Chemical and kinetic equilibrium in cosmology: Facts and myths

Stefano Profumo

Department of Physics, University of California, Santa Cruz, California 95064, USA and Santa Cruz Institute for Particle Physics (SCIPP), Santa Cruz, California 95064, USA

(Received 27 August 2025; accepted 23 September 2025; published 6 November 2025)

We clarify that chemical and kinetic equilibration in the early Universe are distinct: neither implies the other, and the ordering of their decouplings need not be universal. We illustrate this with Standard Model neutrino decoupling, strong-washout leptogenesis, dark–matter scenarios where kinetic decoupling precedes chemical freeze-out (resonant/forbidden, conversion/coannihilation, coscattering), and dark sectors at with temperatures distinct from the visible-sector temperature, with semiannihilation or $3 \to 2$ cannibal dynamics. The moral of the story is simple: Chemical equilibrium governs numbers, kinetic equilibrium governs shapes. In an expanding Universe the operators that control them rarely fade at the same time, and when they do not, the order of decoupling is model dependent. Turning to phase-space evolution whenever momentum selectivity matters is the surest way to obtain robust cosmological predictions.

DOI: 10.1103/lfvd-bkk4

I. INTRODUCTION

In the evolving Universe, the interplay between chemical and kinetic equilibrium governs the abundance and distribution of matter across cosmic time. Understanding these equilibrium concepts is fundamental to cosmology, as they determine everything from primordial nucleosynthesis in the early Universe to star formation in molecular clouds and the ionization structure of the intergalactic medium.

Chemical equilibrium is achieved when the rates of formation and destruction of a given species balance, resulting in a time-independent number density. For a reaction network involving species i, this condition can be expressed as $\frac{dn_i}{dt} = \sum_{\text{formation}} R_{f,i} - \sum_{\text{destruction}} R_{d,i} = 0$, where $R_{f,i}$ and $R_{d,i}$ represent the formation and destruction rates respectively [1,2]. The establishment of chemical equilibrium depends critically on the reaction timescales compared to the relevant dynamical timescales of the system. In cosmological contexts, this often involves comparing reaction rates to the Hubble expansion rate $H(t)$. When reaction timescales are much shorter than the expansion timescale, $\tau_{\text{reaction}} \ll H^{-1}$, chemical equilibrium is maintained [3]. For a generic reaction $A + B \leftrightarrow C + D$, chemical equilibrium implies the law of mass action,

$$\frac{n_C n_D}{n_A n_B} = K_{\text{eq}}(T), \tag{1}$$

where n_i are the number densities of the species and $K_{\text{eq}}(T)$ is the temperature-dependent equilibrium constant. The latter is determined by the ratio of the partition functions of the reactants and products and, in the dilute limit, depends exponentially on the relevant binding energies [4].

Kinetic equilibrium (also termed thermal equilibrium) refers to the establishment of a thermal velocity distribution for particles of a given species. This occurs when collision processes are sufficiently frequent to maintain the characteristic Maxwell-Boltzmann, Fermi-Dirac, or Bose-Einstein distributions appropriate for the particle statistics [5,6]. The condition for kinetic equilibrium requires that the collision timescale be much shorter than other relevant timescales in the system: $\tau_{\text{collision}} = \frac{1}{n\langle\sigma v\rangle} \ll \tau_{\text{system}}$, where n is the particle density, $\langle\sigma v\rangle$ is the thermally averaged collision cross section, and τ_{system} represents the characteristic evolution timescale (e.g., H^{-1} in cosmology) [2]. In kinetic equilibrium, particles maintain a thermal distribution characterized by a single temperature T, with all translational degrees of freedom sharing this temperature. This is a stronger, but distinct, condition than chemical equilibrium and typically requires higher densities and interaction rates.

When kinetic equilibrium is established, the number density of a species can be calculated using statistical mechanics principles. The general expression depends on the quantum statistics obeyed by the particles. For nonrelativistic particles in the classical limit, where the thermal

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

de Broglie wavelength $\lambda_T = h/\sqrt{2\pi mkT}$ satisfies $n\lambda_T^3 \ll 1$, the equilibrium number density is

$$n = g\left(\frac{mkT}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu - mc^2}{kT}\right), \tag{2}$$

where g is the statistical degeneracy factor, which counts the number of internal degrees of freedom of the species (such as spin or polarization states). For clarity, we stress here that g always denotes the internal (statistical) degeneracy factor throughout this work., m is the particle rest mass, μ is the chemical potential, and other symbols have their usual meanings [7].

For fermions, the number density follows Fermi-Dirac statistics

$$n = g\int \frac{d^3p}{(2\pi\hbar)^3}\frac{1}{\exp[\frac{E(p)-\mu}{kT}]+1}, \tag{3}$$

while for bosons the distribution follows Bose-Einstein statistics

$$n = g\int \frac{d^3p}{(2\pi\hbar)^3}\frac{1}{\exp[\frac{E(p)-\mu}{kT}]-1}, \tag{4}$$

where $E(p) = \sqrt{p^2c^2 + m^2c^4}$ is the relativistic energy-momentum relation [3,6].

Several important asymptotic limits are relevant in cosmological applications. In the ultrarelativistic limit ($kT \gg mc^2$), when $\mu \ll kT$, both fermions and bosons exhibit simple power-law scaling with temperature

$$n = \begin{cases} \frac{3\zeta(3)}{4\pi^2} g\left(\frac{kT}{\hbar c}\right)^3 & \text{(fermions)} \\ \frac{\zeta(3)}{\pi^2} g\left(\frac{kT}{\hbar c}\right)^3 & \text{(bosons)} \end{cases} \tag{5}$$

where $\zeta(3) \approx 1.202$ is the Riemann zeta function [2]. In the nonrelativistic limit ($kT \ll mc^2$), the quantum distributions reduce to the classical form given by Eq. (2) when quantum effects are negligible. Finally, when $\mu \gg kT$, fermions become degenerate with number density $n = \frac{g}{6\pi^2}(\frac{\mu}{\hbar c})^3$ for $\mu \gg mc^2$.

These equilibrium expressions form the foundation for calculating particle abundances in various cosmological epochs, from the radiation-dominated early universe through structure formation to the present day [1,2]. However, the relationship between chemical and kinetic equilibrium in cosmological contexts is more nuanced than often assumed in the literature. This paper addresses a few key questions that challenge common misconceptions about the connection between these equilibrium concepts, broadly around the question of whether chemical decoupling always precede kinetic decoupling:

(1) Does early chemical equilibrium imply a thermal velocity distribution at later times?
(2) Must a species that *is* in chemical equilibrium be in kinetic equilibrium with the Standard Model plasma?
(3) If the former does not hold, would chemical equilibrium imply kinetic equilibrium within a dark sector with its own temperature?

We demonstrate that the answer to all the above questions is no, through detailed analysis of concrete examples from neutrino physics, dark matter scenarios, and beyond-Standard Model cosmology. These scenarios illuminate the rich phenomenology possible when chemical and kinetic processes operate on different timescales, commonly encountered in beyond-Standard Model cosmology and dark sector physics. We provide a general framework for determining when chemical and kinetic equilibrium end, and develop detailed examples and models that exemplify these phenomena, analyze their common underlying physics, and explore some of their quantitative observational consequences.

II. GENERAL FORMALISM AND NOTATION

This section collects the kinematic conventions, kinetic equations, rate definitions, and analysis tools used throughout the paper. We aim to make this section fully self-contained by clearly defining every symbol and introducing standard references.

A. Phase space, Liouville flow, and comoving variables

We describe each species i by its phase-space density $f_i(p,t)$ obeying, in a homogeneous FRW background with scale factor $a(t)$,

$$(\partial_t - Hp\cdot\nabla_p) f_i(p,t) = C_i[f], \qquad H \equiv \dot{a}/a, \tag{6}$$

where H is the Hubble expansion rate. Here $C_i[f]$ is the collision operator describing interactions. We use natural units $\hbar = c = k_B = 1$ throughout. For general background on Boltzmann equations in cosmology, see Refs. [3,8].

It is often convenient to work with comoving momentum $q \equiv ap$ or with the dimensionless variable $y \equiv p/T_{\rm cm}$, where $T_{\rm cm} \propto a^{-1}$ is any reference comoving temperature (e.g., the photon temperature in the absence of entropy dumps). The number, energy, and pressure moments are

$$n_i = \frac{g_i}{(2\pi)^3}\int d^3p f_i(p), \qquad \rho_i = \frac{g_i}{(2\pi)^3}\int d^3p E_i(p) f_i(p),$$
$$P_i = \frac{g_i}{(2\pi)^3}\int d^3p \frac{p^2}{3E_i(p)} f_i(p), \tag{7}$$

with $E_i = \sqrt{p^2 + m_i^2}$ and internal degeneracy g_i. The factor g_i counts the number of internal spin, polarization, or color degrees of freedom.

B. Collision operator split and equilibrium concepts

We separate the collision operator C_i into inelastic (number-changing) and elastic (momentum-exchange) pieces,

$$C_i[f] = C_i^{\rm inel}[f] + C_i^{\rm el}[f], \tag{8}$$

where $C^{\rm inel}$ contains decays/inverse decays and $2 \leftrightarrow 2, 3 \leftrightarrow 2, \ldots$ reactions that change $\sum_i N_i$, while $C^{\rm el}$ contains scatterings that conserve particle number but exchange momentum and energy. This decomposition is standard in kinetic theory; see Ref. [9] for neutrino transport and Ref. [10] for dark matter applications.

a. Chemical equilibrium. For a given reaction network $\sum_j \nu_{aj} X_j \leftrightarrow 0$, chemical equilibrium imposes constraints on chemical potentials, $\sum_j \nu_{aj}\mu_j = 0$ for all active processes a. Operationally, letting $\Gamma_{\rm chem}$ denote the slowest relevant per-particle number-changing rate, chemical equilibrium obtains when

$$\Gamma_{\rm chem}(T) \gg H(T). \tag{9}$$

In practice, $\Gamma_{\rm chem}$ can represent annihilation, decay, or conversion rates. As long as these processes are much faster than Hubble expansion, number densities follow their equilibrium values.

b. Kinetic equilibrium. Kinetic equilibrium demands that momentum exchange be sufficiently fast to maintain a thermal distribution for f_i, typically Maxwell-Boltzmann/Fermi-Dirac/Bose-Einstein with some (T_i, μ_i). A robust diagnostic is the momentum-transfer rate γ_p, defined below. Kinetic equilibrium requires

$$\gamma_p^{(X)}(p) \gg H \quad \text{for the momenta that dominate} \quad n_i \quad \text{and} \quad \rho_i, \tag{10}$$

where the superscript (X) indicates the bath mediating kineticization [e.g., Standard Model (SM), or dark sector (DS)]. Because of transport weighting (see the discussion in Sec. II D below), high-p modes typically decouple first. This diagnostic has been extensively used in studies of dark matter kinetic decoupling [10,11]. For distribution-level studies of elastic scattering in both freeze-in and freeze-out scenarios, see Ref. [12]. In agreement with our argument, it was shown that a sufficiently large elastic scattering rate can establish kinetic equilibrium even during freeze-in, where the condition for chemical equilibrium is never satisfied.

C. Reaction rates and thermally averaged kernels

For $2 \leftrightarrow 2$ inelastic reactions $a + b \leftrightarrow c + d$,

$$\Gamma_{a,\rm inel} = \sum_b n_b \langle \sigma v_{\rm M\o l} \rangle_{ab\to\ldots}, \quad \langle \sigma v_{\rm M\o l} \rangle = \frac{\int d\Pi_a d\Pi_b f_a f_b \sigma v_{\rm M\o l}}{\int d\Pi_a d\Pi_b f_a f_b}, \tag{11}$$

with Lorentz-invariant measure $d\Pi = g d^3p/[(2\pi)^3 2E]$ and Møller velocity $v_{\rm M\o l}$. The thermal average $\langle \sigma v_{\rm M\o l} \rangle$ ensures relativistic invariance and is standard in relic density calculations [8].

For decays/inverse decays $X \leftrightarrow ij$,

$$\Gamma_{X,\rm chem} \simeq \Gamma_{X\to ij} \times \frac{n_X}{n_X^{(\rm eq)}(T)}, \tag{12}$$

with finite-density effects included when needed.

For $3 \to 2$ reactions in a self-thermalized sector,

$$\Gamma_{3\to 2} \equiv n^2 \langle \sigma v^2 \rangle_{3\to 2}. \tag{13}$$

During radiation domination,

$$H(T) \simeq 1.66 g_\star^{1/2}(T) \frac{T^2}{M_{\rm Pl}}, \tag{14}$$

where $g_\star(T)$ is the effective number of relativistic degrees of freedom contributing to the energy density [3], and $M_{\rm Pl} \simeq 1.22 \times 10^{19}$ GeV is the (nonreduced) Planck mass.

