pi)^3}\frac{k_\alpha k_\beta}{k} f_{\text{FD}}(k) = p^\alpha\gamma^\beta(Ag_{\alpha\beta} + Bu_\alpha u_\beta), \tag{3.A.8}$$

   where $n_e = 3\zeta(3)T^3/4\pi^2$ is the number density of electrons and positrons, the coefficients A, B can be obtained considering the integral at rest frame of the plasma ($u^\alpha = \delta^\alpha_0$):

$$B = -4A = \frac{4\langle E_e\rangle}{3} \tag{3.A.9}$$

   Therefore, we get

$$\Sigma = \frac{16\sqrt{2}\langle E_e\rangle n_e G_F}{3m_W^2}\left((p\cdot u)\not{u} - \frac{1}{4}\not{p}\right)(1-\gamma_5) \tag{3.A.10}$$

Restoring the overall sign and making a completely similar calculation that involves the

contribution of neutrinos themselves to the self-energy, we get

$$\boxed{\Sigma(p) = -\frac{16\sqrt{2}\langle E_e\rangle n_e G_F}{3m_W^2}\left[1+\frac{m_W^2}{2m_Z^2}\right]\left((p\cdot u)\not{u} - \frac{1}{4}\not{p}\right)(1-\gamma_5)} \tag{3.A.11}$$

This expression agrees with [243] (Eq. (12)) at the rest frame of the plasma, $u^\mu = (1,0,0,0)$ (see also Eq. (20) in [213]).[19]

At rest frame of the plasma, neglecting the neutrino mass, using the relation $\langle E_e\rangle = 7\pi^4 T/180\zeta(3)$, for the correction to the neutrino energy we have

$$\Delta E_\nu(p) = \frac{1}{2p}\text{Tr}[\Sigma u(p)\bar{u}(p)] = -\frac{14\sqrt{2}\pi^2 G_F T^4 p}{45 m_W^2}\left(1+\frac{m_W^2}{2m_Z^2}\right) \tag{3.A.12}$$

This expression agrees with [213, 243].

The self-energy modifies the neutrino propagator: after the resummation we get

$$D_\nu(p) = \frac{1}{\not{p}}\sum_{n=0}^{\infty}\left(-\Sigma\frac{1}{\not{p}}\right)^n = \frac{1}{\not{p} + A\not{c}(1-\gamma_5)} \tag{3.A.13}$$

3.A.2 Derivation of $U_{\mathbf{m}}^2$

Now, consider the general matrix element $\mathcal{M}$ of the interaction of an HNL N with SM particles. It couples to the neutrino via the term $\mathcal{L}_{\text{mixing}} = m_N\theta\bar{N}\nu$, where m_N is the mass of the HNL and $\theta \ll 1$ is the mixing angle. Therefore, $\mathcal{M}$ takes the form

$$\mathcal{M} = \theta m_N \bar{N}(p) D_\nu(p)\gamma^\mu(1-\gamma_5)\ldots, \tag{3.A.14}$$

where $\ldots$ is the interaction dependent part and $\gamma^\mu(1-\gamma_5)$ comes from the neutrino vertex. Let us use the series representation of the neutrino propagator (3.A.13). With the help of the identity

$$\left[\not{c}(1-\gamma_5)\frac{\not{p}}{p^2}\right]^n\gamma^\mu(1-\gamma_5) = \left[2\not{c}\frac{\not{p}}{p^2}\right]^n\gamma^\mu(1-\gamma_5) \tag{3.A.15}$$

we get

$$\mathcal{M} = \theta m_N\bar{N}(p)\frac{1}{\not{p}+2A\not{c}}\gamma^\mu(1-\gamma_5)\cdots = \theta m_N\bar{N}(p)\frac{\not{p}+2A\not{c}}{p^2+4A(p\cdot c)+4A^2c^2}\gamma^\mu(1-\gamma_5)\ldots, \tag{3.A.16}$$

where in the last equality we multiplied the numerator and denominator of the propagator by $\not{p}+2A\not{c}$. Finally, using the dispersion relation for HNLs ($p^2 = m_N^2$), the Dirac equation $\bar{N}\not{p} = m_N\bar{N}$ and neglecting the $4A^2$ term in comparison to $4A$ (valid everywhere for

[19] The expression (3.A.11) is larger from Eq. (12) in [243] by a factor of two, but this is most likely due to a misprint, as Eqs. (21) from [213] and (13) from [243] require twice larger value.

