TAL TECH

DOCTORAL THESIS

Black Holes and Dark Matter in Light of New Infrared and Gravitational Wave Observations

Juan Urrutia Pérez

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2026

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
65/2026

Black Holes and Dark Matter in Light of New Infrared and Gravitational Wave Observations

JUAN URRUTIA PÉREZ

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Science
Department of Cybernetics
This dissertation was accepted for the defense of the degree of Doctor of Philosophy (Applied Physics and Mathematics) on 09 September 2026

Supervisor: Title. Dr. Ville Vaskonen
Laboratory of High Energy and Computational Physics,
National Institute of Chemical Physics and Biophysics (KBFI)

Co-supervisor: Title. Dr. Martti Raidal
Laboratory of High Energy and Computational Physics,
National Institute of Chemical Physics and Biophysics (KBFI)

Opponents: Prof Name: Dr. Diego Blas
Institut de Física d'Altes Energies,
Autonomous University of Barcelona (UAB)

Prof Name: Dr. Peter Johansson,
Department of Physics,
University of Helsinki

Defence of the thesis: 09.10.2026, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology has not been submitted for doctoral or equivalent academic degree.

Juan Urrutia Pérez

---

signature

Copyright: Juan Urrutia Pérez, 2025
ISSN 2585-6898 (paperback)
ISBN 978-9916-80-565-7 (paperback)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-564-0 (PDF)
https://doi.org/10.23658/taltech.65/2026
Printed by Koopia Niini & Rauam

Urrutia Perez, J. (2026). *Black Holes and Dark Matter in Light of New Infrared and Gravitational Wave Observations* [TalTech Press]. https://doi.org/10.23658/taltech.65/2026

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
65/2026

Mustad augud ja tumeaine uute infrapuna- ja gravitatsioonilainete vaatluste valguses

JUAN URRUTIA PÉREZ

TAL TECH
KIRJASTUS

Contents

List of Publications 7

Author's Contributions to the Publications 8

Abbreviations 10

1 Introduction 12
- 1.1 Research objectives of this thesis 15
- 1.2 Structure of the thesis 16

2 Gravitational wave phenomenology 17
- 2.1 Generation of gravitational waves 18
- 2.2 Circular binaries in the Newtonian approximation 19
- 2.3 Beyond the Newtonian approximation 21
- 2.4 Detectability by designed detectors 23
- 2.5 Merger rates and detectability of binaries 26
- 2.6 Pulsar timing arrays 28
  - 2.6.1 The Hellings–Downs correlation 29
  - 2.6.2 The gravitational wave background spectrum 30
  - 2.6.3 The 2023 PTA signal and the current observational status 31
- 2.7 Detector landscape 32

3 Structure formation in different dark matter models 34
- 3.1 The Linear Density Field and its Statistics 34
- 3.2 Suppressed small-scale structures 35
- 3.3 Enhanced small-scale structures 37
- 3.4 Spherical collapse 38
- 3.5 The Press-Schechter Formalism 39
  - 3.5.1 The Basic postulate 39
  - 3.5.2 The Halo mass function 39
- 3.6 The extended Press Schechter formalism 40
- 3.7 Conditional probabilities 41
- 3.8 Ellipsoidal collapse and the moving barrier 43
- 3.9 Ellipsoidal collapse and the Sheth-Mo-Tormen mass function 44
- 3.10 Improving the Sheth-Mo-Tormen approach 44
  - 3.10.1 Conditional first-crossing distribution in the ellipsoidal model 45
  - 3.10.2 A new smoothing filter 46
- 3.11 The halo merger rate 47
- 3.12 The halo dark matter accretion rate 49
- 3.13 Evolution of baryons and black holes through mergers 49

4 Wave optics lensing of gravitational waves 51
- 4.1 The amplification function 51
  - 4.1.1 Geometrical-optics limit 53
  - 4.1.2 Point-mass lens 54
  - 4.1.3 Uniform density sphere 55
  - 4.1.4 Navarro–Frenk–White halo 56
- 4.2 Single lens effects: Computing the cross section 57

5 Massive black holes in our universe 60
5.1 Local Universe 60
5.1.1 PTA merger rate model 61
5.1.2 The binary timescale 63
5.1.3 Imprints of the seeding process at low redshifts 65
5.2 Supermassive Black Holes in the first billion years of the universe 67
5.2.1 Minimal high-redshift galaxy model 68
5.2.2 The UV Luminosity Function at high redshift 70
5.2.3 Black hole co-evolution 73

6 Summary and conclusions 76
6.1 Wave-optics lensing of gravitational waves by dark matter (O1) 76
6.2 Structure formation beyond cold dark matter (O2) 76
6.3 Supermassive black hole binaries and pulsar timing arrays (O3) 77
6.4 Supermassive black holes and galaxies in the early universe (O4) 77
6.5 Main conclusions 77
6.6 Outlook 78

List of Figures 79

References 80

Acknowledgements 91

Abstract 92

Kokkuvõte 95

Appendix 1 99

Appendix 2 129

Appendix 3 139

Appendix 4 149

Appendix 5 159

Appendix 6 173

Appendix 7 189

Appendix 8 199

Appendix 9 209

Appendix 10 221

Appendix 11 233

Curriculum Vitae 242

Elulookirjeldus 246

List of Publications

The present Ph.D. thesis is based on the following publications that are referred to in the text by Roman numbers.

I M. Fairbairn, J. Urrutia, and V. Vaskonen. Microlensing of gravitational waves by dark matter structures. *JCAP*, 07:007, 2023

II J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational wave microlensing by dressed primordial black holes. *Phys. Rev. D*, 108(2):023507, 2023

III J. Urrutia and V. Vaskonen. Dark timbre of gravitational waves. *Phys. Rev. D*, 111(12):123047, 2025

IV G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Did JWST observe imprints of axion miniclusters or primordial black holes? *Phys. Rev. D*, 107(4):043502, 2023

V J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. Starlight from JWST: Implications for star formation and dark matter models. *Astron. Astrophys.*, 702:A109, 2025

VI J. Ellis, M. Fairbairn, G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Prospects for future binary black hole gravitational wave studies in light of PTA measurements. *Astron. Astrophys.*, 676:A38, 2023

VII J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data. *Phys. Rev. D*, 109(2):L021302, 2024

VIII J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Probing Supermassive Black Hole Seed Scenarios with Gravitational-wave Measurements. *Astrophys. J.*, 964(1):11, 2024

IX J. Ellis, M. Fairbairn, G. Hütsi, J. Urrutia, V. Vaskonen, and H. Veermäe. Consistency of JWST black hole observations with NANOGrav gravitational wave measurements. *Astron. Astrophys.*, 691:A270, 2024

X J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. The origin of the JWST supermassive black holes. *Astron. Astrophys.*, 703:A138, 2025

XI J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Constraints on Dark Matter Models from Supermassive Black Hole Evolution. *Phys. Rev. Lett.*, 136(13):131001, 2026

Author's Contributions to the Publications

In high-energy physics phenomenology the authors of a paper are listed alphabetically, so the author order of the publications above does not reflect the individual contributions. In all eleven publications I was one of the main contributors: I took part in formulating the research questions, carried out a substantial part (in most cases the majority) of the analytical and numerical work, produced the figures and co-wrote the manuscripts. My specific contributions to each publication are the following.

I I computed the wave-optics amplification factor for the uniform-sphere, NFW and fuzzy dark matter lens profiles, classified the lensing regimes, performed the detectability analysis (waveform mismatch and lensing probability) for LIGO, ET, LISA, AEDGE and BBO, produced the figures and co-wrote the paper.

II I computed the amplification factor of primordial black holes dressed with particle dark matter halos, derived the constraints from the LIGO–Virgo catalogue and the projections for the Einstein Telescope, produced the figures and wrote the bulk of the manuscript. Except for the part about the PBH dress.

III I proposed the idea, developed the Fresnel-volume argument and the computation of the cumulative lensing signal from many light dark matter halos, carried out all the numerical work, produced the figures and wrote the bulk of the manuscript.

IV I contributed to the computation of the variance of matter perturbations for the axion minicluster and of the halo mass function in models with enhanced small-scale structure, to the comparison with the early JWST galaxy candidates, produced part of the figures and co-wrote the paper.

V I developed the improved extended Press–Schechter formalism (conditional first-crossing distributions and smooth window function), implemented the UV luminosity function model with dust attenuation, produced part of the figures and wrote the bulk of the manuscript.

VI I implemented the supermassive black hole binary population model, computed the distribution of $\Omega_{\rm GW}$ and the predictions for LISA and atom interferometers, produced the figures and co-wrote the paper.

VII I implemented the environmental hardening model and the fit to the NANOGrav 15-year spectrum, contributed to the identification of the population responsible for the signal, produced the figures and co-wrote the paper.

VIII I developed the parametrisation of the occupation fraction for different seeding scenarios, computed the expected event rates and the recoverability of the minimum supermassive black hole mass with LISA and atom interferometers, produced the figures and co-wrote the paper.

IX I performed the fits to the local and JWST supermassive black hole–stellar mass relations, the comparison with the merger-rate model constrained by pulsar timing arrays and the analysis of little red dots and dual AGN, produced the figures and co-wrote the paper.

X I proposed the differential-equation treatment of halo mergers, developed and implemented the semi-analytical seed-evolution model, performed the likelihood scan over the seeding parameters, produced the figures and wrote the bulk of the manuscript.

XI I developed the combined structure-formation and feedback model, performed the fits to the UV luminosity function and to the supermassive black hole observations for cold, warm, fuzzy and enhanced-power dark matter models, derived the resulting constraints, produced the figures and co-wrote the paper.

Abbreviations

AGN	Active galactic nucleus
ALP	Axion-like particle
BH	Black hole
BLR	Broad-line region
CDM	Cold dark matter
CMB	Cosmic microwave background
DM	Dark matter
EdS	Einstein–de Sitter
EPS	Extended Press–Schechter
ET	Einstein Telescope
FDM	Fuzzy dark matter
GR	General relativity
GW	Gravitational wave
HD	Hellings–Downs
HMF	Halo mass function
HST	Hubble Space Telescope
IMBH	Intermediate-mass black hole
IMF	Initial mass function
ISCO	Innermost stable circular orbit
JWST	James Webb Space Telescope
LIGO	Laser Interferometer Gravitational-Wave Observatory
LISA	Laser Interferometer Space Antenna
LVK	LIGO–Virgo–KAGRA
MSP	Millisecond pulsar
NFW	Navarro–Frenk–White
PBH	Primordial black hole
PN	Post-Newtonian
PS	Press–Schechter
PTA	Pulsar timing array
SFR	Star formation rate
SGWB	Stochastic gravitational wave background
SHMR	Stellar-to-halo mass relation
SMBH	Supermassive black hole
SMT	Sheth–Mo–Tormen
SN	Supernova
SNR	Signal-to-noise ratio
ToA	Time of arrival
TT	Transverse-traceless
UVLF	Ultraviolet luminosity function
WDM	Warm dark matter
WIMP	Weakly interacting massive particle

"From space, with its brother time, beneath insistent gravity, feeling matter as a slower space, I wonder with awe at what I do not know."

— Eduardo Chillida

1 Introduction

Throughout the latter half of the twentieth century, the dominant paradigm for probing the fundamental laws of nature was high-energy particle physics: the construction of ever-larger colliders provided access to progressively shorter length scales, culminating in the discovery of the Higgs boson at the Large Hadron Collider in 2012 [39] and the completion of the Standard Model. At the same time that the last pieces of the Standard Model were being found, a parallel revolution was about to happen. The opening of the gravitational-wave (GW) window, with the first direct detection of a binary black-hole (BH) merger by the Advanced LIGO detectors in 2015 [1], implied a profound shift in phenomenological focus: from the microscopic to the cosmic. This happened at the same time that cosmology became a precision science, with no longer a scarcity of data. Both developments have changed the methods of the phenomenologist: new observations and challenges keep coming, and fundamental physics can no longer be disentangled from astrophysics and cosmology, since the universe itself is our laboratory. GWs are not merely a new *channel* of observation; they are the echoes of some of the most energetic events since the Big Bang [51]. BH mergers, neutron-star coalescences, and the stochastic background imprinted by processes in the very early universe all encode information that is fundamentally inaccessible to electromagnetic telescopes. In this sense, GW astronomy completes a programme that began with the opening of the radio, X-ray, and infrared windows: it does not merely extend our reach, it changes the nature of what we can see.

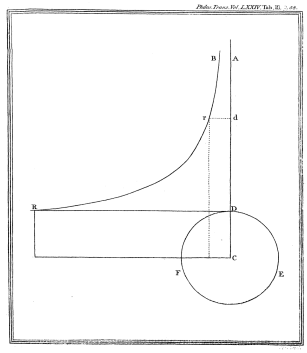


Figure 1: First sketch of the orbits around a "dark star", the precursor of the BH concept. Reproduced from Michell (1784) [99], Philosophical Transactions of the Royal Society; public domain.

The early incarnation of a BH, i.e. the concept of a body so compact that light cannot escape its gravitational pull, was contemplated as early as 1783 by John Michell [99], who claimed "*If there should really exist in nature any bodies, whose density is not less than that of the sun, and whose diameters are more than 500 times the diameter of the sun, since their light could not arrive at us [...] we could have no information from light about their existence*" and, independently, by Pierre-Simon Laplace at the close of the eighteenth century. The modern mathematical framework for such objects emerged from Schwarzschild's exact vacuum solution to Einstein's field equations, obtained in 1916 only months after the publication of general relativity (GR) [122]. The astrophysical reality of BHs became firmly established over the following decades, first as the engines powering quasars and active galactic nuclei [92], and later as the remnants of massive stellar evolution, revealed through radio, X-ray, and optical observations. Today, the existence of supermassive BHs (SMBHs) at the centres of virtually all massive galaxies, with masses ranging from $\sim 10^6$ to $\sim 10^{10} M_\odot$, is considered a cornerstone of observational cosmol-

ogy [86, 117].

GWs have an equally rich history. Einstein himself demonstrated in 1916 and 1918 that linearised GR admits wave solutions propagating at the speed of light, carrying energy away from accelerating mass distributions [47, 48]. Indirect evidence for their existence arrived in 1974 with the discovery of the Hulse–Taylor binary pulsar, whose orbital decay matched the GR prediction for energy loss by gravitational radiation with exquisite precision [90]. Yet direct detection had to wait until the twenty-first century. The milestone observation of GW150914 by Advanced LIGO, a signal consistent with the merger of two stellar-mass BHs, inaugurated GW astronomy as an observational discipline [2]. Since then, the LIGO–Virgo–KAGRA (LVK) collaboration has catalogued more than one hundred compact binary coalescences [6].

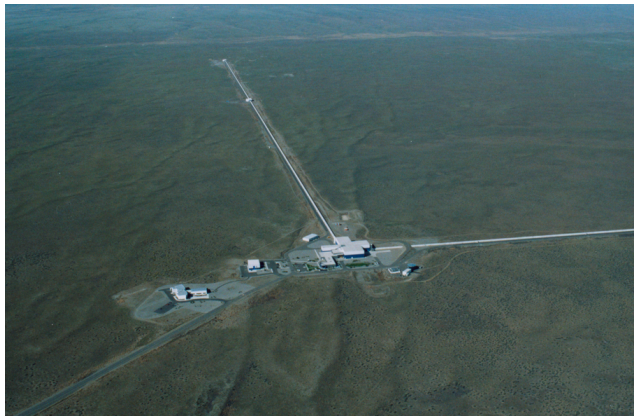


Figure 2: Aerial view of the LIGO Hanford observatory. Image credit: Caltech/MIT/LIGO Laboratory [33].