D. Transport (momentum-exchange) rate and heat exchange

Elastic $2 \to 2$ scattering $it \to it$ contributes to *momentum* and *heat* exchange. The appropriate cross section is transport-weighted,

$$\sigma_{\rm mt}(s) \equiv \int d\Omega (1 - \cos\theta) \frac{d\sigma}{d\Omega}, \tag{15}$$

which suppresses forward scattering. This "transport weighting" ensures that scatterings which only deflect the incoming particle by a very small angle, while perhaps dominating the total elastic cross section, contribute negligibly to momentum transfer. In other words, the factor $(1 - \cos\theta)$ penalizes forward scattering, so that $\sigma_{\rm mt}$ correctly captures the rate at which the momentum of particle i is randomized through interactions with the target bath t (see, e.g., Refs. [11,13,14]).

For a heavy species i scattering on relativistic targets t with temperature T_t,

$$\gamma_p^{(t)}(i) \simeq n_t \langle \sigma_{\rm mt} v \rangle \frac{T_t}{m_i}, \tag{16}$$

up to $\mathcal{O}(1)$ factors. Here n_t is the number density of target particles, $\langle\sigma_{\rm mt}v\rangle$ is the thermal average of the transport cross section, T_t is the bath temperature, and m_i is the mass of the heavy scattering particle. The factor T_t/m_i reflects the fractional momentum transfer per collision in the heavy-light limit. This rate γ_p is often called the "momentum transfer rate" or "drag rate" in the context of dark matter-baryon or dark matter-radiation interactions (see, e.g., Refs. [10,15]).

A corresponding *heat-exchange* (energy-transfer) rate γ_E can be defined analogously, and in small-angle regimes $\gamma_E \sim \gamma_p$ (precise proportionality depends on kinematics). In practice, the ratio γ_E/γ_p depends on how efficiently kinetic energy is exchanged relative to momentum exchange; for example, in Rutherford-like scattering with light targets the two are nearly proportional, whereas for massive targets the proportionality can differ. Both γ_p and γ_E must be computed including plasma effects such as Debye screening, finite-temperature masses, and Coulomb logarithms (see, e.g., Refs. [10,15]). These refinements are essential whenever long-range mediators or gauge interactions are involved.

In summary, γ_p quantifies the rate at which the bulk momentum distribution of a species is isotropized, while γ_E controls the rate of thermal energy exchange with the bath. Kinetic equilibrium is maintained as long as $\gamma_p/H \gg 1$ at the momenta that dominate the number and energy densities. Once γ_p/H falls below unity, elastic scattering is no longer efficient enough to sustain a Maxwell-Boltzmann (or Fermi-Dirac/Bose-Einstein) distribution, and the system kinetically decouples.

For numerical treatments that go beyond such analytical approaches, see, e.g., Ref. [16], where the public code DRAKE was introduced as a dedicated tool to follow departures from local thermal equilibrium at the distribution level.

E. Free-streaming solution after kinetic decoupling

If at some time $t_{\rm kd}$ the appropriate γ_p/H falls below unity across populated momenta and inelastic terms are negligible thereafter, Eq. (6) integrates along characteristics to

$$f_i(p,t) = f_i\left(\frac{a(t)}{a(t_{\rm kd})}p, t_{\rm kd}\right) + \int_{t_{\rm kd}}^{t} dt' S_i^{\rm inel}\left(\frac{a(t)}{a(t')}p, t'\right), \tag{17}$$

where $S^{\rm inel}$ collects any subsequent injection/depletion sources. Thus the postdecoupling spectrum is a redshifted memory of the distribution at kinetic decoupling, plus any later injection; no single (T,μ) need describe it. This statement underlies free-streaming effects of neutrinos [9] and warm dark matter [17].

F. Two-bath cosmologies and temperatures

When a dark sector communicates feebly with the SM, it is useful to track an SM temperature T and a DS temperature T_d with T_d/T evolving by entropy exchange. Portal energy flow $Q_{\rm portal}$ enters the temperature evolution via

$$\dot\rho_{\rm SM} + 3H(\rho_{\rm SM} + P_{\rm SM}) = +Q_{\rm portal}, \tag{18}$$

$$\dot\rho_{\rm DS} + 3H(\rho_{\rm DS} + P_{\rm DS}) = -Q_{\rm portal}, \tag{19}$$

with $Q_{\rm portal}$ obtained from the appropriate collision integrals. Within the DS, species may share T_d (fast intra-DS elastic) or carry distinct kinetic temperatures T_i (marginal intra-DS elastic), in which case one evolves separate heat equations. Such two-temperature setups are standard in hidden-sector cosmology [18,19]. Explicit examples of two-component dark sectors and the impact of DM self-scatterings on equilibration were analyzed numerically in Ref. [20].

G. Exact collision integrals vs. Fokker-Planck/Langevin

a. Full collision integrals ("fBE"). Exact $2 \leftrightarrow 2$ kernels with quantum statistics and (when needed) $3 \leftrightarrow 2$ are mandatory when (i) annihilation/production is sharply momentum-selective (narrow resonances, forbidden/threshold kinematics), (ii) large-angle scatterings matter, or (iii) precision neutrino transport is required. Efficient formulations reduce the nine-dimensional integrals to two-dimensional kernels in $y = p/T_{\rm cm}$.

b. Fokker-Planck (FP)/Langevin. In small-angle regimes (forward t-channel exchange), elastic scattering is accurately captured by an FP operator,

$$C_{\rm el}^{\rm FP}[f] = \partial_{p_i}[A_i(p)f + \partial_{p_j}(B_{ij}(p)f)], \tag{20}$$

with drag A_i and diffusion B_{ij} matched to the exact momentum *and* heat exchange (thereby reproducing γ_p and γ_E). A Langevin limit is efficient for heavy species in a relativistic bath. FP/Langevin should be benchmarked against full kernels near thresholds or when large-angle scatterings contribute appreciably. Full Boltzmann integrals are mandatory in cases of sharp momentum selectivity, while Fokker-Planck approximations apply in forward-scattering regimes. For applications to neutrino transport and DM kinetic decoupling, see Refs. [9,10].

H. Moment hierarchies and controlled closures

Integrating Eq. (6) yields hierarchies for $n, \rho, P, \ldots$. A widely used controlled closure retains number density and a kinetic temperature T_i (or second moment); schematically,

$$\dot{n}_i + 3Hn_i = [\text{inelastic sources/sinks(decays, (co)ann., } 3 \to 2)], \tag{21}$$

$$\frac{3}{2}(\dot{n}_i T_i + n_i \dot{T}_i) + 5Hn_i T_i = -\mathcal{Q}_{\rm el}(T_i - T_{\rm bath}) + \mathcal{Q}_{\rm inel}, \tag{22}$$

where $\mathcal{Q}_{\rm el} \propto \gamma_p$ and $\mathcal{Q}_{\rm inel}$ accounts for energy nonconservation per particle in number-changing reactions (e.g., self-heating from $3 \to 2$, recoil in inverse decays). This closure is accurate if f_i remains close to Maxwellian; when high-p tails or narrow bands dominate chemistry, one must resort to targeted f-space evolution. This controlled closure is accurate as long as distributions remain close to Maxwellian. Deviations require momentum-resolved treatments, as emphasized in [10,21].

I. Chemical networks and chemical potentials

For a network of reactions $\sum_j \nu_{aj} X_j \leftrightarrow 0$ in chemical equilibrium, solve $\sum_j \nu_{aj}\mu_j = 0$ for $\{\mu_j\}$ (with conserved charges as Lagrange multipliers). For example,

(i) semiannihilation $\chi\chi \leftrightarrow \chi\phi$ implies $2\mu_\chi = \mu_\chi + \mu_\phi$;
(ii) $3 \leftrightarrow 2$ cannibalism $\chi\chi\chi \leftrightarrow \chi\chi$ implies $3\mu_\chi = 2\mu_\chi$ so $\mu_\chi = 0$ in the chemically equilibrated DS (if only χ carries number);
(iii) coannihilation/conversions enforce $\mu_\chi \simeq \mu_\psi$.

When $\Gamma_{\rm chem} \gg H$ but $\gamma_p \lesssim H$, these relations hold at the level of *number* densities while f_i need not be thermal. Examples include semiannihilation $\chi\chi \leftrightarrow \chi\phi$, $3 \leftrightarrow 2$ cannibal processes, and coannihilation/conversions, see [22,23]. For conversion-driven freeze-out, distribution-level studies have recently been carried out in Ref. [24], complementing earlier moment-level analyses of semiannihilations and cannibal processes.

III. EXAMPLE SETUP: DECAYS/INVERSE DECAYS WITHOUT ELASTIC SCATTERING

To make the distinction between chemical and kinetic equilibrium explicit, consider a toy model in which a species X, injected, in general, out of thermal equilibrium, interacts with a thermal bath only via number-changing decays and inverse decays

$$X \leftrightarrow Y + Z, \tag{23}$$

where Y and Z are lighter particles in kinetic and chemical equilibrium with the bath. We assume that *elastic* scattering processes involving X,

$$X + \text{SM} \to X + \text{SM}, \tag{24}$$

are negligible.

a. Chemical equilibrium. The decay and inverse-decay rates $\Gamma_{\rm dec}$ are taken to satisfy $\Gamma_{\rm dec} \gg H$, so that the integrated number density n_X closely follows its equilibrium value $n_X^{\rm eq}(T)$. For every X removed by decay, an inverse decay promptly produces another, ensuring detailed balance in the *number* of X particles.

b. Loss of kinetic equilibrium. The momentum distribution $f_X(p,t)$, however, is determined entirely by the kinematics of the inverse decays, which inject X particles at a characteristic momentum p_0 in the bath rest frame. Without frequent elastic scatterings to redistribute momenta, the spectrum does not relax to the Maxwell-Boltzmann form even though $n_X = n_X^{\rm eq}$. This is the simplest realization of the hierarchy

$$\Gamma_{\rm inelastic} \equiv \Gamma_{\rm dec} \gg H \gg \Gamma_{\rm elastic}.$$

c. Including elastic scattering. If we now add an elastic scattering term with rate γ, the steady-state distribution becomes a weighted average of the injection spectrum and the Maxwell-Boltzmann spectrum,

$$f_\infty(p) = \frac{S(p) + \gamma f_{\rm MB}(p)}{\Gamma_{\rm dec} + \gamma}. \tag{25}$$

When $\gamma \to 0$, we recover the purely nonthermal steady state; when $\gamma \gg \Gamma_{\rm dec}$, we recover kinetic equilibrium.

d. Numerical illustration. Figures 1 and 2 show this explicitly for a choice of parameters where the initial nonthermal spectrum is a narrow Gaussian centered away from the thermal peak (shown in green in Fig. 2). We plot the *number spectrum* $dN/dp \propto p^2 f(p)$ on log-log scales, normalized so that the area matches that of the Maxwell-Boltzmann spectrum (chemical equilibrium). The first plot varies γ to show the approach to kinetic equilibrium; the second shows the time evolution toward the steady state for $\gamma/\Gamma_{\rm dec} = 0.5$, in units of the timescale $(\Gamma_{\rm dec} + \gamma)^{-1}$.

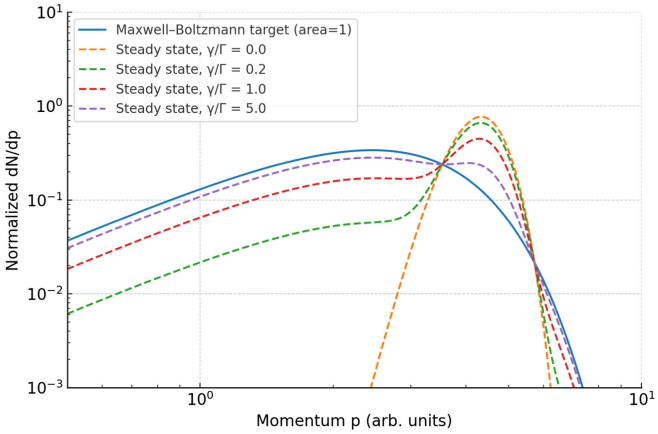


FIG. 1. Steady-state number spectra for various elastic scattering rates $\gamma/\Gamma_{\rm dec}$. All spectra are normalized to the same total number (chemical equilibrium), but only for large γ does the shape match the Maxwell-Boltzmann distribution (kinetic equilibrium).