$T \ll m_W$), we obtain

$$\boxed{\mathcal{M} = \theta_M \bar{N}(p) \left[1 + 2A \left(\frac{(p\cdot u)}{m_N}\not{u} - \frac{1}{4}\right)\right] \gamma^\mu (1-\gamma_5) \dots}, \tag{3.A.17}$$

where we introduced the effective mixing angle θ_M:

$$\theta_M = \frac{\theta}{1 + 4A\left(\frac{(p\cdot u)^2}{m_N^2} - \frac{1}{4}\right)} \approx \frac{\theta}{1 + 2.2\cdot 10^{-8}\left(\frac{T}{1\,\text{GeV}}\right)^4 \left(\gamma_N^2 - \frac{1}{4}\right)}, \tag{3.A.18}$$

with $\gamma_N = E_N/m_N$. *For practical purposes, the second term in the numerator,*

$$\hat{X} = 1 + 2A\left(\frac{(p\cdot u)}{m_N}\not{u} - \frac{1}{4}\right), \tag{3.A.19}$$

may be neglected, see a discussion below.

In the UR limit $\gamma_N \approx p_N/m_N \approx 3.15T/m_N \gg 1$, and we get

$$\boxed{\theta_M(T) \approx \frac{\theta}{1 + a\left(\frac{T}{1\,\text{GeV}}\right)^6 \left(\frac{1\,\text{GeV}}{m_N}\right)^2}}, \quad a = 2.2\cdot 10^{-7} \tag{3.A.20}$$

The value of a fully coincides with that from the literature (see, e.g., [196]).

Role of the numerator. Let us now consider the operator in the square brackets from Eq. (3.A.17):

$$\hat{X} = 1 + 2A\left(\frac{(p\cdot u)}{m_N}\not{u} - \frac{1}{4}\right) \tag{3.A.21}$$

Using $(p\cdot u)\not{u} \simeq E_N$ in the plasma rest frame and approximating $E_N \approx 3.15T$, we find that the second term of the operator cannot be neglected ($\gtrsim 1$) for temperatures

$$T \gtrsim 31\left(\frac{m_N}{1\,\text{GeV}}\right)^{1/5} \tag{3.A.22}$$

In contrast, the second term in the denominator of (3.A.20) becomes non-negligible at

$$T \gtrsim 12.9\left(\frac{m_N}{1\,\text{GeV}}\right)^{1/3} \tag{3.A.23}$$

So there is a temperature domain in which the θ_M is affected by the plasma effects, whereas the numerator can be neglected. However, the numerator prevents the matrix element from huge suppression at large temperatures: the asymptotics of the suppression of the product $\hat{X}\cdot\theta_M$ is m_N/E_N.

3.B Changes in $p \leftrightarrow n$ rates due to the presence of mesons

In this section, we provide details on our estimate of the effect of mesons on BBN.

Pions. The threshold-less processes with charged pions are

$$\pi^- + p \to n + \pi^0/\gamma, \quad \pi^+ + n \to p + \pi^0. \tag{3.B.1}$$

The cross-sections at threshold are [179]

$$\langle \sigma^{\pi^-}_{p\to n} v\rangle \approx 4.3 \cdot 10^{-23} F_c^\pi(T)\ \mathrm{m}^3/s, \quad \frac{\langle \sigma^{\pi^-}_{p\to n} v\rangle}{\langle \sigma^{\pi^+}_{n\to p} v\rangle} \approx 0.9\ F_c^\pi(T), \tag{3.B.2}$$

where F_c^h is the Sommerfeld enhancement of the cross-section due to presence of two oppositely charged particles in the in-state:

$$F_c^h = \frac{x}{1+e^{-x}}, \quad \text{where} \quad x = \frac{2\pi\alpha_{\text{EM}}}{v_e}, \tag{3.B.3}$$

where $v_e \approx \sqrt{\frac{T}{m_h}} + \sqrt{\frac{T}{m_p}}$ is the relative velocity between a nucleon and a meson. F_c is of order of one at $T \simeq 1$ MeV.