The frequency of a GW signal is set by the orbital dynamics of the source: the more massive the binary, the lower the frequency at merger, since two heavier BHs collide at a wider orbital separation. The LVK detectors were designed to target stellar-mass BHs, but a new generation of experiments operating at much lower frequencies is opening additional windows onto the BH mass spectrum. In particular, pulsar timing arrays (PTAs) have recently reported strong evidence for a stochastic GW background in the nanohertz band [10, 16, 144, 116]. This signal is consistent with the inspiral of the heaviest SMBH binaries in the local universe, though contributions from exotic early-universe sources cannot be excluded. Interpreting this background requires both a detailed model of the SMBH binary population and a precise understanding of how those binaries evolve and eventually merge, topics that are central to this thesis.

At the same time, the advent of the James Webb Space Telescope (JWST) [66] has transformed our ability to probe the high-redshift universe, opening an unprecedented infrared window onto the epoch of reionisation and beyond. Looking to greater cosmological distances is equivalent to looking further back in time, and with JWST we are witnessing the formation of the first SMBHs and galaxies. Among the most consequential results of this new observational era has been the discovery of a population of SMBHs at redshifts $z \gtrsim 4$–10 that challenges the conventional picture of SMBH formation [28, 95]. These objects appear over-massive relative to their host galaxies compared to local scaling relations, and their very existence at such early cosmic times challenges standard seeding and growth scenarios.

This goes to the heart of one of the deepest open questions in astrophysics: what is the origin of SMBHs? Two broad classes of formation mechanism have long been debated. In the *light-seed* scenario, SMBHs grow from $\mathcal{O}(10^2\text{–}10^3)\,M_\odot$ remnants of Population III stars that collapsed at $z \gtrsim 15$ [94]. In the *heavy-seed* scenario [142], $\mathcal{O}(10^5)\,M_\odot$ BHs form in the cores of protogalaxies and subsequently merge into the giants observed today. The high-redshift population uncovered by JWST appears to favour either a heavy-seed origin or exceptionally high merger efficiency for light seeds, a conclusion that constitutes another important result of this thesis. Additional complexity arises from the possibility of super-Eddington accretion [20], which could lead to systematic overestimates of SMBH masses inferred from broad-line luminosities, and from the enigmatic *little red dots* [98], compact red sources identified in JWST imaging that defy standard assumptions about high-redshift galaxies and whose connection to SMBH activity remains actively debated.

These questions will converge with the Laser Interferometer Space Antenna (LISA) [15], a proposed GW space detector for the 2040s that will cover the frequency range between PTA experiments and terrestrial detectors. With its extraordinary sensitivity, LISA will detect the majority of BH mergers occurring across cosmic history in the mass range between stellar and the heaviest SMBHs, allowing us to compare GW observations with electromagnetic data from JWST and ultimately uncover the seeding and growth history of the SMBHs we observe at the centres of galaxies today.

Underlying all of these questions about BHs and GWs lies a deeper unsolved problem: the nature of dark matter (DM). The existence of a non-baryonic, gravitationally dominant component of the universe has been inferred from a converging body of evidence accumulated over nearly a century. The anomalous rotation curves of spiral galaxies, first systematically studied by Vera Rubin and collaborators in the 1970s, demonstrated that visible matter alone cannot account for galactic dynamics at large radii [119]. The spatial offset between the X-ray gas and the lensing mass in systems such as the Bullet Cluster provided further compelling evidence that the dominant mass component is collisionless and dark [40]. At the largest scales, the acoustic peaks of the cosmic microwave background (CMB) power spectrum, as measured by the *Planck* satellite, require a DM density $\Omega_{\rm DM}h^2 \simeq 0.120$, accounting for roughly 26% of the total energy budget of the universe [12]. Despite this overwhelming indirect evidence, the microscopic identity of DM remains one of the most pressing unsolved problems in fundamental physics.



Figure 3: JWST image of the Bullet Cluster. Image credit: NASA, ESA, CSA, STScI [103].

The long-favoured Weakly Interacting Massive Particle (WIMP) paradigm [25] has faced increasingly stringent direct-detection constraints without a confirmed signal, now approaching the solar-neutrino floor, motivating renewed interest in a wide landscape of alternative candidates spanning many orders of magnitude in mass: from ultra-light axions

at $m_a \sim 10^{-22}$ eV [76] to primordial BHs [35] with masses potentially reaching $\mathcal{O}(10^2)\,M_\odot$. Beyond what constitutes DM microscopically, there is the question of how it behaves on large scales, which is the focus of this thesis. The standard picture, embodied in the ΛCDM cosmological model [126], is cold DM (CDM), in which the DM particles became non-relativistic early on, have no pressure and do not interact with any other field. A key prediction of CDM is the existence of many light DM halos, $< 10^6\,M_\odot$, that do not contain stars [84]; they are one of the main targets of cosmological searches, since they hold the key to how DM behaves on small scales.

GWs offer a uniquely powerful probe of this landscape, precisely because they travel unimpeded through matter and interact with DM structures only gravitationally. Unlike light, which does not couple to DM or does so only negligibly, GWs respond to DM in exactly the same way as they respond to any other form of mass-energy. A compact binary is a highly coherent GW source for which we possess precise theoretical templates, making small waveform deviations due to DM effects potentially detectable. When a GW propagates through a mass concentration whose Schwarzschild radius is comparable to the wavelength of the signal, diffraction and interference effects become significant and the waveform acquires frequency-dependent modulations, a regime known as *wave-optics lensing* [129]. This is qualitatively distinct from the geometric-optics limit appropriate for electromagnetic radiation, where lensing simply magnifies or demagnifies a source. The wave-optics lensing signal depends sensitively on whether DM is distributed in relatively non-dense CDM halos or concentrated in compact objects such as PBHs, providing a new handle on the DM distribution on small scales.

The evolution of the SMBHs themselves encodes complementary information about DM. The seeding and growth history of SMBHs is tightly connected to the underlying DM model, for a reason that is easy to state. SMBHs grow primarily through two channels: accretion of baryonic matter and mergers with other BHs. Both processes are governed by the large-scale structure, which is in turn dominated by the DM properties. Because BHs, unlike stars, live essentially forever, they retain imprints of their formation conditions deep in the high-redshift universe, when the universe was assembling its first low-mass halos; precisely the regime in which different DM models diverge most sharply. This avenue for probing DM has not been widely explored and, as we shall see, it leads to one of the main results of this thesis.

It is within this rapidly evolving observational landscape (nanohertz GWs from PTAs, infrared light from JWST, and the prospect of millihertz signals from LISA) that the present thesis is situated. The work addresses three interconnected themes: the modelling of the SMBH binary population and its GW emission as constrained by PTA data, the formation and evolution of SMBHs in the high-redshift universe and its dependence on the DM model, and the use of wave-optics GW lensing as a probe of DM structures. Together, these threads constitute a unified programme: using GWs and their sources as precision tools to illuminate some of the most fundamental open questions in modern cosmology.

1.1 Research objectives of this thesis

The overarching aim of this thesis is to use GWs and their sources, in particular BHs, as precision probes of the nature of DM. This aim is pursued through four concrete research objectives, each of which is addressed by a subset of the publications listed on page 7 and referred to below by their Roman numerals.

O1 Wave-optics lensing of GWs as a probe of DM structures. To compute the wave-optics amplification factor for realistic DM lenses (point masses, uniform spheres,

NFW and FDM halos, and PBHs dressed with particle DM halos), to determine which of them can be detected or constrained with present and future GW detectors, and to identify the fundamental limitations of such searches (Papers I, II and III).

O2 **An analytic, DM-model-agnostic description of structure formation.** To extend the extended Press–Schechter (EPS) formalism, through ellipsoidal collapse, improved conditional first-crossing distributions and a new smoothing filter, so that the halo mass function, the halo merger rate and the evolution of the baryonic and BH content of halos can be followed consistently in CDM, WDM, FDM and models with enhanced small-scale power, without resorting to Monte Carlo merger trees (Papers IV, V, X and XI).

O3 **The SMBH binary population and the nanohertz GW background.** To build a model of the SMBH binary merger rate anchored to local SMBH observations and to the EPS halo merger rate, to confront it with the PTA measurements, to quantify the role of environmental effects in the binary evolution, and to extrapolate the model to the lower masses and higher frequencies probed by LISA and atom interferometers, including the imprint of the SMBH seeding mechanism (Papers VI, VII, VIII and IX).

O4 **SMBH and galaxy formation in the first billion years as a test of DM.** To interpret the JWST observations of high-redshift galaxies and SMBHs within the framework of O2, to determine which star formation, feedback and seeding scenarios are compatible with the data, and to establish whether the SMBH population constrains deviations from CDM more strongly than the galaxy population does (Papers IV, V, IX, X and XI).

The main results obtained for each objective and the conclusions of the thesis are collected in Section 6.

1.2 Structure of the thesis

The introductory part of the thesis reviews the theory on which the publications are built. Most of it does not represent original contributions but the theoretical background needed to read the articles; the exceptions are the improvements to the EPS formalism in Sections 3.10–3.13 and the modelling of the SMBH population in Section 5, which were developed in the publications. Sections 2–4 deal with different applications of gravity, from the more fundamental to the more elaborate: we start with the basic theory of GWs and their detection, we follow with structure formation in different DM models, and we end with the wave-optics lensing of GWs by DM structures. Section 5 is about the astrophysics of SMBHs in the local and in the high-redshift universe, which connects structure formation to the GW merger rates and to the JWST observations. Section 6 summarises the original results of the publications and states the main conclusions of the thesis. The abstract, the Estonian summary and copies of the eleven publications close the thesis.

2 Gravitational wave phenomenology

We will start with the most fundamental aspect of this thesis, GWs, one of the most profound predictions of Einstein's GR Theory. Spacetime itself can carry energy in the form of propagating curvature waves; they are to gravitons what classical light is to photons. But unlike any other force, GWs do not propagate *through* spacetime: they *are* spacetime oscillations. The complete solution is encoded in the Einstein equation

$$G_{\mu\nu} \;=\; -8\pi G T_{\mu\nu}, \tag{1}$$

where $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu}$ is the Einstein tensor built from the Ricci tensor $R_{\mu\nu}$ and Ricci scalar $R = g^{\mu\nu}R_{\mu\nu}$, and $T_{\mu\nu}$ is the stress-energy tensor of all matter and radiation fields. The Riemann curvature tensor,

$$R^{\mu}{}_{\nu\alpha\beta} = \partial_\alpha \Gamma^{\mu}_{\nu\beta} - \partial_\beta \Gamma^{\mu}_{\nu\alpha} + \Gamma^{\mu}_{\rho\alpha}\Gamma^{\rho}_{\nu\beta} - \Gamma^{\mu}_{\rho\beta}\Gamma^{\rho}_{\nu\alpha}, \tag{2}$$

is constructed from the Christoffel symbols $\Gamma^{\alpha}_{\mu\nu} = \frac{1}{2}g^{\alpha\rho}(\partial_\nu g_{\mu\rho} + \partial_\mu g_{\nu\rho} - \partial_\rho g_{\mu\nu})$. In the flat Minkowski background, $g_{\mu\nu} = \eta_{\mu\nu} \equiv \mathrm{diag}(-1,1,1,1)$, the Christoffel symbols vanish identically. To study linearised GWs we consider a small perturbation $h_{\mu\nu}$ away from the Minkowski metric

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \qquad |h_{\mu\nu}| \ll 1. \tag{3}$$

Expanding the Christoffel symbols to first order in $h_{\mu\nu}$ and substituting into the Ricci tensor, the Einstein equations (1) become the *linearised Einstein equations*:

$$\Box h_{\mu\nu} - \partial^\lambda \partial_\mu h_{\lambda\nu} - \partial^\lambda \partial_\nu h_{\lambda\mu} + \partial_\mu \partial_\nu h^{\lambda}{}_{\lambda} = -16\pi G \tilde{T}_{\mu\nu}, \tag{4}$$

where $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu$ is the flat d'Alembertian and $\tilde{T}_{\mu\nu} = T_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}T^{\lambda}{}_{\lambda}$ is the trace-reversed stress-energy tensor. The first term almost looks like a wave equation, but we will need to absorb the remaining terms via a gauge transformation. Under an infinitesimal coordinate transformation $x^\mu \to x'^\mu = x^\mu + \xi^\mu$ (with $|\partial_\nu \xi^\mu| \ll 1$) the perturbation transforms as

$$h'_{\mu\nu} = h_{\mu\nu} - \partial_\mu \xi_\nu - \partial_\nu \xi_\mu. \tag{5}$$

Any two solutions related by such a transformation describe the same physics. A particularly convenient choice is the *Lorenz* gauge, defined by

$$2\,\partial^\mu \bar{h}_{\mu\nu} = \partial_\nu \bar{h}^{\mu}{}_{\mu}, \tag{6}$$

where $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h^{\lambda}{}_{\lambda}$ is the trace-reversed perturbation. In this gauge, the linearised equations can be reduced to a wave equation:

$$\Box h_{\mu\nu} = -16\pi G \tilde{T}_{\mu\nu}. \tag{7}$$

This single equation shows immediately that small perturbations of the metric obey a wave equation with propagation speed equal to the speed of light. GWs are therefore massless and travel at c — a prediction confirmed to exquisite precision by the multi-messenger observation of GW170817, which constrained the difference between the GW and electromagnetic propagation speeds to be less than 7×10^{-16} relative to the speed of light. In vacuum ($T_{\mu\nu} = 0$) the wave equation becomes $\Box h_{\mu\nu} = 0$, whose plane-wave solution can be written as

$$h_{\mu\nu}(t,\mathbf{x}) = \int d^3k\, e_{\mu\nu}(\mathbf{k})\, e^{ik_\lambda x^\lambda} + \mathrm{c.c.}, \qquad k^\mu k_\mu = 0, \tag{8}$$

where $e_{\mu\nu}$ is the polarisation tensor. The null condition $k^\mu k_\mu = 0$ implies that the wave-vector is light-like: $k^\mu = (\omega, \mathbf{k})$ with $\omega = |\mathbf{k}|$. The Lorenz gauge condition $2k^\mu e_{\mu\nu} = k_\nu e^\mu{}_\mu$ further reduces the ten independent components of the symmetric tensor $e_{\mu\nu}$ from ten to six. Exploiting the residual gauge freedom one can always pass to the *transverse-traceless* (TT) gauge, in which

$$h^{\rm TT}_{0j} = 0 \quad (\text{spatial}), \qquad h^{\mu\,\rm TT}{}_{\mu} = 0 \quad (\text{traceless}), \qquad \partial^\mu h^{\rm TT}_{\mu\nu} = 0 \quad (\text{transverse}). \tag{9}$$

These three conditions completely fix the gauge and leave only *two* physical, propagating degrees of freedom — the two polarisations of the GW. For a wave propagating in the z-direction they are described by the unit polarisation tensors

$$e^{+}_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad e^{\times}_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{10}$$

corresponding to the *plus* ($+$) and *cross* ($\times$) polarisations, respectively. The general TT waveform is then

$$h^{\rm TT}_{\mu\nu}(t,z) = \int d\omega \left[e^{+}_{\mu\nu} h_+(\omega) + e^{\times}_{\mu\nu} h_\times(\omega) \right] e^{i\omega(t-z)} + \text{c.c.} \tag{11}$$

The physical effect of a passing GW is most easily illustrated by computing the proper distance between two freely falling test particles. For a $+$-polarised wave propagating in the z-direction, particles separated by L_c along the x-axis experience an oscillating proper distance

$$L_x(t) \approx L_c \big(1 + h_+ \cos[\omega(t-z)]\big), \tag{12}$$

while test masses separated along the y-axis experience $L_y(t) \approx L_c(1 - h_+ \cos[\omega(t-z)])$. The two arms are therefore squeezed and stretched in opposite phase — precisely the effect exploited by laser interferometric detectors. This result, together with the helicity argument (a rotation by an angle ψ around the propagation axis maps $h_\pm \to e^{\pm 2i\psi} h_\pm$, confirming that the graviton is a spin-2 particle), shows that GWs carry genuine physical information about the dynamics of their sources. This is the main principle to detect GWs, tiny differences in the space-time between two test masses.