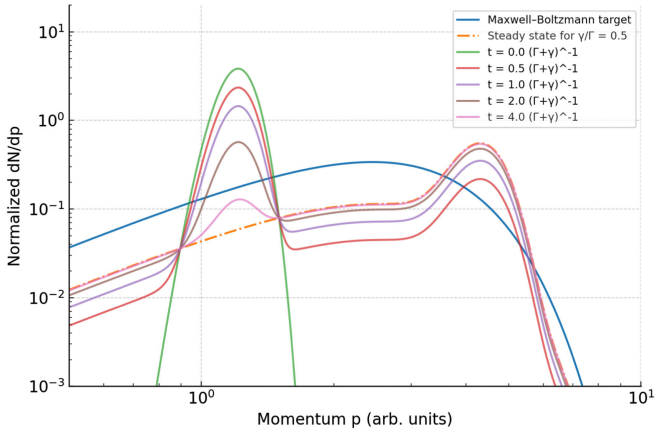


FIG. 2. Time evolution toward the steady state for $\gamma/\Gamma_{\rm dec} = 0.5$. The distribution starts far from thermal (green line) and converges to a mixed nonthermal/thermal shape over a timescale $(\Gamma_{\rm dec} + \gamma)^{-1}$, with the total number fixed at the equilibrium value throughout.

The triple-peaked momentum spectrum in Fig. 2 results from the interplay between (i) elastic scattering with the thermal bath, which drives the distribution toward Maxwell-Boltzmann form, and (ii) inverse decays, which inject particles with a characteristic momentum. The leftmost, green peak is the initial $t = 0$ nonthermal injection spectrum, assumed to be a Gaussian centered below the thermal peak, which vanishes in the steady-state regime shown in Fig. 1.

Thermal peak. Elastic scattering tends to establish a Maxwell-Boltzmann distribution,

$$\frac{dN_{\rm MB}}{dp} \propto p^2 e^{-p^2/(2mT)}, \tag{26}$$

which reaches its maximum at

$$p_{\rm MB} = \sqrt{2mT}. \tag{27}$$

For the parameters used in the figure ($m = 3$, $T = 1$), this gives

$$p_{\rm MB} \simeq \sqrt{6} \simeq 2.45, \tag{28}$$

corresponding to the left-hand peak.

Injection peak. Inverse decays produce particles with a narrow momentum distribution,

$$S(p) \propto \exp\left[-\frac{(p-p_0)^2}{2\sigma^2}\right], \tag{29}$$

so that the number spectrum is

$$\frac{dN_{\rm inj}}{dp} \propto p^2 S(p). \tag{30}$$

Maximizing $\ln(p^2 S(p))$ yields the location of the injection peak,

$$\frac{2}{p} - \frac{p-p_0}{\sigma^2} = 0 \Rightarrow p_{\rm inj} = \frac{p_0 + \sqrt{p_0^2 + 8\sigma^2}}{2}. \tag{31}$$

For the numerical choices $p_0 = 2.5\sqrt{mT} \simeq 4.33$ and $\sigma = 0.3\sqrt{mT} \simeq 0.52$, one finds

$$p_{\rm inj} \simeq 4.45, \tag{32}$$

corresponding to the right-hand peak.

Combined spectrum. The full steady-state distribution is a weighted average of the thermal and injection contributions,

$$\left.\frac{dN}{dp}\right|_{\infty} = \frac{S(p) + \gamma \frac{dN_{\rm MB}}{dp}}{\Gamma + \gamma}, \tag{33}$$

where Γ is the decay/inverse-decay rate and γ the elastic scattering rate. When Γ and γ are comparable, both peaks are clearly visible: the thermal bump at $p \simeq 2.45$ and the injection bump at $p \simeq 4.45$.

IV. DOES PAST CHEMICAL EQUILIBRIUM ENFORCE THERMAL VELOCITIES AT LATER TIMES?

Claim: A species that *was* in chemical equilibrium at an earlier epoch need not possess a thermal (Maxwell-Boltzmann/Fermi-Dirac/Bose-Einstein) velocity distribution at later times, whether or not chemical decoupling happens prior to kinetic decoupling.

A. Kinetic rationale

Let $f(p,t)$ be the phase-space density of the species. Its evolution obeys Eq. (6). Chemical equilibrium at time $t_\star$ constrains the *zeroth moment* (abundance) via $\Gamma_{\rm chem}(t_\star) \gg H(t_\star)$, but it does not determine the future *shape* of f. If, at some later time $t_{\rm kd}$, the momentum-transfer (transport) rate $\gamma_p \sim n_{\rm bath}\langle\sigma_{\rm mt} v\rangle T_{\rm bath}/m$ falls below the Hubble rate for populated momenta, $\gamma_p(p) \ll H$, elastic scattering no longer maintains a thermal form. Neglecting $C_{\rm el}$ afterward, (6) integrates to Eq. (17) where as noted above $\mathcal{S}_{\rm inel}$ represents inelastic sources/sinks per comoving momentum. Equation (17) shows that the postkinetic-decoupling spectrum is a redshifted memory of the shape at $t_{\rm kd}$, plus any later injection/depletion; neither operation forces a return to a single-temperature form. In particular, high-p modes typically violate $\gamma_p(p) \gg H$ first (transport weighting $(1-\cos\theta)$ penalizes forward scatterings), so deviations from thermality seed from the top of the distribution and grow with time. We show two examples drawn from neutrino physics where chemical decoupling

precedes/follows kinetic decoupling, and where the asymptotic momentum distribution is not thermal.

B. Standard-Model neutrinos (MeV epoch)

a. Setup and physical scales. Around the MeV epoch the active neutrinos $\nu = \{\nu_e, \nu_\mu, \nu_\tau\}$ depart from full equilibrium with the electromagnetic (EM) plasma $\{e^\pm, \gamma\}$. Hubble expansion during radiation domination is $H \simeq 1.66 g_*^{1/2} T^2/M_{\rm Pl}$, while weak interaction rates scale roughly as $\sim G_F^2 T^5$ up to order-unity angular/energy weights. Finite-temperature QED corrections modify the EM equation of state and electron mass around $T \sim m_e$, and three-flavor oscillations redistribute energy and distortions among flavors. Momentum-resolved transport (Boltzmann or QKE) with exact $2 \leftrightarrow 2$ weak kernels, finite-T QED thermodynamics, and oscillations provides the modern baseline [25–28].

b. Chemical vs. kinetic rate hierarchy. Two operator classes in the weak collision term behave differently as T falls through m_e:

(i) *Chemical (number-changing):* $\nu\bar{\nu} \leftrightarrow e^+e^-$ controls μ_ν and the possibility of tracking an equilibrium abundance. Its per-particle rate scales as $\Gamma_{\rm chem} \sim n_{e^\pm}\langle\sigma v\rangle_{\nu\bar{\nu}\leftrightarrow e^+e^-} \propto n_{e^\pm} G_F^2 T^2$. As $T \lesssim m_e$, $n_{e^\pm}$ drops exponentially and $\Gamma_{\rm chem}/H$ falls fastest.

(ii) *Kinetic (momentum/energy exchange):* elastic ν-$e^\pm$ scattering, $\nu e^\pm \leftrightarrow \nu e^\pm$, continues to exchange momentum/heat with the EM bath with a transport rate

$$\gamma_p(p) \simeq n_{e^\pm}\langle\sigma_{\rm mt}(\nu e) v\rangle \frac{T}{E_\nu},$$
$$\sigma_{\rm mt} \equiv \int d\Omega (1-\cos\theta)\frac{d\sigma}{d\Omega}, \tag{34}$$

which, although it also weakens as $n_{e^\pm}$ declines, typically remains $\gtrsim H$ for a while *after* pair processes have become inefficient. Neutrino-neutrino scatterings help redistribute within the ν sector but cannot by themselves maintain kinetic coupling to the EM bath.

The generic ordering during decoupling is therefore

$$\Gamma_{\rm chem} \downarrow \text{first} \quad \text{and} \quad \gamma_p \downarrow \text{later},$$

so that chemical decoupling precedes kinetic decoupling.

c. Momentum-space consequences. Solving

$$(\partial_t - Hp\partial_p) f_\nu(p,t) = C_{\rm ann}[f_\nu, f_{e^\pm}] + C_{\rm el}[f_\nu, f_{e^\pm}] + C_{\rm osc}[\varrho], \tag{35}$$

where $C_{\rm osc}$ captures the collision operator associated with neutrino oscillations, reveals small, irreducible spectral distortions in f_ν: (i) high-p modes (in comoving $y \equiv p/T$) remain more strongly coupled and receive a larger share of the late EM entropy flow, generating *larger* deviations there; (ii) flavor oscillations redistribute these distortions, largely equilibrating ν_e and $\nu_{\mu,\tau}$ features by the end of $e^\pm$ annihilation. No single (T_ν, μ_ν) can exactly reproduce the momentum dependence of the relic spectra; a two-parameter ansatz misses the few $\times 10^{-3} - 10^{-2}$ level y-dependent structure seen in full solutions [25–28].

d. Entropy/energy flow and timing. Because ν-$e^\pm$ scattering persists *into* the $e^\pm$ annihilation epoch, part of the entropy released from $e^\pm \to \gamma$ is transferred to neutrinos. Momentum-resolved computations consistently find significant EM $\leftrightarrow$ ν energy exchange well below the naive "instantaneous" decoupling temperature, extending to $T \sim$ 0.1–0.2 MeV in the EM bath. This noninstantaneous reheating of neutrinos is exactly what produces the final-state distortions and the small upward shift in the radiation density.

C. Heavy right-handed neutrinos with strong washout

a. Setup and rate hierarchy. In thermal leptogenesis from a type-I seesaw, heavy Majorana neutrinos N (mass M_N) couple to lepton doublets L and the Higgs H via $\mathcal{L} \supset -y_N \bar{N} L H + \text{H.c.}$ The dominant *inelastic* processes are decays and inverse decays,

$$N \leftrightarrow LH, \qquad \Gamma_D^0 \equiv \frac{|y_N|^2 M_N}{8\pi}, \tag{36}$$

with finite-density/statistical corrections at $T \sim M_N$. The strong-washout regime is characterized by

$$K \equiv \left.\frac{\Gamma_D}{H}\right|_{T=M_N} = \frac{|y_N|^2 M_N}{8\pi}\frac{M_{\rm Pl}}{1.66 g_*^{1/2} M_N^2} \simeq \frac{\tilde{m}}{m_*} \gg 1, \tag{37}$$

so that inverse decays keep n_N close to $n_N^{(\rm eq)}(T)$ around $T \sim M_N$ [29] ("chemical equilibrium").

By contrast, the *kinetic* coupling of N to the SM bath is governed by elastic $2 \to 2$ processes such as $N\ell \to N\ell$ and $NH \to NH$ via t-channel exchange of H or ℓ. These amplitudes scale as y_N^2 and the (differential) cross sections as y_N^4. Moreover, the small-angle dominance at high T makes the transport cross section

$$\sigma_{\rm mt} \equiv \int d\Omega (1-\cos\theta)\frac{d\sigma}{d\Omega}$$

parametrically smaller than $\sigma_{\rm el}$. The momentum-exchange (transport) rate for a heavy N then reads

$$\gamma_p^{(\rm SM)}(N) \simeq n_{\rm bath}(T)\langle\sigma_{\rm mt}^{(\rm SM)} v\rangle \frac{T}{M_N} \sim \frac{|y_N|^4 T^2}{M_N}\langle 1-\cos\theta\rangle, \tag{38}$$

so that, at $T \sim M_N$,

$$\frac{\gamma_p^{(\mathrm{SM})}(N)}{\Gamma_D} \sim |y_N|^2 \frac{T^2}{M_N^2} \langle 1 - \cos\theta \rangle \ll 1 \quad (K \gg 1). \tag{39}$$

Hence one naturally realizes

$$\underbrace{\Gamma_{\mathrm{chem}} \sim \Gamma_D \gg H}_{\text{chemical eq}} \quad \text{while} \quad \underbrace{\gamma_p^{(\mathrm{SM})}(N) \ll H}_{\text{SM-kinetic decoupling}},$$

i.e., N is chemically coupled but not fully thermalized in *momentum* with the SM bath.

b. Momentum-space evolution and relaxation. The phase-space distribution $f_N(p,t)$ obeys

$$(\partial_t - Hp\partial_p) f_N = C_{\mathrm{ID/D}}[f_N; f_L, f_H] + C_{\mathrm{el}}^{(\mathrm{SM})}[f_N; f_L, f_H]. \tag{40}$$

Inverse decays tend to drive f_N toward equilibrium, but their *relaxation rate is energy dependent*

$$\Gamma_{\mathrm{relax}}(E) \sim \Gamma_D \frac{M_N}{E} \times (\text{finite-density factors}), \tag{41}$$

so the highest-p modes equilibrate slowest (time-dilation/phase-space penalty). When $C_{\mathrm{el}}^{(\mathrm{SM})}$ is transport-suppressed as in (38), the net result is

(i) n_N tracks $n_N^{(\mathrm{eq})}$ (chemical equilibrium) for $T \sim M_N$;
(ii) f_N can retain residual high-p distortions (especially if production began nonthermally, e.g., from inflaton or cascade decays) because elastic refilling is too slow;
(iii) once inverse decays thin ($\Gamma_D/H \downarrow$), the nonthermal shape free-streams and simply redshifts [cf. Eq. (17)].

c. Finite-density/thermal corrections. Thermal masses $m_H(T), m_L(T) \sim gT$, Pauli blocking/Bose enhancement, and off-shell scatterings modify both $C_{\mathrm{ID/D}}$ and $C_{\mathrm{el}}^{(\mathrm{SM})}$. Thermal field-theory treatments (HTL/resummed propagators, finite-density cutting rules) confirm that the qualitative hierarchy (39) persists: washout $\propto |y_N|^2$ dominates over SM-elastic transport $\propto |y_N|^4$, with the latter further reduced by transport weighting and small-angle kinematics [30]. These effects also sharpen the energy dependence in (41), reinforcing the late relaxation of high-p modes.

d. Initial conditions and memory. Even in strong washout, the *shape* of f_N can remember nonthermal origins if they preferentially populate hard modes:

(i) *Inflaton/cascade decays.* Production at $p \sim \mathcal{O}(M_\phi/2)$ (with ϕ heavy) seeds a hard f_N; inverse decays thermalize number quickly but not all momenta before T drops.
(ii) *Scattering production (UV freeze-in before equilibration).* If N first appears via $2 \to 2$ scatterings at $T \gg M_N$, the emerging f_N can be nonthermal; once $T \sim M_N$ and $K \gg 1$, inverse decays pin n_N while f_N only partially relaxes.