Kaons. The threshold-less $n \leftrightarrow p$ conversions driven by kaons are

$$\begin{aligned}
K^- + p &\to \Sigma^{\pm/0}/\Lambda + \pi^{\mp/0}/\pi^0 \to n + 2\pi,\\
K^- + n &\to \Sigma^{-/0}/\Lambda + \pi^{0/-}/\pi^- \to n + 2\pi,\\
\bar{K}^0_L + p &\to \Sigma^{0/+}/\Lambda + \pi^{+/0}/\pi^+ \to n + 2\pi,\\
\bar{K}^0_L + n &\to \Sigma^{\pm/0}/\Lambda + \pi^{\mp/0}/\pi^0 \to p + 2\pi,
\end{aligned} \tag{3.B.4}$$

where Λ, Σ are the lightest strange hadronic resonances [178].

Their effect is similar to the one of pions, but with small differences: *(i)* cross-sections of above reactions are higher than the cross-sections of (3.1.2)[20], *(ii)* there is no isotopic symmetry - K^+ mesons do not contribute to $p \leftrightarrow n$ conversion, since there are no threshold-less processes $n + K^+ \to p + X$. Indeed, the process $n + K^+ \to p + K^0$ has the threshold $Q \approx 2.8$ MeV, while the threshold-less processes going through s-quark resonances, similar to (3.B.4), would require resonances with negative strangeness and positive baryon number, that do not exist, *(iii)* neutral kaons do not lose the energy before decaying (however, we follow [178] and approximate the cross-sections by threshold values).

The threshold cross-sections are

$$\langle \sigma^{K^-}_{p\to n} v\rangle \approx 9.6 \cdot 10^{-22} F_c^K(T)\ \mathrm{m}^3/s, \quad \frac{\langle \sigma^{K^-}_{p\to n} v\rangle}{\langle \sigma^{K^-}_{n\to p} v\rangle} \approx 2.46\ F_c^K(T), \tag{3.B.5}$$

[20]The reason is that these reactions have higher available phase space and go through hadronic resonances.

$$\langle\sigma^{K^0}_{p\to n}v\rangle \approx 1.95\cdot 10^{-22}\ \mathrm{m}^3/\mathrm{s}, \quad \frac{\langle\sigma^{K^0_L}_{p\to n}v\rangle}{\langle\sigma^{K^0_L}_{n\to p}v\rangle} \approx 0.41. \tag{3.B.6}$$

Conversion probabilities. A probability for a meson h to convert $p \leftrightarrow n$ before decaying is given by

$$P^h_{\text{conv}} \approx \frac{\langle\sigma^h_{p\leftrightarrow n}v\rangle n_B}{\Gamma^h_{\text{decay}}}, \tag{3.B.7}$$

where Γ^h_{decay} is the decay width and n_B is the baryon number density. The decay widths of mesons are [52]

$$\Gamma^{\pi^\pm}_{\text{decay}} \approx 3.8\cdot 10^7\ \mathrm{s}^{-1}, \quad \Gamma^{K^-}_{\text{decay}} \approx 8.3\cdot 10^7\ \mathrm{s}^{-1}, \quad \Gamma^{K^0_L}_{\text{decay}} \approx 2\cdot 10^7\ \mathrm{s}^{-1} \tag{3.B.8}$$