2.1 Generation of gravitational waves

Solving the inhomogeneous wave equation, Eq. (7), with a retarded Green's function gives the formal solution

$$h_{\mu\nu}(\mathbf{x},t) = 4G \int d^3x' \, \frac{\tilde{T}_{\mu\nu}(\mathbf{x}', t - |\mathbf{x} - \mathbf{x}'|)}{|\mathbf{x} - \mathbf{x}'|}. \tag{13}$$

Projecting onto the TT gauge via the Lambda tensor $\Lambda_{ij,lm}(\hat{x}) = P_{il}P_{jm} - \frac{1}{2}P_{ij}P_{lm}$, where $P_{ij} = \delta_{ij} - \hat{x}_i \hat{x}_j$ is the transverse projector, one obtains

$$h^{\rm TT}_{ij}(\mathbf{x},t) = 4G \Lambda_{ij,lm}(\hat{x}) \int d^3x' \, \frac{T_{lm}(\mathbf{x}', t - |\mathbf{x} - \mathbf{x}'|)}{|\mathbf{x} - \mathbf{x}'|}. \tag{14}$$

For a source localised near the origin, observed at a large distance $r = |\mathbf{x}| \gg |\mathbf{x}'|, 1/\omega$, we can expand $|\mathbf{x} - \mathbf{x}'| \approx r - \mathbf{x}' \cdot \hat{x}$. Retaining only the leading multipole term, one can use the equation-of-motion identity

$$\int d^3x' \, T_{lm}(\mathbf{x}', t - r) = \frac{1}{2} \ddot{Q}_{lm}(t - r), \tag{15}$$

where

$$Q_{lm}(t) \equiv \int d^3x'\, x'_l x'_m\, T_{00}(\mathbf{x}', t) \tag{16}$$

is the *mass quadrupole moment tensor*. Substituting (15) into (14), the far-field GW amplitude in the TT gauge is the very elegant

$$\boxed{h_{ij}^{\rm TT}(\mathbf{x}, t) \approx \frac{2G}{r}\, \Lambda_{ij,lm}(\hat{x})\, \ddot{Q}_{lm}(t-r)} \tag{17}$$

This is the celebrated *quadrupole formula*: in the slow-motion, weak-field (Newtonian) approximation, GWs are sourced exclusively by the *second time derivative* of the mass quadrupole moment. A crucial consequence is that no monopole or dipole GW radiation exists: the monopole is forbidden by mass-energy conservation, and the dipole by linear-momentum conservation. In the next sections, we will elaborate on how to use this formula on binaries to estimate the production of GWs.

2.2 Circular binaries in the Newtonian approximation

The archetype source of GWs is a binary system of two compact objects (BHs or neutron stars) of masses m_1 and m_2 moving on a circular orbit. We work in center-of-mass coordinates. The positions of the two bodies at time t are

$$\mathbf{r}_1 = r_1 \big(\cos[\omega(t-t_0)+\Phi_0],\, \sin[\omega(t-t_0)+\Phi_0],\, 0\big), \tag{18}$$

$$\mathbf{r}_2 = r_2 \big(\cos[\omega(t-t_0)+\pi+\Phi_0],\, \sin[\omega(t-t_0)+\pi+\Phi_0],\, 0\big), \tag{19}$$

where $r_1 = (m_2/M)R$, $r_2 = (m_1/M)R$, $M = m_1 + m_2$ is the total mass, R is the orbital separation, ω the angular frequency, and Φ_0 the orbital phase at the reference time t_0. The quadrupole tensor (16) evaluates to $Q_{lm} = \sum_j m_j r_{j,l} r_{j,m}$ and its second time derivative is

$$\ddot{Q} = 2\mu R^2 \omega^2 \begin{pmatrix} -\cos[2(\omega(t-t_0)+\Phi_0)] & -\sin[2(\omega(t-t_0)+\Phi_0)] & 0 \\ -\sin[2(\omega(t-t_0)+\Phi_0)] & \cos[2(\omega(t-t_0)+\Phi_0)] & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{20}$$

where $\mu = m_1 m_2/(m_1+m_2)$ is the reduced mass. Applying Kepler's third law, $R^3 = GM/\omega^2$, and projecting onto the TT gauge for a wave propagating toward the observer at inclination ι with respect to the orbital angular momentum, one obtains

$$h_+(t) = \frac{4}{r}\,(G\mathscr{M}_c)^{5/3}\,\omega^{2/3}\,\frac{1+\cos^2\iota}{2}\,\cos[2\omega(t-t_0)+\Psi_0]\,, \tag{21}$$

$$h_\times(t) = \frac{4}{r}\,(G\mathscr{M}_c)^{5/3}\,\omega^{2/3}\,\cos\iota\,\sin[2\omega(t-t_0)+\Psi_0]\,, \tag{22}$$

where $\mathscr{M}_c \equiv M^{2/5}\mu^{3/5}$ is the *chirp mass*, arguably the most important parameter extracted from a compact binary coalescence signal, and $\Psi_0 = 2(\Phi_0 + \varphi - \omega r)$ is the fiducial phase. We will denote these two waveforms as

$$h_+(t) = \frac{1+\cos^2\iota}{2} A(t)\cos[\phi(t)]\,, \tag{23}$$

$$h_\times(t) = \cos\iota\, A(t)\,\sin[\phi(t)]\,. \tag{24}$$

Several important physical features emerge from Eqs. (21)–(22). First, the GW frequency is *twice* the orbital frequency, $f_{\rm GW} = \omega/\pi$. Second, a face-on binary ($\iota = 0$) emits circularly

polarised waves ($h_\times/h_+ = 1$), while an edge-on binary ($\iota = \pi/2$) emits linearly polarised waves ($h_\times = 0$). As expected, since the amplitude scales as $\mathscr{M}_c^{5/3}/r$, the more massive and the closer the system is, the easier it is to detect. The GW energy flux, obtained from the stress-energy tensor of the gravitational field at second order in $h_{\mu\nu}$, is

$$\frac{dE_{\rm GW}}{dt} = \frac{G}{5}\left\langle \dddot{Q}_{ij}\dddot{Q}^{ij}\right\rangle. \tag{25}$$

For a circular binary this gives

$$\frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}t} = \frac{32\,G^{7/3}\,\mathscr{M}_c^{10/3}\,\omega^{10/3}}{5}. \tag{26}$$

By energy conservation, the orbital energy $E_{\rm binary} = -Gm_1m_2/(2R)$ must decrease at the same rate. Using Kepler's law to eliminate R in favour of ω, one derives the *chirp equation* governing the secular evolution of the GW frequency:

$$\frac{\mathrm{d}f_{\rm GW}}{\mathrm{d}t} = \frac{96}{5}\,\pi^{8/3}\,(G\mathscr{M}_c)^{5/3}\,f_{\rm GW}^{11/3}\,. \tag{27}$$

Integrating from some reference time t to coalescence at t_c gives

$$f_{\rm GW}(t) = \frac{1}{8\pi}\,(G\mathscr{M}_c)^{-5/8}\left(\frac{5}{t_c-t}\right)^{3/8}. \tag{28}$$

The frequency increases monotonically as the binary loses energy, diverging as $t \to t_c$: this is the *chirp* that gives both the signal its characteristic morphology and the chirp mass its name. Observationally, the time-frequency track (28) encodes $\mathscr{M}_c$ with remarkable precision because it is set entirely by the binary's dynamics. Combining (26) with Kepler's law yields the evolution of the orbital separation

$$\frac{\mathrm{d}R}{\mathrm{d}t} = -\frac{64}{5}\,\frac{G^3\mu M^2}{R^3}\,, \tag{29}$$

whose solution is $R(t) = [R_0^4 - (256G^3\mu M^2/5)(t-t_0)]^{1/4}$, and the coalescence time from an initial separation R_0:

$$t_c - t = \frac{5}{256}\,\frac{R^4}{G^3\mu M^2}. \tag{30}$$

Let us consider an example: for two $1.4\,M_\odot$ neutron stars initially separated by $R_0 \simeq 1\,\mathrm{AU}$ the coalescence time is of order the Hubble time; for a stellar-mass binary detected by LIGO the residual lifetime at entry into the LIGO band ($\sim 10\,$Hz) is of order minutes to seconds. It is worth noting that $t_c - t$ decreases as only a *quarter power* of R, so the inspiral accelerates dramatically during the final moments. This is the famous chirp of GWs. Notice that we can now understand why BHs emit GWs efficiently, but stars do not. Imagine a face–on binary of two stars like the sun orbiting each other, and let us compare the amplitude of the GW at the moment of maximal emission, that is, when the two objects are about to touch, to the same mass binary but made of BHs

$$\frac{h_+(t_{\rm col})_{\rm BH}}{h_+(t_{\rm col})_{\rm sun}} = \frac{R_\odot}{R_s} = \frac{R_\odot}{2GM_\odot/c^2} \sim 2.4\times 10^5\,, \tag{31}$$

where R_s is the Schwarzschild radius and $R_\odot$ is the solar radius. We can clearly see how a solar-mass BH binary emits a GW amplitude 10^5 times larger than a binary of stars of the

same mass, and this is assuming that the stars will hold their shape until the collapse. In reality, tidal forces will destroy them much earlier.

For the analysis of a signal using a GW detector, it is convenient to work in the frequency domain. A particularly simple expression for the waveform can be achieved in the *stationary phase approximation* (valid when the GW phase $\Phi(t)$ evolves slowly compared to a single cycle), the Fourier transform of $A(t)e^{i\phi(t)}$ is

$$\tilde{h}(f) = \int \mathrm{d}t\, A(t) e^{i\phi(t) - i\omega t} \sim \frac{A(\tilde{t})}{\sqrt{\dot{f}_{\mathrm{GW}}}} e^{i\Psi(f)}, \tag{32}$$

where $\tilde{t}$ is defined such that $\dot{\phi}(\tilde{t}) = 2\pi f_{\mathrm{GW}}(\tilde{t})$ and the frequency-domain amplitude is

$$\tilde{A}(f) = \sqrt{\frac{5}{24}}\, \frac{(G\mathscr{M}_c)^{5/6}}{\pi^{2/3} r}\, f^{-7/6}, \tag{33}$$

gives the important *chirp-mass power law* $f^{-7/6}$. Notice that $A(t)$ is dimensionless, whereas $\tilde{A}(f)$ has units of time. On the other hand, the GW phase is

$$\Psi(f) = 2\pi f t_c + \frac{3}{128}\,(\pi G \mathscr{M}_c f)^{-5/3} - \Psi_c - \frac{\pi}{4}. \tag{34}$$

The leading-order (Newtonian) term $\propto (\mathscr{M}_c f)^{-5/3}$ in the phase is the primary carrier of chirp-mass information, while the characteristic amplitude $\tilde{A}(f) \propto f^{-7/6}$ spans the entire sensitive band of a broadband interferometer. Post-Newtonian corrections to the phase (34) enter at successive powers of $(v/c)^2 \propto (GMf)^{2/3}$ and encode the component masses, spins, and tidal deformabilities, topics we return to later. The formulae above assume circular orbits, valid when GW emission has been the dominant driver of binary evolution for long enough to circularise the orbit. For eccentric orbits, Peters [109] derived generalised expressions; most notably, the energy-loss rate becomes

$$\frac{\mathrm{d}E_{\mathrm{GW}}}{\mathrm{d}t} = \frac{32}{5}\, \frac{G^4 \mu^2 M^3}{a^5 (1-e^2)^{7/2}} \left(1 + \frac{73}{24} e^2 + \frac{37}{96} e^4\right), \tag{35}$$

where a is the semi-major axis and e the eccentricity. Since e decreases during inspiral according to

$$\frac{\mathrm{d}e}{\mathrm{d}t} = -\frac{304}{15}\, \frac{G^3 \mu M^2}{a^4}\, e (1-e^2)^{-5/2} \left(1 + \frac{121}{304} e^2\right), \tag{36}$$

binaries circularise efficiently before entering the LIGO band, making the circular approximation well justified for stellar-mass BH mergers.

2.3 Beyond the Newtonian approximation

The adiabatic approximation underlying the derivation of (27), i.e. that the orbit evolves slowly compared to the orbital period, is quantified by the fractional change in the orbital separation during one orbital period:

$$\frac{\Delta R}{R} \approx 3 \times 10^{-3} \left(\frac{\mathscr{M}_c}{M_\odot}\right)^{5/8} \left(\frac{5}{t_c - t}\right)^{5/8}. \tag{37}$$

This shows that, for $t \ll t_c$, the adiabatic condition is excellently satisfied: the orbit changes by less than a percent per cycle over most of the inspiral. However, when

$$t_c - t \lesssim 0.2\,\mathrm{s}\,(\mathscr{M}_c/M_\odot)^{5/3}, \tag{38}$$

the velocities of the binary components approach $v \sim 0.1\,c$ and the Newtonian approximation breaks down. Near coalescence, the objects move at a substantial fraction of the speed of light, and post-Newtonian corrections become essential. Even during the inspiral, such corrections are required to produce waveform templates accurate enough for matched filtering against LIGO data.

The quadrupole formula can be systematically augmented by a post-Newtonian (PN) expansion of both the equations of motion and the radiation field in powers of $(v/c)^2$. At nPN order, corrections to the GW phase scale as $(GM\pi f)^{2n/3}$ relative to the leading-order term. By 3.5PN order (i.e. seven powers of v/c), the inspiral template is accurate enough to confidently model LIGO detections. For illustrative purposes, we quote the 2PN waveform in the frequency domain [42]:

$$\tilde{h}(f) = \sqrt{\frac{5}{24}}\,\frac{\mathscr{M}_z^{5/6}}{\pi^{2/3} D_L}\, f^{-7/6}\, e^{i\Psi(f)}\,, \tag{39}$$

where $D_L = (1+z_s)^2 D_s$ is the luminosity distance, $\mathscr{M}_z = (1+z_s)\mathscr{M}$ is the redshifted chirp mass, and the GW phase reads

$$\begin{aligned}\Psi(f) = {}& 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}(\pi \mathscr{M}_z f)^{-5/3} \\ & \times \left[1 + \frac{20}{9}\left(\frac{743}{336} + \frac{11}{4}\eta\right) v^2 + (4\beta - 16\pi)\, v^3 \right. \\ & \left. + 10\left(\frac{3058673}{1016064} + \frac{5429}{1008}\eta + \frac{617}{144}\eta^2 - \sigma\right) v^4\right],\end{aligned} \tag{40}$$

with $v = (\pi \mathscr{M}_z \eta^{-3/5} f)^{1/3}$. This waveform is characterised by seven parameters: the redshifted chirp mass $\mathscr{M}_z$, the symmetric mass ratio

$$\eta = \frac{m_1 m_2}{(m_1+m_2)^2}\,, \tag{41}$$

the luminosity distance D_L, the spin-orbit and spin-spin coupling parameters β and σ, the coalescence time t_c, and the coalescence phase ϕ_c.