Subsequent leptogenesis then proceeds with $n_N \simeq n_N^{(\mathrm{eq})}$ but a not-quite-Maxwellian f_N, affecting the detailed timing of washout and CP-asymmetry build-up at the $\mathcal{O}(1)$-few × 10% level in precision studies.

e. Numerical parametrics. Using $|y_N| \simeq \sqrt{m_\nu M_N}/v$ with $m_\nu \simeq 0.05$ eV and $v = 174$ GeV gives $|y_N| \sim 4 \times 10^{-3}$ for $M_N = 10^{10}$ GeV. Then $\Gamma_D \sim |y_N|^2 M_N/(8\pi) \sim 10^3$–$10^4$ GeV, while $H(T = M_N) \sim 10^2$ GeV ($g_* \sim 100$). A transport estimate from (38) yields $\gamma_p^{(\mathrm{SM})} \sim \mathcal{O}(1$–$10)$ GeV before the small-angle penalty, hence $\Gamma_D/H \gg 1$ but $\gamma_p^{(\mathrm{SM})}/H \lesssim 0.1$: chemical equilibrium holds; full SM-kinetic equilibrium does not.

f. Flavor and quantum-kinetic aspects. At $T \sim M_N$ one may need a density-matrix (flavor) formulation for L and a proper treatment of soft/collinear corrections in $C_{\mathrm{ID/D}}$. These refinements affect the precise washout timing but not the qualitative separation $\Gamma_D \gg \gamma_p^{(\mathrm{SM})}$ in strong washout. Including them ensures consistent energy conservation and detailed balance in the coupled N-plasma system [29,30].

D. Sufficient conditions for a thermal late-time shape

Thermal velocities at late times *can* occur as a result of a species being in chemical equilibrium if either:

(1) the species remains elastically coupled to a bath with $\gamma_p/H \gg 1$ well after chemical freeze-out (e.g., gauge-strength coupling to a relativistic plasma), or
(2) the species belongs to a self-thermalized dark sector with fast internal $2 \leftrightarrow 2$ (and possibly $3 \leftrightarrow 2$) processes that maintain a common dark temperature $T_d \neq T$.

Absent these conditions, past chemical equilibrium does not constrain the subsequent velocity distribution.

V. DOES CHEMICAL EQUILIBRIUM IMPLY KINETIC EQUILIBRIUM WITH THE STANDARD MODEL PLASMA?

Claim: No. A species can satisfy *chemical* equilibrium conditions while being *kinetically decoupled* from the Standard Model bath. Chemical equilibrium constrains the abundance via number-changing reactions with rate $\Gamma_{\mathrm{chem}} \gg H$, whereas kinetic equilibrium with the SM requires a sufficiently large SM-elastic *transport* rate $\gamma_p^{(\mathrm{SM})} \gg H$ across the occupied momenta. The two are governed by different operators and kinematics (e.g., s-channel decays/annihilations vs. forward-peaked t-channel elastic scatterings), so Γ_{chem} and $\gamma_p^{(\mathrm{SM})}$ can be parametrically and numerically separated.

A. Kinetic rationale

Let $f(p,t)$ denote the species' phase-space density. Its evolution obeys

$$(\partial_t - Hp\partial_p)f = C_{\rm el}^{(\rm SM)}[f] + C_{\rm el}^{(\rm DS)}[f] + C_{\rm inel}[f], \quad (42)$$

where $C_{\rm inel}$ encodes number-changing reactions (to or from SM states), $C_{\rm el}^{(\rm SM)}$ elastic exchange with the SM bath, and $C_{\rm el}^{(\rm DS)}$ elastic exchange *within* a dark sector (if present). Chemical equilibrium only demands $\Gamma_{\rm chem} \gg H$. Kinetic equilibrium *with the SM* demands a large *SM* transport rate,

$$\gamma_p^{(\rm SM)} \simeq n_{\rm SM}\langle\sigma_{\rm mt}^{(\rm SM)}v\rangle\frac{T}{m}, \qquad \sigma_{\rm mt} \equiv \int d\Omega(1-\cos\theta)\frac{d\sigma}{d\Omega}, \quad (43)$$

which is often much smaller than the *inelastic* rate when elastic scattering is t-channel and forward-peaked (transport weighting penalizes small angles), suppressed by extra couplings, or requires heavier mediators [10,31]. If $\gamma_p^{(\rm SM)} \lesssim H$ while $\Gamma_{\rm chem} \gg H$, the species remains chemically tied to the SM but its momentum distribution is *not* thermalized to the SM temperature. If $C_{\rm el}^{(\rm DS)}$ is fast, the species may still be thermal *within a dark bath* at temperature $T_d \neq T$. We discuss two instances below.

B. Conversion-driven/coannihilation freeze-out with a secluded DM state

a. Setup. Let χ be the DM candidate and ψ a slightly heavier coannihilating partner with mass splitting $\Delta \equiv (m_\psi - m_\chi)/m_\chi \ll 1$. The relevant reactions are

$$\text{(i) conversions:} \quad \chi + X \leftrightarrow \psi + X' \ \text{ and } \ \psi \leftrightarrow \chi + X; \quad (44)$$

$$\text{(ii) (co)annihilations:} \quad \chi\psi, \psi\psi \leftrightarrow \text{SM SM}. \quad (45)$$

Here X, X' are SM quanta (or light dark-sector states that remain in thermal contact with the SM). The chemical relation

$$\mu_\chi \simeq \mu_\psi \Rightarrow \frac{n_\chi}{n_\psi} \simeq \frac{n_\chi^{(\rm eq)}}{n_\psi^{(\rm eq)}} \simeq e^{+\Delta x}, \qquad x \equiv \frac{m_\chi}{T}, \quad (46)$$

is enforced by the *inelastic* conversion network (44) provided its per-particle rate

$$\Gamma_{\rm conv} \equiv \Gamma_{\psi\to\chi X} + n_{\rm SM}\langle\sigma v\rangle_{\chi X\leftrightarrow\psi X'} \gg H \quad (47)$$

throughout the epoch of interest. The final DM abundance is then dominantly controlled by (45) (often $\psi\psi \to$ SM), i.e., by the *partner's* efficient annihilations.

b. Chemical vs. kinetic hierarchies. Kinetic equilibrium *with the SM* requires a sufficiently large SM-elastic *transport* rate for χ. In many models the direct elastic portal is t-channel and forward-peaked (heavy mediator, small coupling), so $\sigma_{\rm mt} \ll \sigma_{\rm el}$ and $\gamma_p^{(\rm SM)}(\chi) \lesssim H$ while *chemical* conversions and coannihilations still satisfy (47). The outcome is

$$\Gamma_{\rm chem} \sim \Gamma_{\rm conv} + n_\psi\langle\sigma v\rangle_{\psi\psi} \gg H \ \text{ but } \ \gamma_p^{(\rm SM)}(\chi) \lesssim H, \quad (48)$$

i.e., χ is chemically tied to the SM *via* the partner network, yet kinetically decoupled from the SM bath.

c. Momentum-space picture. Conversions and inverse decays (44) enforce (46) by reshuffling *identities*, not by efficiently randomizing momenta. Inverse decays produce monoenergetic daughters in the parent rest-frame; $2 \leftrightarrow 2$ conversions typically proceed through forward-peaked t-channel exchange, which carries the transport penalty $(1-\cos\theta)$. Consequently, even when n_χ/n_ψ tracks (46), the *shape* of $f_\chi(p)$ need not be thermal at the SM temperature. Depending on the dark elastic coupling, χ can be

(1) *nonthermal*: if χ-ψ elastic scattering is also slow ($\gamma_p^{\chi\leftrightarrow\psi} \lesssim H$), f_χ develops momentum-dependent distortions during freeze-out and then free-streams (redshifts) thereafter;

(2) *thermalized within a dark bath*: if χ-ψ elastic is fast while both species are SM-decoupled, the pair attains a common *dark* temperature $T_d \neq T$ (thermalized-but-secluded), even as (46) is maintained by inelastic conversions to SM quanta.

d. Coupled Boltzmann system. A transparent way to exhibit the split is to evolve the two number densities *and* at least one kinetic moment (e.g., a dark temperature T_d) or, when needed, the full $f(p,t)$. The integrated system reads schematically

$$\begin{aligned}\dot n_\chi + 3Hn_\chi = &-\langle\sigma v\rangle_{\chi\chi\to\rm SM}(n_\chi^2 - n_{\chi,\rm eq}^2)\\ &-\langle\sigma v\rangle_{\chi\psi\to\rm SM}(n_\chi n_\psi - n_{\chi,\rm eq}n_{\psi,\rm eq})\\ &+\Gamma_{\psi\to\chi X}\left[n_\psi - \frac{n_{\psi,\rm eq}}{n_{\chi,\rm eq}}n_\chi\right]\\ &+n_{\rm SM}\langle\sigma v\rangle_{\psi X'\to\chi X}\left[n_\psi - \frac{n_{\psi,\rm eq}}{n_{\chi,\rm eq}}n_\chi\right], \end{aligned} \quad (49)$$

$$\begin{aligned}\dot n_\psi + 3Hn_\psi = &-\langle\sigma v\rangle_{\psi\psi\to\rm SM}(n_\psi^2 - n_{\psi,\rm eq}^2)\\ &-\langle\sigma v\rangle_{\chi\psi\to\rm SM}(n_\chi n_\psi - n_{\chi,\rm eq}n_{\psi,\rm eq})\\ &-(\dot n_\chi + 3Hn_\chi)|_{\rm conv}, \end{aligned} \quad (50)$$

with "conv" picking the conversion terms from (49). Chemical equilibrium between χ and ψ corresponds to the square brackets $\propto [n_\psi - (n_{\psi,\rm eq}/n_{\chi,\rm eq})n_\chi]$ vanishing rapidly by (47), irrespective of the size of the SM-elastic

transport for χ. Kinetic information can be tracked by a dark temperature T_d obeying a heat-exchange equation (when a Maxwellian ansatz is tenable) or by a Fokker-Planck/Langevin operator for $\chi \leftrightarrow \psi$ elastic scattering; momentum-resolved treatments (full f) are required once high-p tails or thresholds dominate the chemistry.

e. Parametrics. If the portal proceeds via a heavy mediator of mass M with coupling $g_\chi g_{\rm SM}$, then

$$\Gamma_{\rm conv} \sim n_{\rm SM} \frac{(g_\chi g_{\rm SM})^2 T^2}{M^4},$$
$$\gamma_p^{(\rm SM)}(\chi) \sim n_{\rm SM} \frac{(g_\chi g_{\rm SM})^2 T^3}{M^4 m_\chi} \times \underbrace{\langle 1-\cos\theta\rangle}_{\ll 1}, \qquad (51)$$

so that $\Gamma_{\rm conv}/\gamma_p^{(\rm SM)} \sim (m_\chi/T)\langle(1-\cos\theta)^{-1}\rangle \gg 1$ for $m_\chi \gg T$ and forward-peaked kinematics. Decays $\psi \to \chi X$ add a term $\Gamma_{\psi\to\chi X} \propto g_\chi^2 \Delta m$ to (47) that scales as g_χ^2 rather than g_χ^4, further widening the chemical-kinetic separation.

f. Phenomenological regimes.

(i) *Coannihilation with conversions fast:* $\Gamma_{\rm conv} \gg H$ and $\langle\sigma v\rangle_{\psi\psi}$ large imply standard Griest-Seckel coannihilation [32], but with the twist that $\gamma_p^{(\rm SM)}(\chi) \lesssim H$; χ is chemically tied to the SM through ψ, yet kinetically SM-decoupled.