Using (3.B.2), (3.B.5), (3.B.8), for the $p \to n$ conversion probabilities we obtain

$$\begin{aligned} &P^{\pi^-}_{\text{conv}}(T) \approx 2.5\cdot 10^{-2}\left(\frac{T}{1\ \text{MeV}}\right)^3, \quad P^{K^-}_{\text{conv}}(T) \approx 2.8\cdot 10^{-1}\left(\frac{T}{1\ \text{MeV}}\right)^3, \\ &P^{K^0_L}_{\text{conv}}(T) \approx 1.6\cdot 10^{-1}\left(\frac{T}{1\ \text{MeV}}\right)^3 \end{aligned} \tag{3.B.9}$$

The largeness of the probabilities is caused by the fact that the decay of mesons proceeds through weak interactions, while the $p \leftrightarrow n$ conversion is mediated by strong interactions. In particular, at $T \gtrsim 2$ MeV kaons participate in the conversion faster than they decay.

3.B.1 Numeric study

To verify the analytic estimate (3.3.24), we numerically solve equation for the neutron abundance X_n, where we include both weak conversion $p \leftrightarrow n$ processes and the meson driven processes (3.B.1)-(3.B.4). The system of equations has the form

$$\begin{cases} \frac{X_n}{dt} = \left(\frac{dX_n}{dt}\right)_{\text{SM}} + \left(\frac{dX_n}{dt}\right)_{\pi} + \left(\frac{dX_n}{dt}\right)_{K^-} + \left(\frac{dX_n}{dt}\right)_{K^0_L}, \\ \frac{dn_{\pi^-}}{dt} = n_N\frac{\text{Br}_{N\to\pi^-}}{\tau_N} - \Gamma^{\pi^-}_{\text{decay}}n_{\pi^-} - \langle\sigma^{\pi^-}_{p\to n}v\rangle(1-X_n)n_Bn_{\pi^-}, \\ \frac{dn_{\pi^+}}{dt} = n_N\frac{\text{Br}_{N\to\pi^+}}{\tau_N} - \Gamma^{\pi^+}_{\text{decay}}n_{\pi^+} - \langle\sigma^{\pi^+}_{n\to p}v\rangle X_nn_Bn_{\pi^+}, \\ \frac{dn_{K^-}}{dt} = n_N\frac{\text{Br}_{N\to K^-}}{\tau_N} - \Gamma^{K^-}_{\text{decay}}n_{K^-} - \langle\sigma^{K^-}_{p\to n}v\rangle(1-X_n)n_Bn_{K^-} - \langle\sigma^{K^-}_{n\to p}v\rangle X_nn_Bn_{K^-}, \\ \frac{dn_{K^0_L}}{dt} = n_N\frac{\text{Br}_{N\to K^0_L}}{\tau_N} - \Gamma^{K^0_L}_{\text{decay}}n_{K^0_L} - \langle\sigma^{K^0_L}_{p\to n}v\rangle(1-X_n)n_Bn_{K^0_L} - \langle\sigma^{K^0_L}_{n\to p}v\rangle X_nn_Bn_{K^0_L} \end{cases} \tag{3.B.10}$$

Here the quantities

$$\left(\frac{dX_n}{dt}\right)_\pi = (1-X_n)n_{\pi^-}\langle\sigma^{\pi^-}_{p\to n}v\rangle - X_n n_{\pi^+}\langle\sigma^{\pi^+}_{n\to p}v\rangle,$$

and (3.B.11)

$$\left(\frac{dX_n}{dt}\right)_K = (1-X_n)n_K\langle\sigma^{K}_{p\to n}v\rangle - X_n n_K\langle\sigma^{K}_{n\to p}v\rangle$$

are the rates of change of X_n due to different mesons ($K = K^-/K^0_L$); n_B is the baryon number density $n_B = \eta_B n_\gamma$. In equations for the number density of mesons n_h, the first term comes from HNLs, the second due to decays of mesons and the last term is due to $p \leftrightarrow n$ conversion. The time-temperature relation and the scale factor dynamics are provided by the solution of Eq. (3.3.19), and the HNL number density may be obtained using Eq. (3.3.17).