A physically important feature of this template is that the GW amplitude falls off as the *luminosity distance* D_L, rather than as the flux $\propto r^{-2}$ familiar from electromagnetic radiation. As a consequence, GW detectors are absolute distance indicators in three-dimensional comoving volume, and compact binary mergers can serve as *standard sirens*: by combining the luminosity distance inferred from the waveform amplitude with a spectroscopic redshift of the host galaxy, one obtains a measurement of the Hubble constant that is entirely independent of the cosmic distance ladder [121].

The 2PN inspiral template remains valid up to the innermost stable circular orbit (ISCO), beyond which the perturbative expansion breaks down. The corresponding frequency is

$$f_{\rm ISCO} = \left[6^{3/2}\pi\, \mathscr{M}_z\, \eta^{-3/5}\right]^{-1}. \tag{42}$$

Beyond $f_{\rm ISCO}$ the binary enters the *merger* phase, in which the two objects plunge together on a dynamical timescale. The GW emission during this phase can no longer be described analytically and requires numerical relativity. Nevertheless, phenomenological models provide accurate closed-form approximations calibrated to numerical simulations. The merger is followed by the *ringdown*: the newly formed remnant, typically a Kerr BH, oscillates in its quasi-normal modes and radiates damped sinusoidal GWs whose

frequency and damping time are uniquely determined by the mass and spin of the remnant, providing a direct test of the no-hair theorem of general relativity. A complete phenomenological amplitude envelope covering the three phases—inspiral, merger, and ringdown—is given by [14]

$$A(f;\theta) = \sqrt{\frac{5}{24}} \frac{\mathscr{M}_z^{\frac{5}{6}}}{\pi^{\frac{2}{3}} D_L} \times \begin{cases} f^{-7/6} & f < f_{\rm merg}\,, \\ f_{\rm merg}^{-1/2} f^{-2/3} & f_{\rm merg} \le f < f_{\rm ring}\,, \\ f_{\rm merg}^{-1/2} f_{\rm ring}^{-2/3} \dfrac{\sigma^2}{4(f-f_{\rm ring})^2+\sigma^2} & f_{\rm ring} \le f < f_{\rm cut}\,. \end{cases} \tag{43}$$

The inspiral segment ($f < f_{\rm merg}$) reproduces the standard chirp power law. The merger segment ($f_{\rm merg} \le f < f_{\rm ring}$) transitions to a shallower power law, reflecting the rapid increase of GW emission as the objects plunge together. The ringdown segment ($f_{\rm ring} \le f < f_{\rm cut}$) is modelled as a Lorentzian centred on the quasi-normal mode frequency $f_{\rm ring}$ with width σ, encoding the exponential damping of the remnant's oscillations in the frequency domain. The transition frequencies $f_{\rm merg}$, $f_{\rm ring}$, $f_{\rm cut}$, and the width σ are parameterised as

$$f_j = \frac{\eta^{3/5}\left(a_j \eta^2 + b_j \eta + c_j\right)}{\pi \mathscr{M}_z}\,, \tag{44}$$

where the coefficients a_j, b_j, and c_j are obtained from fits to numerical relativity waveforms and are tabulated in Table I of [14].

2.4 Detectability by designed detectors

Now we will turn to the phenomenology of detecting GWs. We will differentiate between human-made GW detectors, for which there is good control of the noise (examples are LIGO, the proposed LISA, and atom interferometers), and detectors based on natural phenomena. For these detectors, it is possible to understand the spectral sensitivity properly and compute the properties of a binary if the signal–to–noise ratio (SNR) is high enough. On the other hand, we also have more indirect ways to detect GWs that make use of natural phenomena, like pulsar timing arrays (PTAs), for which the noise and the signal cannot be disentangled and at the present moment, we do not have the sensitivity to detect an individual binary but rather a stochastic background. We will focus first on the detectability of the first class of detectors.

Let's consider a GW propagating from direction $\hat{N}(\theta,\phi)$ with polarisation angle ψ. An L-shaped laser interferometer of arm length L measures the differential strain

$$h(t) = \frac{\Delta L(t)}{L} = \frac{\delta L_x(t) - \delta L_y(t)}{L}, \tag{45}$$

where $\delta L_{x,y}$ are the length changes induced in each arm. In general, the measured strain is a linear combination of the two GW polarisations,

$$h(t) = F_+(\theta,\phi,\psi)\, h_+(t) + F_\times(\theta,\phi,\psi)\, h_\times(t), \tag{46}$$

where the antenna pattern functions of an L-shaped interferometer are

$$\begin{aligned} F_+ &= \frac{1}{2}(1+\cos^2\theta)\cos 2\phi \cos 2\psi - \cos\theta \sin 2\phi \sin 2\psi\,, \\ F_\times &= \frac{1}{2}(1+\cos^2\theta)\cos 2\phi \sin 2\psi + \cos\theta \sin 2\phi \cos 2\psi\,, \end{aligned} \tag{47}$$

with $0 \leq F_{+,\times} \leq 1$ and $F_+^2 + F_\times^2 \leq 1$. For a compact binary inspiral at inclination ι, the signal arriving at the detector can be written compactly as

$$h(t) = \mathscr{F}(\theta, \phi, \psi, \iota) A(t) \cos[\Phi(t) - \varphi_P(\theta, \phi, \psi, \iota)] \,, \tag{48}$$

where

$$\mathscr{F} \equiv \sqrt{F_+^2 \, \frac{(1+\cos^2 \iota)^2}{4} + F_\times^2 \cos^2 \iota} \,, \tag{49}$$

and $A(t)$ is the slowly varying amplitude driven by the chirp dynamics described in the previous section. The factor $\mathscr{F}$ encodes the sensitivity loss due to the sky position and polarization of the source. Because real sources are distributed isotropically, one must average over all angles; this averaging gives

$$\frac{1}{16\pi^2} \int \mathrm{d}\cos\iota \, \mathrm{d}\psi \, \mathrm{d}\phi \, \mathrm{d}\cos\theta \; \mathscr{F}^2 = \frac{4}{25} \,, \tag{50}$$

so the rms detector response is reduced by a factor of $\sqrt{4/25}$ relative to its maximum value. This geometric averaging is the origin of a prefactor $\bar{\omega} = 4/25$ appearing in (58) for L-shaped detectors such as LIGO; for a triangular-shaped detector like LISA, the corresponding factor is $\bar{\omega} = 3/25$. The output of a GW detector contains instrumental noise $n(t)$ and, possibly, a signal $h(t)$,

$$s(t) = n(t) + h(t). \tag{51}$$

The standard technique to extract a weak, known-waveform signal from noisy data is *matched filtering*: one convolves the detector output with a filter function $q(t)$ to form the statistic

$$C(\tau) \equiv \int \mathrm{d}t \; s(t) \, q(t - \tau). \tag{52}$$

Assuming stationary Gaussian noise with one-sided power spectral density $S_n(f)$, defined through $\langle \tilde{n}(f) \tilde{n}^*(f') \rangle = \frac{1}{2} S_n(f) \delta(f - f')$, the signal-to-noise ratio (SNR) is maximized by the Wiener filter $\tilde{q}(f) = \tilde{h}(f) e^{-2\pi i f \tau} / S_n(f)$, which yields

$$\mathrm{SNR} = \sqrt{\langle \tilde{h} | \tilde{h} \rangle} \,, \qquad \langle a | b \rangle \equiv 2 \int_0^{f_{\mathrm{cut}}} \mathrm{d}f \, \frac{a^*(f) b(f) + b^*(f) a(f)}{S_n(f)} \,. \tag{53}$$

This inner product is the natural scalar product in the space of GW templates weighted by the inverse detector noise. The SNR quantifies the detectability of a given signal: a source is deemed confidently detected when $\mathrm{SNR} \gtrsim 8$ for a single detector, or when a consistent SNR is observed in a coincident multi-detector network. For a lower SNR, the detectability is going to depend on the orientation of the binary, as we will explain further below.

Let us now pause for a moment before continuing and elaborate on what we have found so far by looking at a couple of examples. We can see in Fig. 4 a comparison between the dimensionless characteristic strain of the waveforms we have described, given by $|\tilde{h}(f)/\sqrt{f}|$, and compare it to the strain sensitivity expected for LIGO run 5. We see the different power-law behaviour of the inspiral, merger and ringdown and how the merger happens at higher frequency for lighter binaries. For the LIGO frequency range, we can detect the merger of binaries like the ones we have here at $60 M_\odot$, the heavy end of the stellar mass distribution. As we have anticipated, the heavier the binary, the louder the signal is. In the other extreme of the mass range at $5 M_\odot$, which could be a neutron star

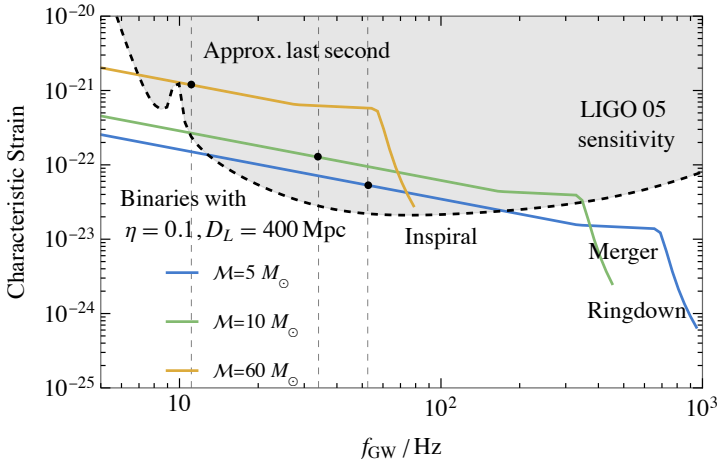


Figure 4: Characteristic strain of three benchmark binaries at a luminosity distance of $400\,\mathrm{Mpc}$*, compared with the expected sensitivity of Advanced LIGO in its fifth observing run (O5).*

binary, we can see only the inspiral phase, and the amplitude is much smaller than for the heavy BH binary. For illustrative purposes, we also show a light BH binary at $10\,M_\odot$. All the binaries are shown at a fixed luminosity distance of $400\,\mathrm{Mpc}$, this is the relevant distance for LIGO, although the majority of binaries are detected at lower redshift. In cosmological scales, this is the local universe, but it does not correspond to our own galaxy. For reference, the local group is around 3 Mpc, and the local supercluster is 250 Mpc, so with advanced LIGO, we can see binaries from structures which are gravitationally independent from ours but still in the local universe. We will see that for LISA the target binaries can be detected even in the first galaxies in the history of the cosmos. For these benchmark binaries, the SNRs are 47, 85 and 300, respectively. We will explain later how this translates into the prospect to detect these binaries.

Once a signal has been detected, one wishes to infer the physical parameters θ of the source from the waveform $\tilde{h}(f;\theta)$. In the high-SNR limit, the posterior distribution on parameters is approximately Gaussian, and the parameter covariance is captured by the *Fisher information matrix* [111],

$$\Gamma_{ij} = \langle \partial_i \tilde{h} | \partial_j \tilde{h} \rangle \,, \tag{54}$$

where the partial derivatives are taken with respect to the waveform parameters. The Fisher matrix provides a lower bound on the variance of any unbiased estimator, and its inverse gives the parameter covariance matrix. For spin parameters, prior information can be incorporated via the replacement $\Gamma_{ij} \to \Gamma_{ij} + (\sigma_j^{(0)})^{-2}\delta_{ij}$, with $\sigma_\beta^{(0)} = 8.5$ and $\sigma_\sigma^{(0)} = 5.0$ following [111]. The standard deviations of the parameter measurements and the correlation matrix then read

$$\Delta j = \sqrt{\Gamma^{-1}_{jj}} \,, \qquad c_{ij} = \frac{\Gamma^{-1}_{ij}}{\Delta i \Delta j} \,. \tag{55}$$

Another technique for inferring the parameter posterior is to use the matched-filtering framework. Consider data that consists of a GW signal ϕ and noise that we characterize by the noise power spectral density $S_{\mathrm{n}}(f)$. The standard matched filtering analysis searches

for the parameters θ of a given template ϕ_T that maximize the log-likelihood [81]

$$\ln \Lambda_T(\phi) = \langle \phi | \phi_T(\theta) \rangle - \frac{1}{2} \langle \phi_T(\theta) | \phi_T(\theta) \rangle \,. \tag{56}$$

For an optimal template $\phi_T = \phi_{\rm opt}$ we can find parameters for which $\phi = \phi_{\rm opt}$. If we know the optimal template, we can calculate how well another template ϕ_T fits the signal by maximizing the log-likelihood (56) over the parameters of the template and compare it to the optimal template by calculating the log-likelihood difference

$$\begin{aligned} \Delta\chi^2 &\equiv 2 \left[\ln \Lambda_{\rm opt}(\phi) - \max_\theta \ln \Lambda_T(\phi) \right] \\ &= \min_\theta \left[\langle \phi_T(\theta) | \phi_T(\theta) \rangle - 2 \langle \phi_{\rm opt} | \phi_T(\theta) \rangle + \langle \phi_{\rm opt} | \phi_{\rm opt} \rangle \right] \\ &= \min_\theta \langle \phi_{\rm opt} - \phi_T(\theta) | \phi_{\rm opt} - \phi_T(\theta) \rangle \\ &= 4 \min_\theta \int \mathrm{d}f \, \frac{|\tilde{\phi}_{\rm opt}(f) - \tilde{\phi}_T(f;\theta)|^2}{S_{\rm n}(f)} \,. \end{aligned} \tag{57}$$

Given a detected waveform $\tilde{\phi}(f;\theta_s)$ with true binary parameters θ_s, we compare it to our template $\tilde{h}_T(f;\theta)$ using the parameters θ, the log-likelihood ratio quantifying the evidence for our estimation is

$$\Delta\chi^2(\theta_s) = 4\bar{\omega} \min_\theta \int \mathrm{d}f \, \frac{|\tilde{h}(f;\theta_s) - \tilde{h}_T(f;\theta)|^2}{S_{\rm n}(f)} \,, \tag{58}$$

where the minimisation is over the parameters of the template, $S_n(f)$ is the noise power spectral density of the detector, and $\bar{\omega}$ is the sky- and inclination-averaged geometric factor discussed above.