(ii) *Conversion-driven freeze-out (CDFO):* $\chi\chi$ annihilation is negligible and the relic density is set by $\chi \leftrightarrow \psi$ conversions shutting off while $\psi\psi$ annihilations remain efficient. Momentum-resolved studies show that χ can be nonthermal or thermal only within the $\{\chi,\psi\}$ subsystem at $T_d \neq T$; relic shifts at the $\mathcal{O}(10\%)$ level are common, larger when high-p modes dominate the conversions [33,34].

(iii) *Coscattering corridor:* when $\chi X \to \psi X'$ (with X thermal) controls the abundance after $\chi\chi$ and $\psi\psi$ annihilations decouple, low-p modes decouple first and f_χ becomes momentum-dependent; chemical coupling persists through (44) even as $\gamma_p^{(\rm SM)}(\chi) \lesssim H$ [35].

g. Representative realizations. Concrete examples include electroweak multiplet DM with a nearly degenerate partner where inelastic conversions proceed via weak interactions but direct χ-SM elastic scattering is mass- or angle-suppressed, and simplified models with a heavy scalar or vector mediator connecting χ/ψ to quarks or leptons. In these cases the parametrics (51) obtain, and the separation in (48) is realized over $x \sim 10$–30 during freeze-out, as explicitly demonstrated in momentum-resolved studies [33,34].

C. s-channel resonant DM annihilation with a suppressed elastic portal

a. Setup and kinematics. Let a Majorana/Dirac DM particle χ annihilate to SM states via an s-channel mediator ϕ (scalar or vector) with mass m_ϕ and width Γ_ϕ, and couplings g_χ (to χ) and $g_{\rm SM}$ (to SM). The annihilation kernel is Breit-Wigner enhanced near $s \simeq m_\phi^2$:

$$\sigma_{\rm ann}(s) v_{\rm rel} = \frac{\mathcal{N} g_\chi^2 g_{\rm SM}^2 \sqrt{1-4m_{\rm fin}^2/s}}{(s-m_\phi^2)^2 + m_\phi^2 \Gamma_\phi^2} \times \Phi(s), \qquad (52)$$

with $\Phi(s)$ a smooth phase-space/initial-state factor and $\mathcal{N}$ a spin-color constant. In the nonrelativistic regime $s \simeq 4m_\chi^2(1+\frac{v^2}{4})$, so proximity to the pole is controlled by

$$\delta \equiv \frac{m_\phi^2 - 4m_\chi^2}{4m_\chi^2}, \qquad v_{\rm res} \simeq 2\sqrt{\max(\delta,0)}, \qquad \Delta v \sim \frac{\Gamma_\phi}{m_\chi},$$

i.e., a *narrow band in velocity space* drives the largest inelastic rate. For $\delta > 0$ the resonance is reachable only from the high-velocity tail; for $\delta < 0$ the pole is kinematically inaccessible and annihilation proceeds off-shell but can still be sharply velocity dependent (and may receive Sommerfeld enhancement if ϕ mediates a long-range force).

b. Chemical vs. kinetic hierarchies. The annihilation (chemical) rate per particle is

$$\Gamma_{\rm chem} \equiv n_\chi \langle\sigma v\rangle_{\rm ann} \simeq n_\chi \int {\rm d}^3 v \sigma_{\rm ann}(v) v f_\chi(\vec{v}), \qquad (53)$$

which can be large when a significant fraction of f_χ overlaps the resonant band ($v \approx v_{\rm res}, \Delta v$). By contrast, *kinetic equilibrium with the SM* requires a large SM-elastic *transport* rate $\gamma_p^{(\rm SM)}$ set by t-channel exchanges off relativistic targets (e.g., $e^\pm, \nu, \pi$). In many simplified or portal models the elastic amplitude is forward-peaked (light mediator) or contact-suppressed by a heavy propagator M

$$\sigma_{\rm mt}^{(\rm SM)} \sim \frac{(g_\chi g_{\rm SM})^2}{(q^2+M^2)^2} \times \underbrace{\langle 1-\cos\theta\rangle}_{\ll 1 \text{ if forward-peaked}} \Rightarrow \gamma_p^{(\rm SM)}$$
$$\propto \frac{(g_\chi g_{\rm SM})^2 T^3}{M^4 m_\chi} \langle 1-\cos\theta\rangle. \qquad (54)$$

Hence one generically finds an epoch (typically $x \equiv m_\chi/T \sim 10$–$30$) with

$$\Gamma_{\rm chem} \gg H \gtrsim \gamma_p^{(\rm SM)}, \qquad (55)$$

i.e., chemical coupling is still active (because of the resonant band) while kinetic exchange with the SM fails.

c. Momentum-space picture and feedback. Because $\sigma_{\rm ann}(v)$ is sharply peaked, annihilations *deplete a narrow shell in momentum space* around $p_{\rm res} \simeq m_\chi v_{\rm res}$. If $\gamma_p^{(\rm SM)} \ll H$, elastic scattering on the SM cannot efficiently refill this shell; f_χ develops a "notch" or dip, and the system

departs from a single-temperature form. Two robust consequences follow:

(1) *Thermal-averaging bias.* The common kinetic-equilibrium estimate $\langle\sigma v\rangle_{\rm ann}^{\rm (KE)} = \langle\sigma v\rangle$ evaluated on a Maxwellian at $T_\chi = T$ (or at a fitted T_χ) *misestimates* the true average in (53) once the resonant band has been sculpted out of f_χ. This shifts the freeze-out condition and the relic density—by factors from $\mathcal{O}(1)$ up to an order of magnitude in sharp/narrow cases—relative to kinetic-equilibrium treatments [21,36,37].

(2) *Velocity-history memory.* The postfreeze-out f_χ retains a memory of the resonant depletion. Late-time observables that depend on v (indirect detection, CMB energy injection) can therefore differ from predictions made with a Maxwellian f_χ at the same number density.

d. Representative realizations.

(i) *Higgs-portal resonance:* $m_\chi \simeq m_h/2$ with annihilation via $h^* \to b\bar{b}, \tau^+\tau^-$. The chemical rate is resonantly enhanced near freeze-out, whereas SM-elastic scattering on relativistic leptons via t-channel h exchange is Yukawa-suppressed ($\propto y_\ell^2$) and contactlike ($\propto m_h^{-4}$), driving $\gamma_p^{\rm (SM)} \lesssim H$ even while $\Gamma_{\rm chem} \gg H$.

(ii) Z' *near threshold:* a narrow vector mediator with $m_{Z'} \simeq 2m_\chi$ and small kinetic mixing ϵ can produce large $\Gamma_{\rm chem}$ while $\chi - e^\pm$ elastic scattering is both ϵ^2- and transport-suppressed, realizing (55).

e. Sommerfeld and bound-state effects. If the mediator participates in the initial-state potential, Sommerfeld factors $S(v)$ can further enhance annihilation at small v (or generate resonant peaks when a near-threshold bound state forms). This *sharpens* the momentum selectivity and strengthens the need for momentum-resolved treatments; $S(v)$ multiplies the Breit-Wigner in (52) and accentuates the depletion of specific velocity shells. Kinetic failure with the SM then occurs even earlier because annihilations preferentially target the lowest-v modes while elastic refilling remains slow.

f. Coupled evolution: from full f to controlled closures. A careful treatment evolves the phase-space density $f_\chi(p,t)$ with the exact $2 \leftrightarrow 2$ annihilation kernel and an elastic operator appropriate to the scattering regime

$$(\partial_t - Hp\partial_p) f_\chi = C_{\rm ann}[f_\chi] + C_{\rm el}^{\rm (SM)}[f_\chi]. \tag{56}$$

In forward-peaked regimes a matched Fokker-Planck/Langevin operator for $C_{\rm el}^{\rm (SM)}$ (drag/diffusion chosen to reproduce momentum and heat exchange) yields reliable *transport* rates and captures the failure of refilling the resonant band. Moment closures (number + kinetic temperature T_χ) can work if $\sigma_{\rm ann}(v)$ is broad; they underperform when the resonance is narrow or when thresholds select the high-p tail [21,36,37].

g. Parametrics and scaling. For a heavy t-channel mediator of mass M,

$$\gamma_p^{\rm (SM)} \sim \frac{(g_\chi g_{\rm SM})^2}{M^4}\frac{T^3}{m_\chi} \times \langle 1 - \cos\theta\rangle,$$

$$\Gamma_{\rm chem} \sim n_\chi \frac{g_\chi^2 g_{\rm SM}^2}{(4m_\chi^2 - m_\phi^2)^2 + m_\phi^2\Gamma_\phi^2},$$

so that $\Gamma_{\rm chem}/\gamma_p^{\rm (SM)} \propto (n_\chi m_\chi/T^3)[\langle 1-\cos\theta\rangle]^{-1} \times [(4m_\chi^2 - m_\phi^2)^2 + m_\phi^2\Gamma_\phi^2]^{-1}$. At $x \sim 20$ this ratio is naturally large unless elastic scattering is gauge-strength and unsuppressed by transport weighting.

In sum, near-threshold s-channel annihilation furnishes a concrete regime where *chemical* coupling is strong—thanks to a narrow velocity band—while *kinetic* coupling to the SM bath is weak due to transport-suppressed elastic scattering. Momentum-resolved studies consistently find pronounced spectral features and sizable relic-density shifts relative to kinetic-equilibrium treatments [21,32,36,37].

D. Sufficient conditions for kinetic equilibrium *with the SM* for species in chemical equilibrium

Kinetic equilibrium with the SM *does* follow from chemical equilibrium if either:

(1) the elastic portal to SM gauge-charged targets is strong enough that $\gamma_p^{\rm (SM)}/H \gg 1$ throughout and after the chemically active epoch (typical for species with unsuppressed gauge interactions), or

(2) the inelastic channel that enforces chemistry *also* implies large-angle elastic scattering on SM targets with comparable rate and no transport suppression.

Absent these conditions, chemical equilibrium does not imply kinetic equilibrium with the SM.

VI. DOES CHEMICAL EQUILIBRIUM IMPLY KINETIC EQUILIBRIUM WITHIN A DARK SECTOR?

Claim: No. In a dark sector (DS) with its own temperature T_d, *chemical* equilibrium among dark species does not, in general, guarantee *kinetic* equilibrium (i.e., thermal velocity distributions) even solely within the DS. Chemical balance fixes chemical potentials through the reaction network, $\sum_j \nu_j \mu_j = 0$ for each active process $\sum_j \nu_j X_j \leftrightarrow 0$, provided the slowest number-changing rate $\Gamma_{\rm chem} \gg H$. By contrast, kinetic equilibrium requires sufficiently rapid *momentum exchange* inside the DS, quantified by transport rates

$$\gamma_p^{\rm (DS)}(i) \simeq \sum_{t\in{\rm DS}} n_t(T_d)\langle\sigma_{{\rm mt},it}^{\rm (DS)} v\rangle \frac{T_d}{m_i} \gg H. \tag{57}$$

Because chemical and transport operators scale differently with couplings, mediator masses and kinematics (e.g., s-channel decays/annihilations vs. forward-peaked t-channel elastic), one can have $\Gamma_{\rm chem} \gg H$ while $\gamma_p^{(\rm DS)} \lesssim H$ for part of the populated momentum range.

A. Kinetic rationale within the DS

Let $f_i(p,t)$ be the phase-space density for dark species i. Its evolution obeys

$$(\partial_t - Hp\partial_p) f_i = C_{\rm el}^{(\rm DS)}[f] + C_{\rm inel}^{(\rm DS)}[f] (+ C_{\rm portal}[f] \text{ if present}). \tag{58}$$

When $C_{\rm inel}^{(\rm DS)}$ is fast it enforces chemical relations (shared μ_i's or their linear constraints), but the *shape* of f_i thermalizes only if $C_{\rm el}^{(\rm DS)}$ provides rapid, transport-efficient momentum exchange across momenta carrying n and ρ. If $\gamma_p^{(\rm DS)} \lesssim H$ after some $t_{\rm kd}^{(\rm DS)}$, the subsequent spectrum is a redshifted memory of f_i at kinetic decoupling plus any later injections, and need not be Maxwell-Boltzmann/Fermi-Dirac at T_d.