During times $t_{\rm eq} \simeq (\Gamma^h_{\rm decay})^{-1} \sim 10^{-8}$ s, which are small in comparison to any other time scale in the system, the solution for n_h reaches the dynamical equilibrium:

$$n_{\pi^-} = \frac{n_N \cdot \mathrm{Br}_{N\to\pi^-}}{\tau_N(\Gamma^{\pi^-}_{\rm decay} + \langle\sigma^{\pi^-}_{p\to n}v\rangle(1-X_n)n_B)}, \quad n_{\pi^+} = \frac{n_N \cdot \mathrm{Br}_{N\to\pi^+}}{\tau_N(\Gamma^{\pi^+}_{\rm decay} + \langle\sigma^{\pi^+}_{n\to p}v\rangle(1-X_n)n_B)}, \tag{3.B.12}$$

$$n_K = \frac{n_N \cdot \mathrm{Br}_{N\to K}}{\tau_N(\Gamma^{K}_{\rm decay} + \langle\sigma^{K}_{p\to n}v\rangle(1-X_n)n_B + \langle\sigma^{K}_{n\to p}v\rangle X_n n_B)}, \tag{3.B.13}$$

where $K = K^-/K^0_L$.

Therefore, we solve a single equation

$$\frac{X_n}{dt} = \left(\frac{dX_n}{dt}\right)_{\rm SM} + \left(\frac{dX_n}{dt}\right)_\pi + \left(\frac{dX_n}{dt}\right)_{K^-} + \left(\frac{dX_n}{dt}\right)_{K^0_L}. \tag{3.B.14}$$

where we use meson number densities given by Eqs. (3.B.12) and (3.B.13) in the meson-driven conversion rates (3.B.11). The results are shown in Fig. 19. Our main result is the right panel of Fig. 19 – it shows that the value $T_0^{\rm min} \simeq 1.50\,\mathrm{MeV}$ and that its variation as a function of the HNL mass is within $\pm 1\%$.

With the help of Eqs. (3.B.2), (3.B.5), we obtain the value of the neutron abundance driven solely by a given meson h. As long as $T \gtrsim T_0$ (see Eq. (3.1.9) and left panel of Fig. 19), the weak interaction processes may be completely neglected, and the resulting X_n are given by

$$X_n^{\pi^\pm} = \frac{\langle\sigma^{\pi^-}_{p\to n}v\rangle \cdot n_{\pi^-}}{\langle\sigma^{\pi^-}_{p\to n}v\rangle \cdot n_{\pi^-} + \langle\sigma^{\pi^+}_{n\to p}v\rangle \cdot n_{\pi^+}} \approx \frac{0.9F_c^\pi(T)}{1+0.9F_c^\pi(T)},$$

$$X_n^{K^-} \approx \frac{2.46F_c^K}{2.46F_c^K+1}, \quad X_n^{K^0_L} \approx 0.32 \tag{3.B.15}$$