2.5 Merger rates and detectability of binaries

We will now introduce another key object in GW phenomenology, the binary merger rate. It gives the rate per comoving volume and binary properties of the merger formation. For the LIGO population, which comes from stellar-mass BHs and neutron stars, we will use the following parametric form

$$\frac{\mathrm{d}R}{\mathrm{d}m_1 \mathrm{d}m_2} = \frac{R_0}{Z_\psi} \, \eta^\beta \, \psi(m_1) \, \psi(m_2) \, \mathrm{SFR}(z) \,, \tag{59}$$

where the redshift evolution is taken to be proportional to the star formation rate [21]. At $z < 2$ this behaves roughly as $\propto (1+z)^{2.7}$, in good agreement with the LIGO-Virgo observations [7]. The normalisation factor Z_ψ is chosen such that $R(z=0) = R_0$. For the BH mass function we adopt a broken power-law shape,

$$\psi(m) \propto \begin{cases} (m/m_2)^{\gamma_1} \,, & m_1 \leq m < m_2 \,, \\ (m/m_2)^{\gamma_2} \,, & m \geq m_2 \,, \end{cases} \tag{60}$$

with the break points fixed at $m_1 = 3\,M_\odot$ and $m_2 = 55\,M_\odot$, corresponding to the lower and upper edges of the BH mass gaps. The lower mass gap separates neutron stars from stellar-mass BHs, while the upper mass gap arises from pair-instability supernova processes that inhibit the formation of BHs in the mass range ~ 55–$130\,M_\odot$. The free parameters of the

model, R_0, β, γ_1 and γ_2, are fitted to the BH binary merger catalogue from the first three LIGO-Virgo observing runs [3, 5, 6] following the procedure described in Ref. [78]. The resulting best-fit values are

$$R_0 = 15\,\mathrm{yr}^{-1}\mathrm{Gpc}^{-3}\,, \quad \beta = 4.8\,, \quad \gamma_1 = -1.8\,, \quad \gamma_2 = -6.0\,. \tag{61}$$

The expected differential rate of compact binary mergers per unit comoving volume element is then

$$\mathrm{d}\lambda = \frac{1}{1+z_s}\frac{\mathrm{d}R}{\mathrm{d}m_1\,\mathrm{d}m_2}\frac{\mathrm{d}V_c}{\mathrm{d}z_s}\,\mathrm{d}m_1\,\mathrm{d}m_2\,\mathrm{d}z_s\,, \tag{62}$$

where the factor $(1+z_s)^{-1}$ converts from the source-frame rate to the observed rate, accounting for cosmological time dilation. The number of detectable events given the merger rate and a given detector is

$$\bar{N}_{\mathrm{events}} = \mathscr{T}\int \mathrm{d}\lambda\; p_{\mathrm{det}}\left(\frac{\mathrm{SNR}_c}{\mathrm{SNR}(\theta_s)}\right)\,, \tag{63}$$

where $\mathscr{T}$ is the observation time, p_{det} is the detection probability [67], SNR is the signal-to-noise ratio [63] we have introduced before. The probability of detection accounts for all the possible orientations of the binary with respect to the detector and it is computed at the optimal SNR, SNR_c is normally assumed to be 8, so that $p_{\mathrm{det}}(x<1)$ saturates to 1, making the detection of binaries with $\mathrm{SNR} > \mathrm{SNR_c}$ almost certain.

As discussed earlier, BH binaries can generate two qualitatively distinct classes of signal: long, nearly monochromatic signals from the slowly-evolving inspiralling phase and relatively short signals from the merger and ringdown. The detections of these two types of sources need to be considered separately. The method we have described so far does not take into account the observation time and works very well for LIGO, in which almost all the binaries end in the merger and the observation time is much longer than the signal duration. The reason is that to see an inspiral with LIGO, it should come from a lighter compact object, and we do not expect compact binaries lighter than neutron stars nearby, except for PBH binaries, which have not been detected.

As we will see, a big focus of our work is going to be to predict merger rates for heavier BHs in lower frequency detectors. In this case, there can be a wide range of masses and we expect to detect both the mergers of the heaviest BHs that a given detector can probe and the inspiral of those which are lighter. In this case, the observation time needs to be taken into account. The number of BH binaries at redshift z emitting in a given frequency band can be estimated from the time the binary spends in that frequency band, which gives [124]

$$\frac{\mathrm{d}N}{\mathrm{d}\mathscr{M}_z\mathrm{d}\eta\,\mathrm{d}z\mathrm{d}\ln f} = \frac{\mathrm{d}\lambda}{\mathrm{d}\mathscr{M}_z\mathrm{d}\eta\mathrm{d}z}\left|\frac{\mathrm{d}\tau}{\mathrm{d}\ln f}\right|\,, \tag{64}$$

where, assuming circular orbits and orbital decay via GW emission [108], the coalescence time of an inspiralling binary emitting GWs with frequency $(1+z)f$ is

$$\tau(f) = \frac{5}{256\pi^{\frac{8}{3}}\mathscr{M}_z^{\frac{5}{3}}}f^{-\frac{8}{3}}\,. \tag{65}$$

Notice that it is the same as Eq.(30) after some algebra and plugging values suitable for cosmological scales. From this, it follows that the number of detectable nearly monochromatic GW signals that arise from inspiralling binaries is then given by

$$N_{\mathrm{insp}} = \int_{\mathscr{T}+\tau_{\mathrm{min}}}^{\infty}\mathrm{d}\tau\int\mathrm{d}\lambda\; p_{\mathrm{det}}\left[\frac{\mathrm{SNR}_c}{\mathrm{SNR}(\tau,z)}\right]\,, \tag{66}$$

where $\tau_{\rm min}$ is the arbitrary time to merge that defines an "inspiral" and the SNR for a detector characterised by the noise power spectrum $S_n(f)$ is given by

$$\mathrm{SNR}(\tau,z) = \sqrt{\int_0^{\mathscr{T}} \mathrm{d}t\, \frac{2|h(\tau-t,z)|^2}{S_n(f(\tau-t,z))}}\,. \tag{67}$$

If we performed a similar exercise as the one we did for the neutron star binary, we would see that the timescale for an SMBH binary quickly reaches the Hubble time. The timescale for a circular binary to coalesce is given by

$$t_{\rm GW} = \frac{\dot{E}_{\rm GW}}{E_{\rm GW}} = 4\,\tau(f)\,, \tag{68}$$

for example, a binary of $\mathscr{M}_{\rm c} \sim 10^8\,M_\odot$ will take a Hubble time to merge if the binary is separated at only $0.04\,\mathrm{pc}$. This is, together with the merger rate, the second key ingredient for GW predictions of heavy binaries. A mechanism to lower the timescale and allow for efficient merger is necessary. In the next section, we will introduce structure formation, which is going to be key to predicting the merger rate of heavy binaries, and we will later turn back to the binary timescales when comparing to the PTA observations. But first, let us introduce another kind of detector, PTAs.

2.6 Pulsar timing arrays

The frequency range of the detectors is essentially set by their arm length. As we have shown, designed detectors fall at the stellar-mass compact binaries for the case of LIGO and can be extended to heavier binaries in LISA when they are broadened in space. However, as we showed in Eq. (30), the coalescence time of a circular binary scales as $t_c \propto R^4/(G^3\mu M^2)$: so for the SMBH binaries with total masses $M \sim 10^8$–$10^{10}\,M_\odot$, the ones at the centres of the heaviest galaxies, which could form as a by-product of galaxy mergers, the GW frequency at coalescence falls in the *nanohertz* regime, $f_{\rm GW} \sim 10^{-9}\,\mathrm{Hz}$. These are plainly inaccessible to terrestrial detectors, since the arm length should be of the order of 10 parsecs. Fortunately, nature has already deployed a network of cosmic clocks that are sensitive to exactly this frequency band: an ensemble of millisecond pulsars distributed across the Milky Way, monitored as a PTA.

Millisecond pulsars (MSPs) are neutron stars that have been spun up by accretion from a companion star to rotation periods $P \sim 1$–$10\,\mathrm{ms}$. After this recycling phase they settle into an extremely stable rotational state, with fractional period derivatives $\dot{P}/P \lesssim 10^{-19}\,\mathrm{s}^{-1}$, making them among the most precise natural clocks known. The emission mechanism of a pulsar channels electromagnetic radiation into a narrow beam that sweeps across the sky once per rotation; an observer aligned with the beam records a sharp radio pulse with each rotation. A precise model of the pulsar ephemeris, encoding the rotation period and its derivatives, the pulsar's astrometric position, proper motion, parallax, and, for pulsars in binary systems, the Keplerian and post-Keplerian orbital elements, allows one to predict the arrival times (ToAs) of future pulses with sub-microsecond accuracy over baselines of years to decades. The *timing residual* is defined as the difference between the observed and predicted ToA,

$$R_\alpha(t) \equiv t^\alpha_{\rm obs} - t^\alpha_{\rm pred}(t)\,. \tag{69}$$

A passing GW perturbs the spacetime metric along the line of sight between the pulsar and the observatory. This stretches or compresses the null geodesic of the photon, inducing a

non-zero timing residual. The redshift of the pulse train due to the GW is obtained from the geodesic equation [46]

$$z_\alpha(t) \equiv \frac{\mathrm{d}}{\mathrm{d}t} R_\alpha(t) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \sum_\lambda F_\lambda^\alpha(\hat{k})\, h_\lambda(t, \mathbf{0})\,, \tag{70}$$

where the antenna pattern function for a pulsar with sky position $\hat{p}$ and a GW propagating along $\hat{k}$ is

$$F_\lambda^\alpha(\hat{k}) \simeq \frac{\hat{p}^i \hat{p}^j}{2(1+\hat{k}\cdot\hat{p})}\, e_{ij}^\lambda(\hat{k})\,. \tag{71}$$

This expression holds in the limit $\lambda_{\mathrm{GW}} \ll d_\alpha$, where d_α is the distance to the pulsar. Since PTA frequencies are set by the inverse of the observation time ($T_{\mathrm{obs}} \sim 10$–$30\,\mathrm{yr}$), the corresponding wavelengths are $\lambda_{\mathrm{GW}} \sim c/f \sim 10\,\mathrm{pc}$, comfortably shorter than the kpc-scale distances to the monitored MSPs [46]. Consequently, the so-called *pulsar term*, which encodes the phase of the GW at the pulsar location, is incoherent across different pulsar pairs and averages away in the cross-correlation analysis described below. This simplification — the Earth term dominance — is the key structural difference between PTAs and interferometers.

The contrast with laser interferometry cannot be overstated. In a broadband interferometer, the signal from a compact binary coalescence is a transient event: it lasts seconds to minutes, and the instrument records a clean stretch of noise before and after the event. This allows one to characterise the noise power spectral density $S_n(f)$ directly from the data stream, without the GW signal present, and to then search for an excess power above this floor using the matched-filtering framework of Section 2.2. Signal and noise are, in principle, cleanly separable.

In a PTA, no such separation exists. The lowest accessible frequency is approximately $f_{\mathrm{min}} = 1/T_{\mathrm{obs}}$, and the highest is $f_{\mathrm{max}} \sim 1/(2\Delta t)$, where $\Delta t \sim 2$–4 weeks is the typical cadence between observations. The entire spectrum from nHz to μHz is therefore probed over the duration of the experiment. A low-frequency GW with a period comparable to T_{obs} is permanently present in the data: one cannot wait for it to switch off. Moreover, any unmodelled deterministic process in the pulsar — stochastic spin noise, dispersion measure variations due to the interstellar medium, instrumental systematics— will appear as a red noise process in the timing residuals, mimicking the spectral shape of a GW background. The fundamental challenge of PTA data analysis is therefore not to detect a signal above an empty noise floor, but to disentangle several simultaneously present stochastic processes with similar spectral shapes. As a consequence, even after a common-spectrum process (CSP) has been identified in the PTA data, it cannot be interpreted as a GW background without an additional discriminator. That discriminator is the *spatial correlation* of the timing residuals across different pulsars, to which we now turn.

2.6.1 The Hellings–Downs correlation

One of the most important results in PTA analysis is the classical Hellings–Downs correlation. Consider two pulsars α and β at sky positions $\hat{p}_\alpha$ and $\hat{p}_\beta$, separated by an opening angle γ defined by $\cos\gamma = \hat{p}_\alpha \cdot \hat{p}_\beta$. If both timing residuals are driven by the same stochastic GW background, their redshifts $z_\alpha(t)$ and $z_\beta(t)$ will be correlated because both pulsars sample the same spacetime metric perturbation at the solar system barycentre (the Earth term). The cross-correlation is

$$s_{\alpha\beta} \equiv \langle z_\alpha(t)\, z_\beta(t)\rangle = \int k^2\, \mathrm{d}k\, P_\lambda(|\mathbf{k}|) \int \frac{\mathrm{d}\hat{k}}{4\pi} \sum_\lambda F_\lambda^\alpha(\hat{k})\, F_\lambda^\beta(\hat{k})\,, \tag{72}$$

where $P_\lambda(k)$ is the GW power spectrum, defined through

$$\langle h_\lambda(\mathbf{k}) h^*_{\lambda'}(\mathbf{k}')\rangle = (2\pi)^3 \delta_{\lambda\lambda'} \delta^{(3)}(\mathbf{k}+\mathbf{k}')\, P_\lambda(|\mathbf{k}|)\,, \tag{73}$$

for an isotropic, unpolarised background. Performing the angular integral over the GW propagation direction $\hat{k}$ analytically, using the antenna pattern functions (71), yields the celebrated *Hellings–Downs* (HD) curve [74]:

$$\boxed{\Gamma_{\rm HD}(\gamma) = \frac{1}{4} + \frac{1}{12}\cos\gamma + \frac{1}{2}(1-\cos\gamma)\ln\left(\frac{1-\cos\gamma}{2}\right)} \tag{74}$$

Several features of this function deserve emphasis. First, $\Gamma_{\rm HD}(0) = 1/2$ for a pair of coincident pulsars (excluding the trivial self-correlation at $\gamma = 0$ for $\alpha = \beta$): even for pulsars at the same sky position, the correlation is not unity, because the independent pulsar terms introduce uncorrelated noise. Second, the function is not monotone: it has a zero crossing near $\gamma \approx 49^\circ$ and a minimum near $\gamma \approx 90^\circ$, producing the characteristic quadrupolar pattern. Third, and most importantly, the functional form (74) is a direct and unique consequence of the tensor nature of the metric perturbation: any alternative theory of gravity that predicts additional polarisation modes (scalar, vector) would yield a distinct angular correlation pattern. The HD correlation is therefore the *smoking gun* signature of general-relativistic GWs; it cannot be mimicked by any pulsar-intrinsic or instrumental noise process, because those sources of noise are by construction uncorrelated across different pulsars. This is the key point: while the total power in the timing residuals may be dominated by noise, the *angular pattern* of the cross-correlations is the clean observable.

Instrumental noise processes, clock noise, radiometer noise, and polarisation calibration errors are uncorrelated between different telescopes and different pulsars, contributing zero to $s_{\alpha\beta}$ for $\alpha \neq \beta$. Astrophysical noise processes intrinsic to the pulsar itself, rotational irregularities and spin noise are also uncorrelated between different MSPs.

2.6.2 The gravitational wave background spectrum

In the same way that for LIGO we have introduced the observed merger rate, we will do the same for the PTAs. A big focus of the thesis is going to be to improve upon this simple estimate, but by introducing it, we will have a grasp of the fundamental properties of a stochastic background. Although, as we will see, it is connected to the merger rate, it is also fundamentally different. While the merger rate was about computing the differential number of binaries per unit of time, a stochastic background is a time-independent ensemble of many sources contributing to form a background of GWs. The amplitude of the background is normally defined with respect to the critical density given by

$$\Omega_{\rm GW}(f) \equiv \frac{1}{\rho_c}\frac{{\rm d}\rho_{\rm GW}}{{\rm d}\ln f} = \frac{2\pi^2}{3H_0^2} f^2 h_c^2(f)\,. \tag{75}$$

For a stochastic GW background (SGWB) sourced by an incoherent superposition of GW emitters, the characteristic strain spectrum can be conventionally parametrised as a power law,

$$h_c(f) = A_{\rm GWB}\left(\frac{f}{f_{\rm yr}}\right)^{(3-\gamma)/2}, \tag{76}$$

where $f_{\rm yr} = 1\,{\rm yr}^{-1} \approx 31.7\,{\rm nHz}$ is the reference frequency and $A_{\rm GWB}$ is the amplitude at that frequency. Equation (76) follows from Phinney's theorem [110], which expresses the present-day GW energy density as an integral over the cosmological emission history of a

population of compact binaries. Its derivation rests on a set of assumptions that should be made explicit before the spectrum is used as a benchmark. (i) Statistical stationarity: the differential merger rate density is treated as a smooth, time-translation invariant function of the chirp mass and mass ratio over the observation timescale and over the light-crossing time of the relevant frequency bin; (ii) incoherent superposition: the number of binaries per logarithmic frequency interval is large enough that Poisson fluctuations from individually bright sources are subdominant; (iii) continuous chirp, (iv) circular, GW-driven inspirals. Departures from any of (i)–(iv) modify the spectral index, so the value $\gamma = 13/3$ should be regarded as a benchmark expectation rather than a robust prediction. In fact, as we will see, a great part of the content of this thesis is about going beyond this approximation and showing why it is not a good ansatz.