B. Two-state DS with conversions but slow DS-elastic

a. Field content and reactions. Consider a secluded dark sector (DS) with temperature $T_d \neq T$, containing a DM state χ and a slightly heavier partner ψ with $\Delta \equiv (m_\psi - m_\chi)/m_\chi \ll 1$, plus light dark radiation φ that efficiently self-thermalizes and sets T_d. The key processes are

$$\text{(inelastic conversions)} \quad \chi \leftrightarrow \psi + \varphi, \quad \chi\varphi \leftrightarrow \psi\varphi', \tag{59}$$

$$\text{(dark elastic)} \quad \chi\varphi \leftrightarrow \chi\varphi, \quad \chi\psi \leftrightarrow \chi\psi, \quad \psi\varphi \leftrightarrow \psi\varphi, \tag{60}$$

$$\text{(number depletion)} \quad \psi\psi \leftrightarrow \text{light DS} \quad (\text{or SM}). \tag{61}$$

We assume φ is light enough (or gaugelike) that (60) keeps φ internally thermal with negligible chemical potential.

b. Chemical equilibrium at T_d vs. kinetic equilibrium within the DS. The inelastic network (59) enforces chemical relations

$$\mu_\chi \simeq \mu_\psi \Rightarrow \frac{n_\chi}{n_\psi} \simeq \frac{n_\chi^{(\rm eq)}(T_d)}{n_\psi^{(\rm eq)}(T_d)} \simeq e^{+\Delta x_d}, \qquad x_d \equiv \frac{m_\chi}{T_d}, \tag{62}$$

provided the slowest per-particle conversion rate satisfies

$$\Gamma_{\rm conv}^{(\rm DS)} \equiv \Gamma_{\psi\to\chi\varphi} + n_\varphi \langle \sigma v \rangle_{\chi\varphi\leftrightarrow\psi\varphi'} \gg H. \tag{63}$$

By contrast, *kinetic* equilibrium of χ *within the DS* requires a large *transport* rate from (60)

$$\gamma_p^{(\rm DS)}(\chi) \simeq \sum_{t=\{\varphi,\psi\}} n_t(T_d) \langle \sigma_{{\rm mt},\chi t}^{(\rm DS)} v \rangle \frac{T_d}{m_\chi} \gg H,$$
$$\sigma_{\rm mt} \equiv \int d\Omega (1-\cos\theta) \frac{d\sigma}{d\Omega}, \tag{64}$$

with the $(1-\cos\theta)$ weighting penalizing forward scattering. A broad and generic hierarchy then emerges

$$\Gamma_{\rm chem}^{(\rm DS)} \sim \Gamma_{\rm conv}^{(\rm DS)} \gg H \text{ while } \gamma_p^{(\rm DS)}(\chi) \lesssim H, \tag{65}$$

whenever elastic exchange is dominated by t-channel exchange of a heavier mediator or suffers additional coupling/angle suppressions.

c. Parametrics from a simple mediator model. Take conversions from a Yukawa-like interaction $y\psi\chi\varphi$ and DS elastic mediated by a heavy vector/scalar of mass M_d with coupling g_d. For $\Delta \ll 1$ (small Q-value two-body decay) one finds

$$\Gamma_{\psi\to\chi\varphi} \sim \frac{y^2}{16\pi} \Delta m_\chi, \qquad n_\varphi \langle \sigma v \rangle_{\chi\varphi\to\psi\varphi'} \sim y^2 T_d, \tag{66}$$

while the *transport* rate from χ-φ elastic scattering scales as

$$\gamma_p^{(\rm DS)}(\chi) \sim n_\varphi \frac{g_d^4 T_d^2}{M_d^4} \frac{T_d}{m_\chi} \times \langle 1-\cos\theta \rangle. \tag{67}$$

Hence

$$\frac{\Gamma_{\rm conv}^{(\rm DS)}}{\gamma_p^{(\rm DS)}(\chi)} \sim \left(\frac{y^2}{g_d^4}\right) \left(\frac{M_d^4}{T_d^2}\right) \left(\frac{m_\chi}{T_d}\right) \langle 1-\cos\theta \rangle^{-1},$$

which is generically $\gg 1$ for $m_\chi \gg T_d$ and forward-peaked elastic exchange (small $\langle 1-\cos\theta \rangle$), realizing (65) over the freeze-out window $x_d \sim 10$–30.

d. Momentum-space picture. Inverse decays $\chi\varphi \to \psi$ and decays $\psi \to \chi\varphi$ reshuffle *identities* and inject quanta at characteristic momenta set by the Q-value ($\sim \Delta m_\chi$). When $\gamma_p^{(\rm DS)}(\chi) \lesssim H$, elastic refilling is inefficient and $f_\chi(p)$ develops nonthermal features: (i) a hard tail from decay-produced χ, (ii) momentum-selective dips/excesses from the $2 \leftrightarrow 2$ conversion kernel. Even while (62) holds at the level of number ratios, the velocity distribution of χ need not be Maxwell-Boltzmann at T_d.

e. Coupled evolution (number, heat, and, when needed, full f). A minimal yet informative closure supplements the number densities with a separate kinetic temperature for the heavy sector, T_χ, in addition to the DS bath temperature T_d (carried by φ)

$$\dot{n}_\chi + 3Hn_\chi = -\mathcal{C}_{\rm dep}[n_\chi, n_\psi; T_d] + \mathcal{C}_{\rm conv}[n_\chi, n_\psi; T_d], \tag{68}$$

$$\frac{3}{2}(\dot{n}_\chi T_\chi + n_\chi \dot{T}_\chi) + 5Hn_\chi T_\chi = -\underbrace{\mathcal{Q}_{\rm el}}_{\propto \gamma_p^{(\rm DS)}(\chi)(T_\chi - T_d)} + \underbrace{\mathcal{Q}_{\rm conv}}_{\text{energy injection/extraction from conversions}} + \ldots \tag{69}$$

Here $\mathcal{C}_{\rm dep}$ lumps number-depleting reactions (often dominated by $\psi\psi \to$ light DS/SM), while $\mathcal{C}_{\rm conv}$ enforces (62) when (63) holds. The elastic heat-exchange term $\mathcal{Q}_{\rm el}$ drives $T_\chi \to T_d$ with rate $\sim\gamma_p^{(\rm DS)}$, whereas $\mathcal{Q}_{\rm conv}$ accounts for the nonthermal energy transfer carried by decay/ inverse-decay quanta. This two-temperature (cBE) system is accurate if f_χ remains close to Maxwellian; otherwise a momentum-resolved treatment,

$$(\partial_t - Hp\partial_p) f_\chi = C_{\rm conv}[f_\chi, f_\psi, f_\varphi] + C_{\rm el}^{(\rm DS)}[f_\chi, f_\varphi],$$

is required. In small-angle regimes $C_{\rm el}^{(\rm DS)}$ can be implemented as a matched Fokker-Planck/Langevin operator that reproduces the exact momentum/heat exchange and yields $\gamma_p^{(\rm DS)}$ directly.

f. Regimes and outcomes.

(i) *Nonthermal* χ: if both χ-φ and χ-ψ elastic are slow, (69) gives $T_\chi/T_d \to$ constant $\gtrsim 1$ (or a slowly varying ratio) and f_χ develops persistent high-p features set by conversion kinematics. The relic density can shift relative to a single-T_d closure when the chemistry is sensitive to high-p modes.

(ii) *Thermalized-but-secluded pair* $\{\psi, \varphi\}$*:* if ψ-φ elastic is fast while χ-φ is slow, the bath and ψ share T_d and Maxwellian shapes; χ tracks (62) number-wise but remains kinetically misaligned (nonthermal or at a distinct T_χ).

(iii) *Recovery of DS kinetic equilibrium:* if at later times a lighter mediator turns on (phase transition) or the forward peak is screened (finite-T_d mass), $\gamma_p^{(\rm DS)}$ can exceed H again, partially thermalizing χ at late x_d; imprints of earlier nonthermality may nevertheless survive in the low-v tail.

g. Representative realizations. (i) A dark-$U(1)$ with heavy dark photon A' mediating elastic scattering (g_d) and a Yukawa $y\psi\chi\varphi$ driving conversions; (ii) Z_2-odd doublet of fermions with a light pseudoscalar φ enabling $\psi \to \chi\varphi$ while elastic χ-φ proceeds via heavy scalar exchange; (iii) multi-TeV χ with $\Delta \sim 0.05$ and $M_d \gg m_\chi$, realizing (65) for $x_d \sim 10$–30.

Note that this DS example is the secluded analog of conversion/coannihilation and coscattering dynamics discussed for SM-coupled sectors (see Sec. V A); the kinetic diagnostics (transport-weighted rates, FP/Langevin) follow the methodology developed for early kinetic decoupling and momentum-resolved freeze-out [10,31]. Dark-sector temperature evolution and cannibal phases provide additional structure when $3 \to 2$ is active [23,38]; conversion-driven freeze-out and related multistate dynamics are discussed in [33,34].

C. Semiannihilation, Z_3-like DS: Chemistry without full kineticization

a. Field content and reactions. Consider a secluded dark sector (DS) with temperature $T_d \neq T$, containing a stable DM particle χ (stabilized by a Z_3-like symmetry), a light dark state ϕ that sets T_d via fast self-interactions, and semiannihilation (SA)

$$\chi\chi \leftrightarrow \chi\phi, \tag{70}$$

supplemented by rapid ϕ-self-thermalization (e.g., $\phi\phi \leftrightarrow \phi\phi$ and, optionally, $\phi\phi \leftrightarrow \phi\phi\phi$ or $\phi \to$ light DS) and *suppressed* elastic $\chi - \phi$ scattering (heavy t-channel mediator or forward-peaked kinematics). Equation (70) changes the χ number by -1 per reaction and links the chemical potentials,

$$2\mu_\chi = \mu_\chi + \mu_\phi \Rightarrow \mu_\chi = \mu_\phi, \tag{71}$$

so that, in the presence of fast SA and any ϕ-number-violating processes (e.g., ϕ decay or $3 \leftrightarrow 2$ in the ϕ-sector), one obtains $\mu_\phi \simeq 0$ and thus $\mu_\chi \simeq 0$. If ϕ only undergoes $2 \leftrightarrow 2$ self-scattering, its number is conserved and the common value $\mu_\chi = \mu_\phi$ can be nonzero; in either case, *chemical* balance is enforced at the DS temperature T_d when the slowest number-changing rate satisfies $\Gamma_{\rm chem}^{(\rm DS)} \gg H$.

b. Rate hierarchy and transport. Define the SA depletion rate per χ as

$$\Gamma_{\rm SA} \equiv n_\chi \langle\sigma v\rangle_{\chi\chi\to\chi\phi}(T_d, T_\chi), \qquad \Gamma_{\rm chem}^{(\rm DS)} \sim \Gamma_{\rm SA}, \tag{72}$$

where the thermal average may depend on both T_d (for ϕ) and the *actual* χ spectrum (see below). Kinetic equilibrium *within the DS* requires a large *transport* rate for momentum exchange between χ and the bath,

$$\gamma_p^{(\rm DS)}(\chi) \simeq n_\phi(T_d)\langle\sigma_{\rm mt}(\chi\phi \to \chi\phi)v\rangle \frac{T_d}{m_\chi}, \qquad \sigma_{\rm mt} \equiv \int d\Omega (1 - \cos\theta)\frac{d\sigma}{d\Omega}. \tag{73}$$

With heavy t-channel mediators or light mediators producing forward peaking, $\sigma_{\rm mt} \ll \sigma_{\rm el}$ and thus $\gamma_p^{(\rm DS)}(\chi) \lesssim H$ over $x_d \equiv m_\chi/T_d \sim 10$–$30$ even when $\Gamma_{\rm SA} \gg H$. Therefore one naturally realizes

$$\Gamma_{\rm chem}^{(\rm DS)} \gg H \quad \text{while} \quad \gamma_p^{(\rm DS)}(\chi) \lesssim H, \tag{74}$$

i.e., chemical equilibrium in the DS without full kineticization of χ at T_d.

c. Momentum-space picture: injection vs. refilling. Semiannihilation is *momentum-selective*. In the nonrelativistic regime, two nearly at-rest χ's produce a final-state χ and a (typically relativistic) ϕ. The outgoing χ receives a kinetic kick set by the Q-value

$$Q \simeq m_\chi - m_\phi + \mathcal{O}(T_d) \approx m_\chi \quad (\text{if } m_\phi \ll m_\chi),$$

so SA *injects* a population of hard χ at $p \sim \sqrt{2m_\chi Q}$, while inverse SA, $\chi\phi \to \chi\chi$, repopulates specific momentum bands governed by the ϕ bath. When $\gamma_p^{(\mathrm{DS})}(\chi) \lesssim H$, elastic $\chi - \phi$ scattering cannot efficiently isotropize/thermalize these injections. The result is a nonthermal $f_\chi(p)$ featuring (i) a high-p tail from $\chi\chi \to \chi\phi$ and (ii) momentum-band features from $\chi\phi \to \chi\chi$; both features simply redshift once DS kinetic decoupling is complete.

d. Energetics and dark self-heating. Each SA event converts approximately one rest mass m_χ into kinetic energy shared by χ and ϕ. In a two-temperature closure the heavy component obeys

$$\frac{3}{2}(\dot{n}_\chi T_\chi + n_\chi \dot{T}_\chi) + 5Hn_\chi T_\chi = -\underbrace{\mathcal{Q}_{\mathrm{el}}}_{\propto \gamma_p^{(\mathrm{DS})}(\chi)(T_\chi - T_d)} + \underbrace{\mathcal{Q}_{\mathrm{SA}}}_{\simeq \frac{1}{2} m_\chi \Gamma_{\mathrm{SA}} n_\chi} + \ldots \tag{75}$$

where $\mathcal{Q}_{\mathrm{el}}$ thermalizes $T_\chi \to T_d$ at a rate set by (73), while $\mathcal{Q}_{\mathrm{SA}} > 0$ *heats* the heavy component (the $\frac{1}{2}$ reflects that one χ is removed per reaction). If $\gamma_p^{(\mathrm{DS})} \ll \Gamma_{\mathrm{SA}}$ during freeze-out, semiannihilation drives T_χ/T_d above unity ("partial cannibalization by SA") and accentuates nonthermal high-p features; the bath ϕ also receives heat, but if $g_{*,d}$ is large its temperature responds only mildly.

e. Coupled number evolution with detailed balance. The number density evolves as

$$\dot{n}_\chi + 3Hn_\chi = -\langle\sigma v\rangle_{\chi\chi\to\chi\phi}[f_\chi, f_\phi] n_\chi^2 + \langle\sigma v\rangle_{\chi\phi\to\chi\chi}[f_\chi, f_\phi] n_\chi n_\phi, \tag{76}$$

with *distinct* thermal averages for forward and inverse processes once f_χ is nonthermal or $T_\chi \neq T_d$. Detailed balance at the microphysical level still relates the *kernels*, but the phase-space-averaged rates differ from those computed assuming a Maxwellian at T_d. In practice:

(i) A single-temperature closure ($T_\chi = T_d$) *over*-mixes momenta and misestimates the forward/inverse balance when injection shapes are important.