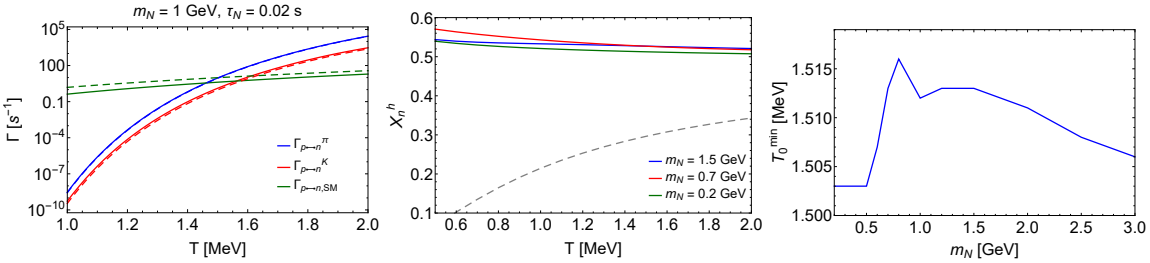

Figure 19: *Left panel*: the behavior of the $p \to n$ (solid lines) and $n \to p$ (dashed lines) conversion rates in the case of pion and kaon driven conversions and SBBN. We consider HNLs mixing with e flavor, mass $m_N = 1$ GeV and lifetime $\tau_N = 0.02$ s as an example. *Middle panel*: the temperature dependence of the neutron abundance X_n assuming that its evolution is completely dominated by the meson driven $p \leftrightarrow n$ conversions. We consider HNLs mixing with e flavor and different masses: $m_N = 200$ MeV (only pions are present), $m_N = 700$ MeV (pions and charged kaons are present), $m_N = 1.5$ GeV (pions, charged and neutral kaons are present). The dashed gray line denotes the value of the neutron abundance at equilibrium in SBBN. *Right panel*: the HNL mass dependence of the temperature $T_0^{\min}$.

The values of $X_n^{\pi^-/K^-}$ grow with the decrease of the temperature due to the growth of the Coulomb factor F_c, which enhances the rate of the $p \to n$ process.

The quantities (3.B.15) provide us the qualitative estimate of the value of X_n in presence of different mesons, Fig. 19. Below the kaon production threshold, $X_n^h = X_n^{\pi^\pm}$. At larger masses, in order to find X_n^h we need to set the whole right hand-side of Eq. (3.B.10) to zero. Below the K_L^0 production threshold (which occurs at $m_N = m_\phi$), the value of X_n^h grows, since charged kaons tend X_n to higher values than $X_n^{\pi^-}$. Above the neutral kaon production threshold, the ratio $\mathrm{Br}_{N\to K^-}/\mathrm{Br}_{N\to\pi^-}$ increases (Fig. 10) and X_n^h grows further. However, kaons K_L^0, that are present in small amounts, somewhat diminish this growth.

The value of $X_n^h(m_N)$ provide us the mass dependence of $T_0^{\min}(m_N)$, which is the smallest temperature allowed by observations (c.f. Fig. 2). We show it in Fig. 19 (right panel).

Let us now comment on the approximations of this approach. If HNLs disappear from the plasma before neutrinos froze out, the evolution of the neutron abundance and subsequent nuclear reactions proceed exactly as in SBBN case (albeit with modified initial value of X_n at $T = T_0^{\min}$).

Indeed, the onset of nuclear reactions is determined by the dynamical balance between reactions of deuterium synthesis and dissociation. This balance depends on the value of η_B. The latter gets diluted by the factor ζ due to decays of HNLs, see Section 3.3.1. However, we fix η_B at the beginning of nuclear reactions to be the same as measured by CMB. This of course means that η_B has been ζ^{-1} times higher before decays of HNLs, but no observables can probe the value of η_B in this epoch.

Another ingredient that affects dynamics of nuclear reactions is the time-temperature relation, traditionally encoded in the value of N_{eff}. If HNLs have $\tau_N \simeq 0.02$ s, neutrinos

are in equilibrium during the decay of the most of HNLs, and therefore they do not change neither N_{eff} nor weak $p \leftrightarrow n$ conversion, see detailed analysis in [170]. As a result, the evolution of primordial plasma below T_0^{min} is governed by the SBBN equations, and our prediction of the upper bound on the allowed HNL lifetimes is conservative. HNLs with larger lifetimes do change N_{eff} and rates. However, the net effect of this impact is an increase of the ^{4}He abundance [64, 170], and therefore the predictions of our approach in the increase of the ^{4}He abundance, which does not include changes in these quantities, are conservative.