The spectral index γ encodes the physics of the source. An ensemble of SMBH binaries hardening *solely* by GW emission on circular orbits is expected to produce a background with [110], $\Omega_{\rm GW}(f) \propto f^{2/3}$ corresponding to the characteristic frequency evolution $\mathrm{d}f/\mathrm{d}t \propto f^{11/3}$ from the chirp equation (27). This can be seen immediately: since each binary contributes $\mathrm{d}E_{\rm GW}/\mathrm{d}\ln f \propto f^{2/3}$ (from the quadrupole formula and Kepler's third law), and the number of binaries per logarithmic frequency interval scales as $\mathrm{d}N/\mathrm{d}\ln f \propto |\tau_{\rm GW}(f)|^{-1} \propto f^{-8/3}$ (from Eq. (65)), the total energy density per logarithmic frequency interval $\Omega_{\rm GW}(f) \propto (\mathrm{d}E_{\rm GW}/\mathrm{d}\ln f) \times (\mathrm{d}N/\mathrm{d}\ln f) \propto f^{2/3}$ follows. At the frequencies accessible to PTAs, the effective dimensionless GW amplitude is

$$h_c(f_{\rm yr}) \equiv A_{\rm GWB} \sim 10^{-15}\text{-}10^{-14}\,, \tag{77}$$

corresponding to timing residuals of order $\delta t \sim h_c/f \sim 10\text{-}100\,\mathrm{ns}$, comparable to the timing precision of the best-monitored MSPs. This remarkable coincidence of signal and sensitivity is what makes PTAs viable GW detectors.

2.6.3 The 2023 PTA signal and the current observational status

In 2023, four independent PTA collaborations simultaneously announced evidence for a stochastic GW background through the detection of the HD spatial correlation [10, 17, 116, 144, 9]. The NANOGrav 15-year data set (NG15) reported a signal with characteristic strain amplitude $A_{\rm GWB} = 2.4^{+0.7}_{-0.6} \times 10^{-15}$ at $f_{\rm yr}$ and spectral index $\gamma = 3.2 \pm 0.6$ [10]. The

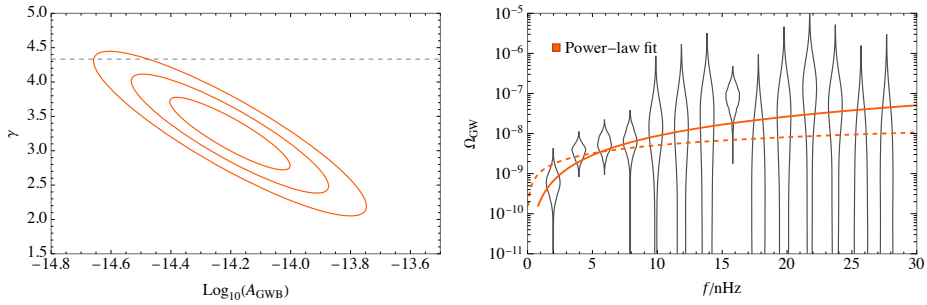


Figure 5: Left panel: Posteriors on $A_{\rm GWB}$ and γ from a fit of the power law of Eq. (76) to the NANOGrav 15-year measurements; the dashed line marks the value $\gamma = 13/3$ predicted for a GW-driven population of inspiralling SMBH binaries. Right panel: Best-fit power law (solid) compared with the NANOGrav 15-year measurements (violins); the dashed line shows the best fit when $\gamma = 13/3$ is imposed. Adapted from Paper VII [53], © 2024 American Physical Society.

spectral shape of the signal is of particular importance for identifying the source. Pure

GW-driven circular SMBH binaries predict $\gamma = 13/3$ in the power-law parametrisation of the timing residual spectrum $P(f) \propto f^{-\gamma}$. We can see in Fig.5 the posteriors on $A_{\rm GWB}$ and γ from fitting a simple power-law parametrisation to the NANOGrav data, which is compatible with a softer spectrum ($\gamma < 13/3$ at 99% confidence level), hinting that the SMBH binaries may be experiencing additional environmental energy losses beyond GW emission at the frequencies probed by the PTA band. Binaries at separations above the "final parsec" — the separation below which GW emission dominates the inspiral — interact with their stellar and gaseous environments through dynamical friction, stellar hardening, and viscous drag from circumbinary discs [52]. These processes accelerate the binary evolution, depleting the low-frequency end of the GW background and steepening the spectral tilt, precisely consistent with the NG15 observations [52]. We will return to this point later, where we model the SMBH binary population and its environmental evolution. On the right panel of Fig.5, we can see how the best fit looks compared to the data. It is especially relevant to see that the reason why the GW-only evolution, shown with a dashed line, cannot reproduce the observations properly is that it is too flat and cannot match the lowering of the amplitude at smaller frequencies. It compensates for it with a lower amplitude but a worse overall fit.

2.7 Detector landscape

Now that we have introduced the basics of GWs, and we have shown the difference between LIGO binaries and stochastic backgrounds such as those detectable with PTAs, it is going to be useful to introduce the landscape of planned detectors. We can see in Fig.6 the sensitivity of the different detectors as compared to the extension of the model we will develop to fit the PTAs. We are getting ahead of ourselves here, but the point is just to compare the frequencies that each detector is sensitive to and what kind of BH mass it approximately corresponds to.

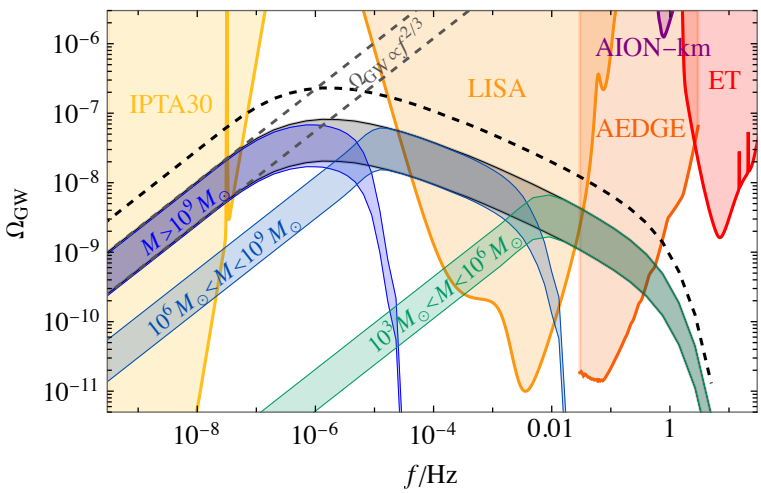


Figure 6: The GW energy density spectrum $\Omega_{\rm GW}$ as a function of frequency for merging SMBHs of different mass ranges (coloured bands), compared with the design sensitivity curves of a representative set of current and future GW observatories. Reproduced from Paper VI [54], © ESO 2023.

At the lowest frequencies, $f \sim 10^{-9}$–10^{-7} Hz, PTAs — represented here by the International PTA (IPTA30) sensitivity curve — are sensitive to GWs with periods of order

years to decades. The characteristic frequency scale of a merging binary is set by its total mass; binaries in this band correspond to the final inspiral of the heaviest SMBHs, with total masses $M \gtrsim 10^9\,M_\odot$, and it is precisely this population that is responsible for the stochastic GW background recently detected at nHz frequencies [11, 52].

Bridging the gap between PTAs and ground-based detectors, the space-borne laser interferometer LISA will operate in the millihertz band, $f \sim 10^{-5}$–10^{-1} Hz. The arm-length of 2.5×10^6 km, set by five-million-kilometre laser links between three free-falling spacecraft, optimises sensitivity around $f \sim 10^{-3}$ Hz and makes LISA the natural observatory for mergers of SMBHs in the mass range $10^4\,M_\odot \lesssim M \lesssim 10^7\,M_\odot$, reaching SNRs of order 10^4 for the heaviest events.

Filling the so-called "mid-band" between LISA and the ground-based network, $f \sim 0.01$–10 Hz, are the proposed atom interferometers AION-km and AEDGE. These detectors exploit the quantum-mechanical interference of laser-cooled atomic clouds rather than macroscopic mirrors, and their baseline lengths, from one kilometre on the ground (AION-km) to a thousand-kilometre satellite separation (AEDGE), place their peak sensitivity squarely in the frequency window where intermediate-mass black hole (IMBH) binaries with $10^3\,M_\odot \lesssim M \lesssim 10^6\,M_\odot$ radiate most of their GW energy. As the figure illustrates, AEDGE has superior reach for the lower end of this mass range, while LISA and AEDGE have complementary and overlapping coverage around $M \sim 10^5\,M_\odot$.

At the highest frequencies, $f \sim 1$–10^3 Hz, ground-based laser interferometers, LVK network and the proposed Einstein Telescope (ET), are sensitive to GW emission from stellar-mass and light intermediate-mass compact binaries with $M \lesssim 10^3\,M_\odot$. The ET, a cryogenic underground instrument with 10 km triangular arms, will extend the accessible volume by roughly two orders of magnitude in luminosity distance relative to current detectors.

Taken together, the ensemble of detectors depicted spans more than twelve orders of magnitude in binary total mass, providing a continuous spectroscopic window onto BH populations from stellar remnants to the most massive objects in the Universe. We note that the PTA frequency band naturally connects to the lower-frequency regime of space-based interferometers such as LISA, which will be sensitive to GWs in the μHz–mHz band. LISA will be able to resolve individual SMBH binary coalescences at high redshift, while PTAs measure the incoherent background from the entire unresolved population. Together, the two probes provide complementary windows on the SMBH mass function and its cosmological evolution, a point that we will explore in this thesis. In order to do this, we will introduce structure formation, which informs us about the galactic mergers. LIGO binaries are a function of the star formation rate since they reflect the stars that die and become BHs, but the SMBH merger and evolution cannot be disentangled from structure formation and as we will discuss, the DM model.

3 Structure formation in different dark matter models

Galaxy mergers drive the coalescence of the SMBHs residing at their centres, and the gravitational evolution of large-scale structure that governs these mergers is itself dictated by DM, which dominates the gravitational mass budget of the universe. Beyond the two-body problem, gravity admits no closed-form analytical solution, and one must resort to numerical methods. N-body simulations have risen to this challenge over the past few decades with spectacular success, resolving the hierarchical assembly of structure across cosmic time with ever-increasing fidelity. Yet a complementary and remarkably powerful analytical approach exists.

The universe emerged from the epoch of recombination at $z \sim 1100$ as a nearly featureless matter distribution, with tiny density fluctuations imprinted on the CMB. The statistical properties of these primordial perturbations are encoded in the CMB anisotropies, and under the assumption of linear growth their subsequent evolution offers a surprisingly accurate window onto cosmic structure— a striking instance of an inherently gravitational problem yielding to a far simpler, non-gravitational perturbative treatment. In this section, we develop the analytical framework for structure formation: beginning with linear perturbation theory and the matter power spectrum, proceeding through the spherical and ellipsoidal collapse models, and culminating in the Press–Schechter and extended Press–Schechter (EPS) formalisms for the halo mass function (HMF). We further improve upon the classical EPS treatment by introducing a revised window function and a new fitting form for the ellipsoidal collapse barrier. From this foundation, we derive the HMF, the halo–halo merger rate, the accretion rate onto forming BHs, and the evolution of the baryonic matter and BHs across cosmic time.

3.1 The Linear Density Field and its Statistics

Let us begin with the description of the original density fluctuations of matter, which will seed the first structures. Inflation predicts that the primordial density perturbations form a Gaussian random field with a nearly scale-invariant power spectrum. In the matter era, the fractional density contrast $\delta(\mathbf{x},t) \equiv [\rho(\mathbf{x},t) - \bar{\rho}(t)]/\bar{\rho}(t)$ evolves according to the linearised continuity, Euler and Poisson equations, which in comoving coordinates give

$$\ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\,\delta = 0\,, \tag{78}$$

where $H = \dot{a}/a$ is the Hubble parameter and $a(t)$ is the scale factor. This equation admits a growing solution $\delta(\mathbf{x},t) = D(t)\,\delta_0(\mathbf{x})$, where $D(t)$ is the *linear growth factor*, conventionally normalised to unity at present t_0, and $\delta_0(\mathbf{x})$ is the density field linearly extrapolated to $z = 0$. In ΛCDM, $D(t)$ must be obtained by numerical integration of Eq. (78), but it scales approximately as the scale factor in the matter-dominated era. Because the density field is a Gaussian random field, its statistical properties are completely characterised by the two-point function, or equivalently by the power spectrum $P(k)$, defined through

$$\langle \delta(\mathbf{k})\,\delta^*(\mathbf{k}')\rangle = (2\pi)^3\,P(k)\,\delta_{\mathrm{D}}^{(3)}(\mathbf{k}-\mathbf{k}')\,, \tag{79}$$

where $\delta(\mathbf{k})$ is the Fourier transform of $\delta(\mathbf{x})$ and δ_{D} is the Dirac distribution. The variance of the density field follows as

$$\sigma^2 \equiv \langle \delta^2 \rangle = \frac{1}{2\pi^2}\int_0^\infty P(k)\,k^2\,\mathrm{d}k = \int_0^\infty \Delta^2(k)\,\frac{\mathrm{d}k}{k}\,, \tag{80}$$

where $\Delta^2(k) \equiv k^3 P(k)/(2\pi^2)$ is the dimensionless power spectrum. A central quantity in what follows is the *smoothed* density field. For a window function $W(\mathbf{x};R)$ of character-

istic scale R, the smoothed field is

$$\delta(\mathbf{x};R) \equiv \int \delta(\mathbf{x}')\, W(\mathbf{x}-\mathbf{x}';R)\, \mathrm{d}^3x' \,, \tag{81}$$

which in Fourier space reduces to $\tilde{\delta}(\mathbf{k};R) = \tilde{\delta}(\mathbf{k})\tilde{W}(kR)$. The three most commonly employed filters are the real-space top-hat (for which the corresponding mass is $M = \frac{4}{3}\pi R^3\bar{\rho}$), the Gaussian, and the sharp k-space filter; their Fourier transforms are, respectively,

$$\tilde{W}_{\mathrm{TH}}(kR) = \frac{3}{(kR)^3}\left[\sin(kR) - kR\cos(kR)\right], \tag{82}$$

$$\tilde{W}_{\mathrm{G}}(kR) = \exp\left(-\frac{(kR)^2}{2}\right), \tag{83}$$

$$\tilde{W}_{\mathrm{sk}}(kR) = \Theta(1-kR), \tag{84}$$

where Θ is the Heaviside step function. We will elaborate later on the problems of these window functions when considering exotic DM models, and we will motivate our proposed window function. The *mass variance* of the smoothed field is then

$$\sigma^2(M) \equiv \sigma^2(R) = \frac{1}{2\pi^2}\int_0^\infty P(k)\,\tilde{W}^2(kR)\,k^2\,\mathrm{d}k \,, \tag{85}$$

and it plays a fundamental role in the Press-Schechter formalism. The cosmological parameter σ_8 is defined as the root-mean-square of the linearly extrapolated density field at $z = 0$ when smoothed with a top-hat filter of radius $R = 8\,h^{-1}$ Mpc; its currently favoured value is $\sigma_8 \simeq 0.81$ [13]. In hierarchical CDM cosmologies, $\sigma(M)$ is a monotonically *decreasing* function of mass: small scales are more strongly fluctuating than large ones, which implies that low-mass structures collapse first and subsequently merge into progressively more massive haloes — the bottom-up, or hierarchical, picture of structure formation. This is not always true and we will explore both fuzzy dark matter (FDM) and warm dark matter (WDM), for which there is a hard cut-off of the small scales because there are processes preventing the collapse of the lightest halos. We will elaborate on this further in the next sections.