(ii) A two-temperature (cBE) system $(n_\chi, T_\chi; T_d)$ captures leading effects if f_χ remains nearly Maxwellian; otherwise a momentum-resolved treatment is required.

f. Parametric example. Take a Z_3-like interaction $\lambda\chi^3\phi + \mathrm{H.c.}$ generating SA at

$$\langle\sigma v\rangle_{\mathrm{SA}} \sim \frac{|\lambda|^2}{16\pi m_\chi^2},$$

and elastic $\chi - \phi$ scattering mediated by a heavy state of mass M_d with coupling g_d

$$\gamma_p^{(\mathrm{DS})}(\chi) \sim n_\phi \frac{g_d^4 T_d^2}{M_d^4} \frac{T_d}{m_\chi} \times \langle 1 - \cos\theta\rangle.$$

For $m_\chi \gg T_d$ and forward peaking, $\Gamma_{\mathrm{SA}}/\gamma_p^{(\mathrm{DS})} \propto (m_\chi/T_d)(M_d^4/g_d^4)\langle 1 - \cos\theta\rangle^{-1} \gg 1$, realizing (74) throughout freeze-out.

g. Regimes and outcomes.

(i) *Chemistry fast, DS-elastic slow:* n_χ tracks its chemical trajectory at T_d [set by (71)] while f_χ exhibits nonthermal high-p shoulders and $T_\chi > T_d$. Relic-density predictions using $T_\chi = T_d$ can err at the $\mathcal{O}(10\%)$–$\mathcal{O}(1)$ level when SA is sharply momentum-selective.

(ii) *Late partial rekineticization:* If $\gamma_p^{(\mathrm{DS})}$ grows relative to H (e.g., screening of the forward peak, mediator becoming lighter, or n_ϕ rising during a DS phase transition), the hard tail is partially erased at $x_d \gtrsim x_{\mathrm{fo}}$; observables sensitive to the low-v tail (indirect/CMB) can still retain memory of early nonthermality.

Semiannihilation provides a clean intra-DS realization of "chemistry without kineticization": number-changing is efficient, but DS-elastic transport can be parametrically weaker. The energetics mimic cannibal self-heating to be discussed below (Sec. VI C) but proceed through $2 \leftrightarrow 2$ reactions that *inject* hard χ rather than preferentially depleting low-p modes. For model scaffolding and phenomenology of semiannihilation in Z_3-like sectors, see, e.g., [22]; kinetic diagnostics and transport-weighted treatments follow [10,31].

D. Cannibal sector caveat: $3 \to 2$ chemistry with marginal $2 \to 2$

a. Field content and standard assumption. Consider a secluded dark sector (DS) at temperature $T_d \neq T$ with a single nonrelativistic species χ undergoing efficient number-changing

$$\chi\chi\chi \leftrightarrow \chi\chi, \tag{77}$$

and (possibly) elastic self-scattering $\chi\chi \leftrightarrow \chi\chi$. In the usual *cannibal* regime one assumes fast $2 \to 2$ elastic so that f_χ remains Maxwell-Boltzmann at a common T_d, while (77) maintains chemical equilibrium and heats the sector,

yielding the familiar $T_d \propto 1/\ln a$ during the nonrelativistic phase [23,38].

b. Chemical vs. kinetic hierarchy inside the DS. Let the slowest number-changing per-particle rate be

$$\Gamma_{3\to2}^{(\mathrm{DS})} \equiv n_\chi^2 \langle \sigma v^2 \rangle_{3\to2}(T_d), \tag{78}$$

which can readily satisfy $\Gamma_{3\to2}^{(\mathrm{DS})} \gg H$ around freeze-out. Kinetic equilibrium *within the DS* requires a large *transport* rate from elastic self-scattering,

$$\gamma_p^{(\mathrm{DS})}(\chi) \simeq n_\chi \langle \sigma_{\mathrm{mt}}(\chi\chi \to \chi\chi) v \rangle \frac{T_d}{m_\chi},$$
$$\sigma_{\mathrm{mt}} \equiv \int d\Omega (1-\cos\theta) \frac{d\sigma}{d\Omega}, \tag{79}$$

which is suppressed if the dominant $2 \to 2$ topology is forward-peaked t-channel or proceeds via a heavy mediator. When

$$\Gamma_{3\to2}^{(\mathrm{DS})} \gg H \quad \text{but} \quad \gamma_p^{(\mathrm{DS})}(\chi) \lesssim H, \tag{80}$$

the DS maintains chemical relations and number depletion ("chemistry") while failing to fully isotropize/thermalize momenta ("kinetics") on a Hubble time.

c. Momentum-space picture. In the nonrelativistic regime, the $3 \to 2$ process is exothermic: two survivors share the rest-mass energy of the third. If elastic transport is slow, repeated $3 \to 2$ events deplete low-p quanta and *inject* kinetic energy into the surviving pair, pushing weight toward intermediate/high p. Absent rapid $\chi\chi \leftrightarrow \chi\chi$ to redistribute momenta,

$$(\partial_t - Hp\partial_p) f_\chi = C_{3\leftrightarrow2}[f_\chi] + C_{\mathrm{el}}^{(\mathrm{DS})}[f_\chi],$$
$$\gamma_p^{(\mathrm{DS})} \ll H \Rightarrow f_\chi(p,t) \quad \text{nonthermal}, \tag{81}$$

with characteristic signatures: (i) selective depletion at low p (where $3 \to 2$ is most phase-space efficient in the NR limit), (ii) a self-heated shoulder at higher p from the exothermic recoil. These features then *redshift* after intra-DS kinetic decoupling.

d. Energetics and temperature evolution beyond the textbook limit. Write the number and heat equations (integrated Boltzmann hierarchy) for a heavy NR component

$$\dot{n}_\chi + 3Hn_\chi = -\langle \sigma v^2 \rangle_{3\to2}(n_\chi^3 - n_\chi^2 n_{\chi,\mathrm{eq}}(T_d)), \tag{82}$$

$$\frac{3}{2}(\dot{n}_\chi T_\chi + n_\chi \dot{T}_\chi) + 5Hn_\chi T_\chi$$
$$= -\underbrace{\mathcal{Q}_{\mathrm{el}}}_{\propto \gamma_p^{(\mathrm{DS})}(\chi)(T_\chi - T_d)} + \underbrace{\mathcal{Q}_{3\to2}}_{\text{self-heating from } 3\to2} + \ldots \tag{83}$$

The standard cannibal result $T_\chi = T_d \propto 1/\ln a$ follows when $\mathcal{Q}_{\mathrm{el}}$ enforces $T_\chi = T_d$ at all times. When $\gamma_p^{(\mathrm{DS})} \lesssim H$, however, $\mathcal{Q}_{\mathrm{el}}$ cannot enforce $T_\chi \to T_d$ and (83) yields $T_\chi/T_d > 1$ with a slower-than-logarithmic cooling of the *heavy* component, reflecting incomplete kineticization of the self-heating. This modifies both the instantaneous $\langle \sigma v^2 \rangle_{3\to2}$ (through its temperature/momentum dependence) and the mapping between comoving entropy conservation in the DS and the $T_d(a)$ trajectory familiar from [23,38].

c. Parametrics and when the split occurs. Assume $3 \to 2$ arises from a contact operator with strength α_d/m_χ^5 (NR scaling), while elastic proceeds via a heavy mediator of mass M_d and coupling g_d

$$\Gamma_{3\to2}^{(\mathrm{DS})} \sim \alpha_d \frac{n_\chi^2}{m_\chi^5}, \qquad \gamma_p^{(\mathrm{DS})}(\chi) \sim n_\chi \frac{g_d^4 T_d^2}{M_d^4} \frac{T_d}{m_\chi} \times \langle 1-\cos\theta \rangle. \tag{84}$$

At $x_d \equiv m_\chi/T_d \simeq 10$–$30$ one then finds

$$\frac{\Gamma_{3\to2}^{(\mathrm{DS})}}{\gamma_p^{(\mathrm{DS})}(\chi)} \propto \left(\frac{\alpha_d}{g_d^4}\right)\left(\frac{M_d^4}{m_\chi^4}\right)\left(\frac{m_\chi}{T_d}\right)\langle 1-\cos\theta\rangle^{-1} \gg 1$$

for modest couplings and forward-peaked elastic, realizing (80) over the epoch where $3 \to 2$ controls the chemistry.

f. When $3 \to 2$ can self-thermalize (and when not). A broad $3 \leftrightarrow 2$ kernel with sizable momentum transfer *can* contribute to kineticization; in some UV completions the same diagrams that deplete number also exchange momentum efficiently across p. However, near kinematic thresholds, for narrow resonances, or if the $3 \to 2$ matrix element is dominated by configurations with small momentum transfer, the induced momentum exchange per event is insufficient to maintain a Maxwellian shape on a Hubble time without the help of $2 \to 2$ elastic. These corners are precisely where a momentum-resolved check is warranted rather than assuming a single T_d.

E. Remarks on sufficient conditions

Chemical *and* kinetic equilibrium within the DS is ensured if at least one of the following holds:

(1) Some DS elastic channel with light target (e.g., dark photon, light scalar) yields $\gamma_p^{(\mathrm{DS})} \gg H$ throughout the chemically active epoch for all heavy species.
(2) Number-changing processes themselves (e.g., broad-s $2 \leftrightarrow 2$ or sufficiently frequent $3 \leftrightarrow 2$ with wide kinematics) also furnish efficient momentum exchange across populated p without strong forward peaking.

Absent these, DS chemistry can be maintained (shared μ_i at T_d) while DS kinetics fails (nonthermal f_i), even when the DS is fully secluded from the SM.

Note that dark freeze-out in a self-thermalized sector (with the *assumption* of fast DS elastic) is developed in [23,38]. Momentum-exchange diagnostics via transport cross sections and FP/Langevin operators—directly applicable to DS kinetics—are discussed in [10,31]. Semiannihilation frameworks providing chemistry with potentially weak DS elastic appear in Z_3-like models (e.g., [22]); DS conversion/coannihilation analogs of Sec. VI A are analyzed in conversion-driven freeze-out contexts [33,34].

VII. DISCUSSION AND CONCLUSIONS

This paper has separated, both conceptually and operationally, two notions that are often conflated in early Universe dynamics: chemical equilibrium, which governs number densities through the slowest number-changing process, and kinetic equilibrium, which governs the shape of the momentum distribution through transport. The clean way to keep them apart is to track the rate that controls chemistry, $\Gamma_{\rm chem}$, alongside the transport-weighted momentum-exchange rate, γ_p, and to compare each to the Hubble rate over the momenta that actually carry number and energy. Because inelastic and elastic operators scale differently with couplings, propagators, angular structure, and thresholds, there is no generic implication in either direction. It is common to encounter parameter regions where $\Gamma_{\rm chem} \gg H$ while $\gamma_p \lesssim H$, or the converse, and it is equally common that the two ratios cross unity at different times so that chemical and kinetic decoupling proceed in different orders.

The Standard Model neutrino sector provides a precise illustration of these ideas. Around the MeV epoch, pair processes that control the neutrino chemical potentials thin first, while elastic scattering on $e^{\pm}$ continues to exchange momentum and energy into the era of $e^{\pm}$ annihilation. Momentum-resolved transport then predicts small, irreducible spectral distortions—largest at high comoving momenta—that cannot be absorbed into any single choice of (T_ν, μ_ν). The effect is minute but robust: it underlies the SM prediction $N_{\rm eff} \simeq 3.044 - 3.045$ and is a reminder that a species may have been fully equilibrated in the past and yet not retain an exactly thermal relic spectrum.