3.B.2 Numeric approach for long-lived HNLs

The total system of equations for HNLs, mesons, SM plasma and nuclei reads

$$\begin{cases} n_N = \left(\frac{a_{N,\text{dec}}}{a}\right)^3 n_{N,\text{dec}} \cdot e^{-t/\tau_N}, \\ \dot{a}(t) = a(t) \cdot H(t), \\ \frac{dT_{\text{EM}}}{dt} + HT_{\text{EM}} = \Gamma_{\text{EM}\leftrightarrow\nu} \frac{\rho_{\text{EM}}}{d\rho_{\text{EM}}/dT_{\text{EM}}} + \frac{\rho_N \epsilon_{\text{EM}}}{\tau_N}, \\ \frac{dT_\nu}{dt} + HT_\nu = -\Gamma_{\text{EM}\leftrightarrow\nu} \frac{\rho_\nu}{d\rho_\nu/dT_\nu} + \frac{\rho_N \epsilon_\nu}{\tau_N}, \\ \dot{X}_i = \sum_{j,k} N_i \left(\Gamma_{j\to ki} \prod_j \frac{Y_j^{N_j}}{N_j!} - \Gamma_{ki\to j} \prod_k \frac{Y_k^{N_k}}{N_k!} \right) \end{cases} \tag{3.B.16}$$

Here, $X_i \equiv n_i/n_B$, N_i denotes the stoichiometric coefficient, $j \to kl$ is the shortland notation for

$$j_1 + \cdots + j_p \to i + k_1 + \cdots + k_q, \tag{3.B.17}$$

and $\prod_k \frac{Y_k^{N_k}}{N_k!}$ is the shortland notation for

$$\prod_k \frac{Y_k^{N_k}}{N_k!} \equiv \frac{Y_{k_1}^{N_{k_1}} \ldots Y_{k_q}^{N_{k_q}}}{N_{k_1}! \ldots N_{k_q}!} \tag{3.B.18}$$

$\Gamma_{j\to ik}$ are the reaction rates of SBBN reactions governed the evolution of $d, t,^3$He,^{4}He,^{7}Be,^{7}Li, as well as weak $p \leftrightarrow n$ rates from from [48], and meson-driven dissociation rates, which we use from [182].[21] The number density of mesons evolve due to Eqs. (3.B.10), where in addition to $p \leftrightarrow n$ rates there are now also nuclear dissociation rates. Our results for nuclear abundances in SBBN are in perfect agreement with predictions from [48].

We neglect the change of weak SM rates, since in presence of long-lived HNLs with $\tau_N \gg 1$ s they do not change at temperatures $T \simeq \mathcal{O}(1\,\text{MeV})$ at which weak interaction processes are important.

[21] We of course do not include the inverse reactions in which mesons occur, since these reactions are endotermic and practically impossible.

3.C Temperature Evolution Equations

Here we provide the relevant equations for the time evolution of the neutrino and photon temperatures in the presence of decaying FIPs. Assuming a Fermi-Dirac distribution for neutrinos, the equations read [184, 198]:

$$\frac{dT_\nu}{dt} + 4HT_\nu = \frac{(1-\xi_{\text{EM,eff}})\frac{\rho_{\text{FIP}}}{\tau_{\text{FIP}}} + \Gamma_{\nu\leftrightarrow\text{EM}}(T_\nu, T_{\text{EM}})}{d\rho_\nu/dT_\nu} \tag{3.C.1}$$

$$\frac{dT_{\text{EM}}}{dt} + \frac{(4H\rho_\gamma + 3H(\rho_e + p_e))}{d\rho_e/dT + d\rho_\gamma/dT} = \frac{\xi_{\text{EM,eff}}\frac{\rho_{\text{FIP}}}{\tau_{\text{FIP}}} - \Gamma_{\nu\leftrightarrow\text{EM}}(T_\nu, T_{\text{EM}})}{d\rho_e/dT + d\rho_\gamma/dT} \tag{3.C.2}$$

$$\frac{d\rho_{\text{FIP}}}{dt} + 3H\rho_{\text{FIP}} = -\frac{\rho_{\text{FIP}}}{\tau_{\text{FIP}}}\,, \tag{3.C.3}$$

where $\xi_{\text{EM,eff}}$ is given in Eq. (3.2.6), ρ_i is the energy density of particle i, τ_{FIP} is the FIP lifetime and $\Gamma_{\nu\leftrightarrow\text{EM}}(T_\nu, T_{\text{EM}}) = \left(\Gamma_{\nu_e\leftrightarrow\text{EM}} + 2\Gamma_{\nu_\mu\leftrightarrow\text{EM}}\right)/3$ is the energy density exchange rate averaged over neutrino flavours, given by Eqs. (2.12a) and (2.12b) in [198].