3.2 Suppressed small-scale structures

The hierarchical build-up of structure described above is a hallmark of the standard CDM paradigm, in which the DM power spectrum extends to arbitrarily small scales. Two well-motivated alternatives predict a sharp suppression of structure below a characteristic mass scale, and both play an important role in the analysis presented later in this thesis.

The first is FDM, in which the DM is a coherent condensate of an ultra-light bosonic field with mass $m_{\mathrm{FDM}} = \mathcal{O}(10^{-22}\,\mathrm{eV})$. The quantum nature of FDM modifies the perturbation equations at the level of the field's equations of motion. The DM field ψ obeys the Schrödinger–Poisson system,

$$i\hbar\,\partial_t\psi = -\frac{\hbar^2}{2m_{\mathrm{FDM}}a^2}\nabla^2\psi + m_{\mathrm{FDM}}\,\Phi\,\psi \,, \qquad \nabla^2\Phi = 4\pi G\bar{\rho}\,a^2\,\delta \,, \tag{86}$$

where a is the scale factor and Φ is the Newtonian gravitational potential. Writing $\psi = \sqrt{\rho/m_{\mathrm{FDM}}}\,e^{i\theta}$ and defining the velocity field $\mathbf{v} = (\hbar/m_{\mathrm{FDM}}a)\nabla\theta$, the Schrödinger–Poisson

system is exactly equivalent to the continuity and Euler equations for a pressureless fluid plus an additional *quantum pressure* term. The resulting fluid equations read

$$\dot{\delta} + \frac{1}{a}\nabla\cdot\big[(1+\delta)\mathbf{v}\big] = 0\,, \tag{87}$$

$$\dot{\mathbf{v}} + \frac{\dot{a}}{a}\mathbf{v} + \frac{1}{a}(\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{a}\nabla\Phi + \frac{\hbar^2}{2m_{\rm FDM}^2 a^3}\nabla\left(\frac{\nabla^2\sqrt{\rho}}{\sqrt{\rho}}\right), \tag{88}$$

where the last term in Eq. (88) is the quantum pressure, sometimes called the quantum Bohm potential. In the linear regime $\delta \ll 1$, Fourier transforming and working to first order yields the linearised equation for the density contrast,

$$\ddot{\delta}_k + 2H\dot{\delta}_k = \left(4\pi G\bar{\rho} - \frac{\hbar^2 k^4}{4m_{\rm FDM}^2 a^4}\right)\delta_k\,. \tag{89}$$

The first term on the right-hand side is the familiar gravitational instability term, while the second is the quantum pressure contribution. The two balance at the *Jeans wavenumber*, defined by setting the bracket to zero. Evaluating this condition at matter–radiation equality one obtains [97]

$$k_{\rm J} = \frac{66.5\,{\rm Mpc}^{-1}}{(1+z_{\rm eq})^{1/4}}\left[\frac{\Omega_{\rm FDM}h^2}{0.12}\right]^{1/4}\left[\frac{m_{\rm FDM}}{10^{-22}\,{\rm eV}}\right]^{1/2}. \tag{90}$$

Modes with $k \gtrsim k_{\rm J}$ oscillate rather than grow, so the power spectrum is exponentially suppressed on scales below the corresponding Jeans mass. Equivalently, the effective sound speed of the FDM fluid is $c_s^2 \simeq \hbar^2 k^2/(4m_{\rm FDM}^2 a^2)$, which is scale-dependent and vanishes in the limit $m_{\rm FDM}\to\infty$, recovering the pressureless CDM limit. Because the linearised Eq. (89) does not capture non-linear effects and the mismatch between the Schrödinger–Poisson and fluid representations at high densities, in practice one employs an *effective-fluid approximation* whose parameters are calibrated against the full solution of the Klein–Gordon equation [75, 106]. We employ the fitting formula of [106],

$$T_{\rm F}(k) = \frac{\sin x^n}{x^n(1+Bx^{6-n})}\,, \qquad x = A\,\frac{k}{k_{\rm J}}\,, \tag{91}$$

where $A = 2.22\,m_{22}^{1/25+1/(1000\ln m_{22})}$, $B = 0.16\,m_{22}^{-1/20}$, and $m_{22}\equiv m_{\rm FDM}/10^{-22}\,{\rm eV}$. An additional feature of FDM that distinguishes it from CDM at the non-linear level is the formation of *solitonic cores* at the centres of DM halos. These are coherent, quasi-stationary solutions of the Schrödinger–Poisson system whose central density is roughly constant and whose radius scales inversely with the halo mass. While the mass within the soliton can be a non-negligible fraction of the total halo mass, the soliton is generally less dense than the NFW cusp that CDM simulations predict, and this difference in the central concentration has consequences for GW wave optics effects.

The second class of models features *Warm Dark Matter* (WDM), in which the DM particles are semi-relativistic at production, with a characteristic thermal mass $m_{\rm WDM} = \mathcal{O}({\rm keV})$. Unlike FDM, the suppression of structure in WDM is a purely classical, kinetic effect: the finite thermal velocity of the particles prevents them from clustering on scales smaller than their free-streaming length. At the level of the Boltzmann hierarchy, the WDM phase-space distribution $f(\mathbf{x},\mathbf{p},t)$ obeys

$$\frac{\partial f}{\partial t} + \frac{p_i}{a m_{\rm WDM}}\frac{\partial f}{\partial x^i} - m_{\rm WDM}\frac{\partial\Phi}{\partial x^i}\frac{\partial f}{\partial p_i} = 0\,, \tag{92}$$

i.e. the collisionless Boltzmann equation in the expanding universe. Unlike CDM, whose distribution function is effectively a cold delta function in momentum space, WDM possesses a thermal Fermi–Dirac spectrum. The non-zero velocity dispersion damps perturbations on scales below the free-streaming length

$$\lambda_{\rm fs} \sim \left(\frac{m_{\rm WDM}}{\rm keV}\right)^{-4/3} {\rm Mpc}\,, \tag{93}$$

which washes out structure on the corresponding scales. In a similar way as with FDM, computing the precise cut-off requires tracking the relativistic-to-non-relativistic transition through the full Boltzmann equations. This yields the well-known transfer function [27, 71, 140]

$$T_{\rm W}(k) = \left[1 + (\alpha k)^{2\mu}\right]^{-5/\mu}\,, \tag{94}$$

with $\mu = 1.12$ and

$$\alpha = 0.049 \left[\frac{m_{\rm WDM}}{\rm keV}\right]^{-1.11} \left[\frac{\Omega_{\rm DM}}{0.25}\right]^{0.11} \left[\frac{h}{0.7}\right]^{1.22} h^{-1}\,{\rm Mpc}\,. \tag{95}$$

Unlike the FDM Jeans scale, which depends on the field equations of motion and is therefore a dynamical cut-off that evolves with redshift, the WDM free-streaming length is set at production and is essentially a geometric property of the primordial phase-space distribution. Both physical mechanisms thus lead to the same qualitative consequence, a plateau in $\sigma^2(M)$ and a truncated halo mass function at low masses, but they differ in the slope of the transition, the shape of the transfer function, and the nature of the halo cores they produce.

In both models, the modified matter power spectrum takes the form

$$\mathscr{P}_{\rm sup}(k) = T_J(k)^2\, \mathscr{P}_{\rm CDM}(k)\,, \tag{96}$$

where $\mathscr{P}_{\rm CDM}(k)$ is computed using the transfer function of [49]. The practical effect is to flatten $\sigma(M)$ below the respective Jeans or free-streaming mass scale, suppressing halo formation in a regime that is precisely the target of the observational probes discussed in Sections 5.2 and 4.

3.3 Enhanced small-scale structures

Certain DM models predict the opposite behaviour: a significant *excess* of power on small scales relative to CDM. We consider two such scenarios that are relevant to the work in this thesis. In axion-like particle (ALP) models where the global $U(1)$ symmetry is broken after inflation, the Kibble mechanism generates large-amplitude small-scale density fluctuations that seed the formation of *axion mini-clusters* [85]. This mechanism imprints a white-noise contribution to the matter power spectrum [60, 62, 50], and the overall spectrum can be approximated as

$$\mathscr{P}_{\rm enh}(k) = \mathscr{P}_{\rm CDM}(k) + \mathscr{P}_{\rm CDM}(k_c)\,, \tag{97}$$

where the single parameter k_c encodes the scale above which the white-noise contribution dominates. For ALPs this characteristic scale is related to the particle mass by [77]

$$k_c \approx 3\,h\,{\rm Mpc}^{-1} \left[\frac{m_a}{10^{-18}\,{\rm eV}}\right]^{0.6}\,, \tag{98}$$

with an upper cutoff at $k_{\rm cut} \simeq 300\,{\rm Mpc}^{-1}(m_a/10^{-18}\,{\rm eV})^{1/2}$ set by the horizon scale at the onset of axion oscillations. Throughout the parameter space we explore, this cutoff lies well above k_c.

A similar white-noise enhancement arises in scenarios where a significant fraction of DM resides in *heavy primordial BHs* (PBHs). Each PBH acts as a discrete, massive seed that imprints a Poisson-noise contribution to the power spectrum [79, 44]. The characteristic scale of this contribution is determined by the mean comoving separation between PBHs,

$$k_c \approx 6\,h\,{\rm Mpc}^{-1}\left[\frac{f_{\rm PBH}\,m_{\rm PBH}}{10^4\,M_\odot}\right]^{-0.4}, \qquad k_{\rm cut} = 900\,h\,{\rm Mpc}^{-1}\left[f_{\rm PBH}\,\frac{10^4\,M_\odot}{m_{\rm PBH}}\right]^{1/3}, \tag{99}$$

where $f_{\rm PBH}$ is the DM fraction in PBHs. Below $k_{\rm cut}$, there is, on average, fewer than one PBH per comoving volume element, and the seed effect [37, 36, 34] drives the growth of DM overdensities. The white-noise regime is realised when $k_c < k_{\rm cut}$, which requires $f_{\rm PBH} > 10^{-4}(m_{\rm PBH}/10^4\,M_\odot)^{-0.09}$.

3.4 Spherical collapse

We have started showing the statistics of density perturbations and how they change for different DM models. Now, we will turn to how these densities can collapse into virialised objects and essentially decouple from the expansion. The simplest model for the formation of a gravitationally bound DM halo is the *spherical top-hat collapse* model. Consider a spherical region of uniform overdensity $\delta_i > 0$ embedded in a uniformly expanding background. The evolution of the shell at radius r enclosing mass M is governed by

$$\ddot{r} = -\frac{GM}{r^2} + \frac{\Lambda}{3}\,r, \tag{100}$$

where the cosmological constant term $\Lambda/3$ is negligible during matter domination. Combining this with energy conservation, one obtains the parametric solution in an Einstein–de Sitter (EdS) universe, $\Omega_m = 1$

$$r(\theta) = \frac{r_i}{2\delta_i}\,(1-\cos\theta), \tag{101}$$

$$t(\theta) = \frac{\pi}{2H_i(1+\delta_i)^{1/2}}\,\frac{\theta-\sin\theta}{(1+\delta_i)^{1/2}}, \tag{102}$$

where $\theta \in [0, 2\pi]$ is a development angle and r_i, H_i, δ_i are the initial radius, Hubble parameter and overdensity, respectively. The shell detaches from the Hubble flow at *turn-around* when $\dot{r} = 0$, which occurs at $\theta = \pi$, after which it re-collapses. Formally, the sphere reaches zero radius at $\theta = 2\pi$. The physical picture is that shell-crossing and violent relaxation convert the ordered infall into a virialised, quasi-equilibrium DM halo before the singularity is reached. By the virial theorem, the final virialised radius is $r_{\rm vir} = r_{\rm ta}/2$, where $r_{\rm ta}$ is the turn-around radius. The mean density inside the virialised object can be computed from the parametric solution, giving a characteristic overdensity

$$\Delta_c \equiv \frac{\rho_{\rm halo}}{\bar{\rho}(t_{\rm col})} \simeq 18\pi^2 \simeq 178 \tag{103}$$

in an EdS universe. In ΛCDM the virial overdensity is a weak function of redshift and cosmological parameters; a common approximation is $\Delta_c(z) \simeq 18\pi^2 + 82(\Omega_m(z)-1) - 39(\Omega_m(z)-1)^2$, which approaches the EdS value during matter domination.

The key quantity from the perspective of the Press-Schechter formalism is the *critical linear overdensity for collapse*, δ_c. If we linearly extrapolate the initial overdensity δ_i forward to the time $t_{\rm col}$ at which the sphere formally collapses, we obtain the overdensity that the density field *would have* if it had continued to evolve linearly. For an EdS universe one finds the remarkable result that this linearly extrapolated overdensity is a *universal constant*

$$\delta_c \equiv \delta_{\rm lin}(t_{\rm col}) = \frac{3}{20}(12\pi)^{2/3} \simeq 1.686\,. \tag{104}$$

This value is largely insensitive to the background cosmology: in ΛCDM, δ_c changes by less than 1% over the redshift range $0 \leq z \leq 5$. The universality of $\delta_c \simeq 1.686$ is the cornerstone of the Press-Schechter formalism. A region in the linear density field with $\delta_0(\mathbf{x}) > \delta_c/D(t)$ will have collapsed by time t to form a virialized DM halo; equivalently, using the convention in which the linear density field is held fixed and the collapse barrier evolves, $\delta_c(t) = \delta_c/D(t)$.

3.5 The Press-Schechter Formalism

With the spherical collapse model establishing a universal collapse threshold $\delta_c \simeq 1.686$, and the linear density field providing a probabilistic description of the initial conditions, Press & Schechter (1974) [112] proposed a remarkably elegant framework for computing the abundance of DM haloes as a function of mass and redshift.