Heavy right-handed neutrinos in strong-washout leptogenesis illustrate the inverse separation. Inverse decays proportional to $|y_N|^2$ pin the abundance close to equilibrium at $T \sim M_N$, but the elastic processes that would efficiently thermalize momenta are $|y_N|^4$ and typically forward-peaked, so their transport rate is parametrically small. As a result, the momentum distribution can remain only partially relaxed even while the number density is chemically equilibrated. Any residual nonthermal structure simply redshifts after inelastic processes thin, and although strong washout remains intact, the detailed timing of washout and the efficiency factor can shift at the level relevant for precision studies.

Dark matter scenarios make the nontrivial ordering of decouplings especially transparent. When annihilation is concentrated by kinematics or resonances—near a narrow s-channel pole or into forbidden final states—chemical processing can remain efficient in a narrow band of velocities or in the high-velocity tail, even as elastic scattering on relativistic Standard Model targets is transport-suppressed by heavy propagators or forward peaking. In such regions one literally sees the kinetic decoupling occur first: elastic refilling fails, momentum-space "notches" open near the favored annihilation velocities, or the high-p tail erodes, and only later does chemical freeze-out complete once the sculpted distribution can no longer sustain $\Gamma_{\rm chem} \gtrsim H$. The relic abundance inferred assuming kinetic equilibrium can then differ by factors of order unity, and in extreme cases by an order of magnitude, from the result obtained using the actual phase-space distribution.

Closely related hierarchies appear in coannihilation and conversion-driven settings, and in coscattering corridors. Conversions enforce chemical relations between a dark matter state and a slightly heavier partner, tying chemical potentials together and maintaining the equilibrium number ratio while the final depletion proceeds through the partner's annihilations. If the direct elastic portal of the dark matter to the Standard Model is weak, the species is chemically tied to the bath through its partner but kinetically decoupled from it. In coscattering, an endothermic upscattering $\chi X \to \psi X'$ controls the abundance after annihilations thin; low-momentum modes fall below threshold and decouple first, and the depletion proceeds by nibbling away at the high-momentum tail. Here, too, kinetic decoupling precedes chemical decoupling in a way that is visible directly in the evolving $f(p)$.

All of these phenomena have analogs within secluded dark sectors at a temperature $T_d \neq T$. Chemical equilibration among dark species can be maintained by conversions, semiannihilation, or $3 \to 2$ reactions, while intrasector elastic scattering is too weak to enforce Maxwellian shapes. In such cases the sector has a well-defined chemical network at T_d but not a common kinetic temperature, and the heavy component retains nonthermal features—self-heated shoulders from $3 \to 2$ or injection shoulders from semiannihilation—even as the chemical constraints are satisfied. The textbook cannibal result $T_d \propto 1/\ln a$ relies on fast elastic self-scattering; when transport is marginal, the heavy component cools more slowly than the bath and a single-temperature closure is no longer reliable.

These examples motivate a simple methodological lesson. The reliable analysis is to compute $\Gamma_{\rm chem}(T)$ for the process that actually sets the abundance and to compute $\gamma_p(T)$ to the relevant bath using transport cross sections or a properly matched Fokker-Planck/Langevin description that reproduces both momentum and heat exchange.

One should plot $\Gamma_{\rm chem}/H$ and γ_p/H on the same axis through the epoch of interest and, whenever possible, resolve the momentum dependence of γ_p/H, because the highest-momentum modes typically fail first. The solver should match the physics: full collision integrals in momentum space are warranted when chemistry is momentum-selective—resonances, thresholds, coscattering—or when precision neutrino transport is at stake; Fokker-Planck/Langevin captures small-angle elastic dynamics efficiently and provides clean transport diagnostics; moment closures that evolve number density and a kinetic temperature are acceptable only while the spectrum remains close to Maxwellian. It is good practice to expose spectral evidence with snapshots of $p^2 f$ or $f/f_{\rm MB}$ and to compare thermal averages taken over the evolving distribution to their kinetic-equilibrium surrogates, because any bias there feeds directly into the inferred relic abundance or radiation density. In two-bath cosmologies one should also report the evolution of T_d/T, and, when intrasector elastic is marginal, species-specific kinetic temperatures relative to T_d.

The phenomenological consequences of getting this right are concrete. In resonant and forbidden windows, early kinetic decoupling changes the relic density by order-one factors relative to equilibrium estimates, shifting the couplings that reproduce today's abundance. In coannihilation and conversion-driven settings, the differences are often at the few-to-ten percent level but can become larger near thresholds. In the radiation sector, momentum-resolved neutrino transport underlies the small but precise shift in $N_{\rm eff}$ and fixes the baseline against which late injections or low reheating scenarios must be tested. Nonthermal late-time velocity distributions alter indirect-detection rates for velocity-dependent operators, including Sommerfeld-enhanced channels, and change the mapping to CMB energy-injection constraints. If kinetic decoupling occurs while a species is semirelativistic, broadened or anisotropic distributions modify free-streaming scales and small-scale structure relative to thermal templates; in dark sectors with $T_d \neq T$, the combination of seclusion, self-heating, and nonthermal shapes imprints characteristic cutoffs.

The central lesson is straightforward: chemical equilibrium describes numerical abundances, while kinetic equilibrium captures distributional forms, and in our expanding Universe, the processes governing each typically decouple at different epochs. By maintaining both diagnostics in parallel—and invoking phase-space evolution when momentum-dependent effects become significant-one establishes the most reliable pathway to robust, testable cosmological predictions.

ACKNOWLEDGMENTS

This work was inspired by a (correct) remark by Joshua Ruderman on the occasion of his lectures at the pre-SUSY2025 summer school; I acknowledge feedback from Joshua Ruderman, Pierce Giffin, and Grant Roberts. S. P. is partly supported by the U.S. Department of Energy Grant No. de-sc0010107.

DATA AVAILABILITY

No data were created or analyzed in this study.

[1] P. J. E. Peebles, *Principles of Physical Cosmology* (Princeton University Press, Princeton, NJ, 1993).
[2] S. Weinberg, *Cosmology* (Oxford University Press, Oxford, 2008).
[3] E. W. Kolb and M. S. Turner, *The Early Universe*, Frontiers in Physics (Addison-Wesley, Reading, MA, 1990), Vol. 69.
[4] F. H. Shu, *The Physics of Astrophysics. Volume 1: Radiation* (University Science Books, Mill Valley, CA, 1991).
[5] J. Bernstein, *Kinetic Theory in the Expanding Universe*, Cambridge Monographs on Mathematical Physics (Cambridge University Press, Cambridge, England, 1988).
[6] S. Dodelson, *Modern Cosmology* (Academic Press, San Diego, CA, 2003).
[7] R. K. Pathria and P. D. Beale, *Statistical Mechanics*, 3rd ed. (Academic Press, Boston, MA, 2011).
[8] P. Gondolo and G. Gelmini, Cosmic abundances of stable particles: Improved analysis, Nucl. Phys. **B360**, 145 (1991).
[9] A. D. Dolgov, Neutrinos in cosmology, Phys. Rep. **370**, 333 (2002).
[10] T. Bringmann and S. Hofmann, Thermal decoupling of wimps from first principles, J. Cosmol. Astropart. Phys. 04 (2007) 016; J. Cosmol. Astropart. Phys. 03 (2016) E02(E).
[11] C. Boehm, A. Riazuelo, S. H. Hansen, and R. Schaeffer, Interacting dark matter disguised as warm dark matter, Phys. Rev. D **66**, 083505 (2002).
[12] Y. Du, F. Huang, H.-L. Li, Y.-Z. Li, and J.-H. Yu, Revisiting dark matter freeze-in and freeze-out through phase-space distribution, J. Cosmol. Astropart. Phys. 04 (2022) 012.
[13] S. Hofmann, D. J. Schwarz, and H. Stoecker, Damping scales of neutralino cold dark matter, Phys. Rev. D **64**, 083507 (2001).
[14] M. Vogelsberger, J. Zavala, F.-Y. Cyr-Racine, C. Pfrommer, T. Bringmann, and K. Sigurdson, Ethos—An effective theory of structure formation: Dark matter physics as a possible explanation of the small-scale cdm problems, Mon. Not. R. Astron. Soc. **460**, 1399 (2016).

[15] T. R. Slatyer and C.-L. Wu, General constraints on dark matter decay from the cosmic microwave background, Phys. Rev. D **98**, 023013 (2018).
[16] T. Binder, T. Bringmann, M. Gustafsson, and A. Hryczuk, Dark matter relic abundance beyond kinetic equilibrium, Eur. Phys. J. C **81**, 577 (2021).
[17] P. Bode, J. P. Ostriker, and N. Turok, Halo formation in warm dark matter models, Astrophys. J. **556**, 93 (2001).
[18] X. Chu, T. Hambye, and M. H. G. Tytgat, Thermal relic dark matter in scenarios with large annihilation cross sections, J. Cosmol. Astropart. Phys. 05 (2012) 034.
[19] J. L. Feng, M. Kaplinghat, H. Tu, and H.-B. Yu, Hidden charged dark matter, J. Cosmol. Astropart. Phys. 07 (2009) 004.
[20] A. Hryczuk and M. Laletin, Impact of dark matter self-scattering on its relic abundance, Phys. Rev. D **106**, 023007 (2022).
[21] T. Binder, T. Bringmann, M. Gustafsson, and A. Hryczuk, Early kinetic decoupling of dark matter: When the standard way of calculating the thermal relic density fails, Phys. Rev. D **96**, 115010 (2017).
[22] F. D'Eramo and J. Thaler, Semi-annihilation of dark matter, J. High Energy Phys. 06 (2010) 109.
[23] D. Pappadopulo, J. T. Ruderman, and G. Trevisan, Dark matter freeze-out in a nonrelativistic sector, Phys. Rev. D **94**, 035005 (2016).
[24] S. Chatterjee and A. Hryczuk, Conversions in two-component dark sectors: A phase space level analysis, J. High Energy Phys. 07 (2025) 279.
[25] G. Mangano, G. Miele, S. Pastor, T. Pinto, O. Pisanti, and P. D. Serpico, Relic neutrino decoupling including flavor oscillations, Nucl. Phys. **B729**, 221 (2005).
[26] M. Escudero, Neutrino decoupling beyond the standard model: Cmb constraints on the dark matter mass with a fast and precise $n_{\rm eff}$ evaluation, J. Cosmol. Astropart. Phys. 02 (2019) 007.
[27] K. Akita and M. Yamaguchi, A precision calculation of relic neutrino decoupling, J. Cosmol. Astropart. Phys. 08 (2020) 012.
[28] M. Cielo, M. Escudero, G. Mangano, and O. Pisanti, Neff in the standard model at NLO is 3.043, Phys. Rev. D **108**, L121301 (2023).
[29] G. F. Giudice, A. Notari, M. Raidal, A. Riotto, and A. Strumia, Towards a complete theory of thermal leptogenesis in the SM and MSSM, Nucl. Phys. **B685**, 89 (2004).
[30] M. Beneke, B. Garbrecht, M. Herranen, and P. Schwaller, Finite number density corrections to leptogenesis, Nucl. Phys. **B838**, 1 (2010).
[31] Y. Ali-Haïmoud, Boltzmann-Fokker-Planck formalism for dark-matter–baryon scattering, Phys. Rev. D **99**, 023523 (2019).
[32] K. Griest and D. Seckel, Three exceptions in the calculation of relic abundances, Phys. Rev. D **43**, 3191 (1991).
[33] M. Garny, J. Heisig, B. Lülf, and S. Vogl, Coannihilation without chemical equilibrium, Phys. Rev. D **96**, 103521 (2017).
[34] M. Garny, J. Heisig, M. Hufnagel, B. Lülf, and S. Vogl, Conversion-driven freeze-out: Dark matter genesis beyond the WIMP paradigm, Proc. Sci., CORFU2018 (**2019**) 092 [arXiv:1904.00238].
[35] R. T. D'Agnolo, D. Pappadopulo, and J. T. Ruderman, Fourth exception in the calculation of relic abundances, Phys. Rev. Lett. **119**, 061102 (2017).
[36] K. Ala-Mattinen, M. Heikinheimo, K. Kainulainen, and K. Tuominen, Momentum distributions of cosmic relics: Improved analysis, Phys. Rev. D **105**, 123005 (2022).
[37] A. Aboubrahim, M. Klasen, and L. P. Wiggering, Forbidden dark matter annihilation into leptons with full collision terms, J. Cosmol. Astropart. Phys. 08 (2023) 075.
[38] M. Farina, D. Pappadopulo, J. T. Ruderman, and G. Trevisan, Phases of cannibal dark matter, J. High Energy Phys. 12 (2016) 039.