3.D Comment on "Massive sterile neutrinos in the early universe: From thermal decoupling to cosmological constraints" by Mastrototaro et al.

After our work was submitted, the paper [197] appeared that studies the impact of HNLs with masses $m_N < m_\pi$ on N_{eff}. The authors of this work used numerical simulations in order to obtain N_{eff} and disagree with our conclusion that N_{eff} can decrease even if most of the HNL energy is injected into neutrinos. They have presented an analytic argument in their Appendix C which aims to demonstrate that our conclusion on N_{eff} is wrong. They start with a toy model in Eq. (C.1) that describes the evolution of the distribution function of neutrinos f_ν:

$$x\partial_x f_\nu(E_\nu, x) = \frac{1}{H}\left[S(x, E_\nu) + \varsigma^2 G_F^2 T^4 E_\nu (f_{\text{eq}} - f_\nu)\right]\,, \tag{3.D.1}$$

where $x = ma$ (with a the scale factor and $m = 1\,\text{MeV}$), H is the Hubble rate, ς is a constant and $S(x) > 0$ is the source term from decays of HNLs. The second term in the brackets describes the interactions between neutrinos and EM particles, where f_{eq} is the equilibrium distribution function resulting from the interaction dynamics of neutrinos and EM particles in the presence of HNLs.

Their argument as to why $N_{\rm eff}$ cannot decrease goes as follows: as far as the source injecting rate $S(x, E_\nu)$ and the collision rate $G_F^2 T^4 E_\nu$ are much higher than the Hubble rate, the solution of Eq. (3.D.1) may be given in terms of the quasi-static solution:

$$f_\nu \approx f_{\rm eq} + \frac{S}{G_F^2 T^4 E_\nu} \,. \tag{3.D.2}$$

In the limiting case $S \ll G_F^2 T^4 E_\nu$, the solution is just $f_\nu = f_{\rm eq}$, while in the opposite case $f_\nu \gg f_{\rm eq}$. The authors conclude that in any case $f_\nu \geq f_{\rm eq}$ and thus $\Delta N_{\rm eff} \geq 0$. However, while this argument may be applicable at very early times when neutrinos are in perfect equilibrium, it is no longer valid at temperatures $T = \mathcal{O}(1\ \text{MeV})$, when they start to decouple. During the decoupling process, the dynamics of the equilibration between neutrinos and EM particles, i.e., the energy transfer between the two sectors, becomes very important and is not captured by Eq. (3.D.1).

We reiterate our argument as to why $N_{\rm eff}$ can decrease when FIPs inject most of their energy into neutrinos, but now from the point of view of the neutrino distribution function (see also the right panel of Fig. 4 and the surrounding text for a similar discussion). Before the decay of the FIP, the neutrino distribution function is the same as the equilibrium distribution, $f_\nu = f_{\rm eq}$. Right after the decay of the FIP, the neutrino distribution at high energies becomes $f_\nu > f_{\rm eq}$, while at low energies it is still $f_\nu = f_{\rm eq}$. During the thermalisation, high-energy neutrinos interact with both low-energy neutrinos and EM particles. In this process, the temperature of the equilibrium distribution function $f_{\rm eq}$ increases. Now, neutrinos in the high-energy tail of f_ν interact efficiently, see Eq. (3.2.5), and $f_\nu \longrightarrow f_{\rm eq}$ for such neutrinos. But at low energies, neutrinos do not interact efficiently anymore to catch up with the increase of $f_{\rm eq}$, which eventually leads to $f_\nu < f_{\rm eq}$ in this energy range. Given that these low-energy neutrinos contribute the most to $N_{\rm eff}$, it means that $\Delta N_{\rm eff}$ can become negative.