3.5.1 The Basic postulate

Let $\delta_0(\mathbf{x})$ be the density field linearly extrapolated to the present day, and let $\delta(\mathbf{x};M)$ be the same field smoothed on a mass scale M using a top-hat filter. Since $\delta_0(\mathbf{x})$ is a Gaussian random field, the smoothed field $\delta(\mathbf{x};M)$ is also Gaussian at each point $\mathbf{x}$, with zero mean and variance $\sigma^2(M)$. The probability density function of δ at a given location is therefore

$$p(\delta;M)\,\mathrm{d}\delta = \frac{1}{\sqrt{2\pi}\,\sigma(M)} \exp\!\left(-\frac{\delta^2}{2\sigma^2(M)}\right) \mathrm{d}\delta\,. \tag{105}$$

The Press-Schechter postulate states that *the fraction of mass that at time t is locked up in haloes with mass greater than M is equal to the probability that the smoothed linear density field exceeds the collapse threshold $\delta_c(t)$*:

$$F(>M,t) = 2\,P[\delta(\mathbf{x};M) > \delta_c(t)] = 2\int_{\delta_c(t)}^{\infty} p(\delta;M)\,\mathrm{d}\delta = \mathrm{erfc}\!\left(\frac{\delta_c(t)}{\sqrt{2}\,\sigma(M)}\right), \tag{106}$$

where $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = (2/\sqrt{\pi})\int_x^\infty e^{-t^2}\mathrm{d}t$ is the complementary error function. The factor of two is a *fudge factor* introduced by Press & Schechter themselves to correct for the fact that, as noted above, integrating the Gaussian only over positive δ accounts for only half the total mass. Underdense regions are surrounded by larger overdense regions and will eventually be swept up into virialised objects; the factor of two is meant to account for this mass. Below, we will see that the excursion set formalism provides a rigorous derivation of this factor.

3.5.2 The Halo mass function

The HMF $n(M,t)$, the comoving number density of haloes with masses in the range $[M, M+\mathrm{d}M]$, follows from the Press-Schechter fraction by

$$n(M,t)\,\mathrm{d}M = \frac{\bar{\rho}}{M}\left|\frac{\partial F(>M,t)}{\partial M}\right| \mathrm{d}M\,, \tag{107}$$

where $\bar{\rho}$ is the mean matter density of the universe. Substituting Eq. (106) and using the chain rule,

$$\frac{\partial F}{\partial M} = \frac{\partial}{\partial \sigma}\mathrm{erfc}\left(\frac{\delta_c}{\sqrt{2}\sigma}\right) \cdot \frac{\mathrm{d}\sigma}{\mathrm{d}M} = \sqrt{\frac{2}{\pi}}\,\frac{\delta_c}{\sigma^2}\, e^{-\delta_c^2/(2\sigma^2)}\,\frac{\mathrm{d}\sigma}{\mathrm{d}M}\,, \tag{108}$$

so that

$$\boxed{n(M,t) = \sqrt{\frac{2}{\pi}}\,\frac{\bar{\rho}}{M^2}\,\frac{\delta_c}{\sigma(M)}\,\exp\left(-\frac{\delta_c^2}{2\sigma^2(M)}\right)\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right|}\,. \tag{109}$$

This is the *Press-Schechter mass function*. It is conventional to introduce the peak-height variable $\nu \equiv \delta_c(t)/\sigma(M)$ and the multiplicity function $f(\nu)$, such that

$$n(M,t)\,\mathrm{d}M = \frac{\bar{\rho}}{M^2}\,f_{\rm PS}(\nu)\left|\frac{\mathrm{d}\ln\nu}{\mathrm{d}\ln M}\right|\mathrm{d}M\,, \qquad f_{\rm PS}(\nu) = \sqrt{\frac{2}{\pi}}\,\nu\, e^{-\nu^2/2}\,. \tag{110}$$

Note that ν encodes both the mass and time dependences through $\delta_c(t) \propto D^{-1}(t)$ and $\sigma(M)$; the time evolution of the halo abundance is thus encoded entirely in the single parameter ν. Several important features follow immediately from Eq. (109):

- *Low-mass end ($M \ll M_*$, $\nu \ll 1$):* Since $\sigma(M)$ increases as M decreases in a CDM cosmology, $\nu \to 0$ as $M \to 0$, and $n(M) \propto M^{-2}$ modulo logarithmic corrections. The mass function rises steeply towards low masses, predicting an abundance of sub-galactic and sub-stellar haloes.

- *High-mass end ($M \gg M_*$, $\nu \gg 1$):* The exponential factor $e^{-\nu^2/2}$ causes an exponential suppression of the halo abundance above M_*. Massive haloes are rare precisely because the probability of finding a region with $\delta > \delta_c$ decreases rapidly with the mass scale.

- *Mass normalisation:* The mass fraction $\int_0^\infty f_{\rm PS}(\nu)\,\mathrm{d}\nu/\nu = 1$ (by construction, once the factor of 2 is included), ensuring that all matter is eventually locked up in collapsed haloes.

The factor of two in the Press-Schechter formalism, while empirically motivated, lacks a rigorous derivation within the original framework. Moreover, the PS ansatz suffers from an internal inconsistency known as the *cloud-in-cloud problem*: if a mass element at position $\mathbf{x}$ lies above the threshold δ_c when the field is smoothed on scale M_1, but below the threshold when smoothed on a larger scale $M_2 > M_1$, to which halo does it belong? These difficulties were resolved by Bond et al. (1991) [29] through the *excursion set formalism*, also known as *Extended Press-Schechter (EPS) theory*.

3.6 The extended Press Schechter formalism

The EPS ansatz states that *the fraction of mass elements whose trajectory first upcrosses the barrier δ_c at a value of S between S_1 and $S_1 + \mathrm{d}S$ equals the fraction of mass that at time t resides in haloes with masses between M and $M + \mathrm{d}M$, where M and S are related by* $S = \sigma^2(M)$. This prescription resolves the cloud-in-cloud problem elegantly. A trajectory that crosses the barrier at $S = S_1$ and then returns below it at $S > S_1$ is interpreted as a mass element that is incorporated into a halo of mass M_1 at time t and has not yet been incorporated into a larger halo. The first upcrossing picks out the *smallest* enclosing halo, which is the physically relevant one.

For a flat barrier δ_c and Markovian random walks, the first-upcrossing distribution $f_{\rm FU}(S,\delta_c)$ can be computed analytically. The key insight is that, for a Markovian walk, any trajectory that crosses δ_c at some $S_0 < S_1$ and then returns to $\delta < \delta_c$ at S_1 has an equally likely mirror image that *starts* above δ_c. The fraction of trajectories with first upcrossing between S and $S + \mathrm{d}S$ is

$$f_{\rm FU}(S,\delta_c)\,\mathrm{d}S = \frac{1}{\sqrt{2\pi}}\,\frac{\delta_c}{S^{3/2}}\,\exp\left(-\frac{\delta_c^2}{2S}\right)\mathrm{d}S\,. \tag{111}$$

This is related to the PS multiplicity function by $f_{\rm FU}(S,\delta_c) = \frac{1}{2S}\,f_{\rm PS}(\nu)$ with $\nu = \delta_c/\sqrt{S}$. Substituting into the EPS mass function

$$n(M,t)\,\mathrm{d}M = \frac{\bar{\rho}}{M}\,f_{\rm FU}(S,\delta_c)\left|\frac{\mathrm{d}S}{\mathrm{d}M}\right|\mathrm{d}M \tag{112}$$

reproduces exactly the Press-Schechter mass function Eq. (109), but *without any fudge factor*: the factor of two arises naturally from the method-of-images argument, which accounts for trajectories that were previously above the barrier (i.e., mass elements that were once part of a larger collapsed structure at an earlier time and have since been dispersed, only to re-collapse into a smaller halo).

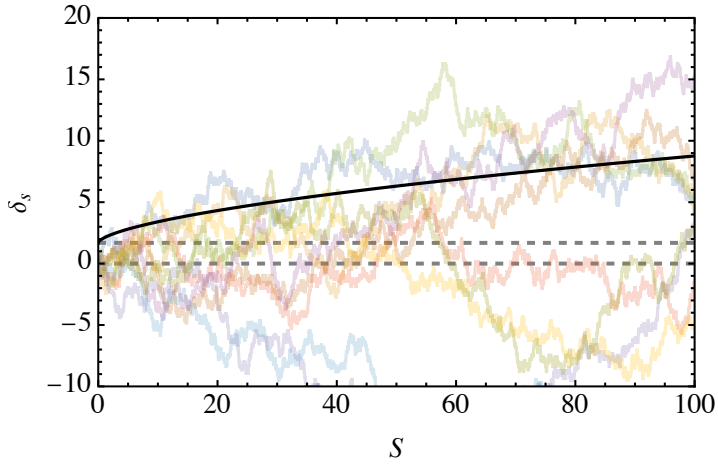


Figure 7: Random path demonstration to illustrate the EPS formalism, the dashed lines represent the spherical collapse and the moving barrier, the elliptical collapse.

To make the argument explicit, consider the ensemble of all random-walk trajectories, like in Fig.7 by looking at the dashed uniform barrier. A trajectory B that, at variance S_1, lies below the barrier but has previously crossed it at $S < S_1$, is a *missed* trajectory in the naive PS counting: the mass element appears to be uncollapsed at scale M_1, but it has a history of having been part of a larger halo. The mirror image B' of B reflected across δ_c is, by the Markovian property, equally probable. Counting B' alongside B effectively doubles the counting of trajectories that cross the barrier at S_1, precisely providing the factor of two.

3.7 Conditional probabilities

Although the EPS formalism is a convenient way to derive halo merger trees and thereby BH histories, it has its drawbacks, mainly that halos with masses below $\sim 10^8\,M_\odot$ are too

abundant and computationally expensive to include (although this can be improved by an adaptive resolution [131]). For this thesis, I have aimed to overcome these drawbacks by applying the EPS formalism directly to the evolution of halos without invoking the halo merger tree step, which breaks the math behind the EPS formalism.

As we have started by describing in this section, in CDM, the variance of matter perturbations $S(M)$ is a monotonically decreasing function of the halo mass M: $S(M) \to 0$ as $M \to \infty$. Thus, we can replace M by $S \equiv S(M)$ and the matter overdensity field δ around each spatial point evolves along a path in the $(S(M), \delta)$ plane that originates from $(0,0)$. If the path is not initiated in $(0,0)$ but instead at (S,δ), the probability to get to the point (S',δ'), $S' > S$ becomes $P(\delta'|\delta,S,S') = P(\delta'-\delta|S'-S)$. The probability for a path starting at (S,δ) to cross the threshold $\delta' = \delta_c(z') > \delta = \delta_c(z)$ at variance larger than S' is, in the same way as before but given by

$$P_{\rm FC}(\delta',S'|\delta,S) = 1 - {\rm erf}\left[\frac{\delta'-\delta}{\sqrt{2(S'-S)}}\right], \tag{113}$$

and the probability for the path to cross the threshold δ' in the range $(S', S'+{\rm d}S')$ is

$$\begin{aligned} p_{\rm FC}(\delta',S'|\delta,S){\rm d}S' &= \frac{{\rm d}P_{\rm FC}(\delta',S'|\delta,S)}{{\rm d}S'}{\rm d}S' \\ &= \frac{\delta'-\delta}{\sqrt{2\pi(S'-S)^3}} e^{-\frac{(\delta'-\delta)^2}{2(S'-S)}}\,{\rm d}S'. \end{aligned} \tag{114}$$

From this one can compute what fraction of the mass M of a halo at z on average comes from a halo of mass M' in the past, $z' > z$. If instead, we want to compute how the mass of the halo M' at z' is going to be distributed in heavier halos $M > M'$ in some later time $z < z'$, we need to invert the conditional probability (114). This gives the probability that a path that starts at (δ',S') crosses the threshold δ in the range $(S, S+{\rm d}S)$ (see Eq. (2.16) in [88]):

$$\begin{aligned} p_{\rm FC}(\delta,S|\delta',S'){\rm d}S &= p_{\rm FC}(\delta',S'|\delta,S)\frac{p_{\rm FC}(\delta,S|0,0)}{p_{\rm FC}(\delta',S'|0,0)}{\rm d}S \\ &= p_{\rm FC}(\delta',S'|\delta,S)\left(\frac{S'}{S}\right)^{\frac{3}{2}}\left(\frac{\delta}{\delta'}\right)e^{\frac{\delta'^2}{2S'}-\frac{\delta^2}{2S}}\,{\rm d}S. \end{aligned} \tag{115}$$

A change of variables then gives the probability that the halo whose mass at z' is M' ends up being a part of a halo in the mass range $(M, M+{\rm d}M)$ at $z < z'$:

$${\rm d}P(M,z|M',z') \equiv p_{\rm FC}(\delta,S|\delta',S')\left|\frac{{\rm d}S}{{\rm d}M}\right|{\rm d}M. \tag{116}$$

This allows us to compute the evolution of the HMF forward in time from $z' > z$ to z as

$$\frac{{\rm d}n(z)}{{\rm d}M} = \int_0^M {\rm d}M' \frac{{\rm d}n(z')}{{\rm d}M'}\frac{M'}{M}\frac{{\rm d}P(M,z|M',z')}{{\rm d}M}. \tag{117}$$

It is easy to check that this formula holds by inserting the HMF in the right-hand side and performing the integral. We will elaborate on a similar formula to track the evolution of baryons and SMBH inside galaxies.

3.8 Ellipsoidal collapse and the moving barrier

The spherical collapse model is a powerful conceptual tool, but it rests on the assumption of perfect spherical symmetry, which is never realized in a generic cosmological density field. In reality, the tidal field of the surrounding large-scale structure breaks spherical symmetry, and the collapse of a generic overdense region is expected to be *triaxial*, or *ellipsoidal*. While the PS mass function was in acceptable agreement with the N-body simulations available up to the late 1990s, the Millennium Simulation [127] and subsequent runs with improved statistics revealed systematic discrepancies: the PS mass function *over-predicts* the abundance of low-mass haloes and *under-predicts* the abundance of massive haloes, with the discrepancies growing at higher redshifts.

The ellipsoidal collapse model was developed by Zel'dovich [145] and subsequently cast in a form directly applicable to Press-Schechter theory [125]. The starting point is the observation that the initial tidal field at a given point can be decomposed into its eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ (with $\lambda_1 + \lambda_2 + \lambda_3 = \delta$), which determine the collapse rates along the three principal axes. Zel'dovich's approximation for the displacement field,

$$\mathbf{x}(\mathbf{q}, t) = \mathbf{q} - D(t)\,\nabla\phi(\mathbf{q}), \tag{118}$$

where $\mathbf{q}$ is the Lagrangian (initial) position and ϕ is the potential, predicts that collapse first occurs along the axis with the largest eigenvalue. The first collapse produces a sheet (*Zel'dovich pancake*), followed by collapse along a second axis to form a filament, and finally along the third axis to form a compact knot, i.e. the DM halo.

The ellipsoidal collapse of an object characterised by overdensity δ and tidal ellipticity e and prolateness p (defined from the eigenvalues of the tidal tensor) requires a higher overdensity at the moment of final collapse than the spherical case. To see this quantitatively, note that an oblate body with $\lambda_3 < \lambda_2 \sim \lambda_1$ has its longest axis resisting collapse; the extra tidal stretching must be overcome by a larger initial overdensity. Sheth, Mo & Tormen showed that the critical linear overdensity for ellipsoidal collapse is well approximated by

$$\delta_{\rm ec}(e, p, S) \approx \delta_c \left[1 + 0.47 \left(\frac{5(e^2 \pm p^2)}{\delta_c^2/\sigma^2} \right)^{0.615} \right], \tag{119}$$

where $e \equiv (\lambda_1 - \lambda_3)/(2\delta)$ and $p \equiv (\lambda_1 - 2\lambda_2 + \lambda_3)/(2\delta)$ are the ellipticity and prolateness of the tidal field. The critical overdensity is always larger than $\delta_c = 1.686$, and increases for more asymmetric configurations. In the spherical limit $e = p = 0$, Eq. (119) reduces to the standard result.

The distributions of e and p are determined by the statistics of the Gaussian density field. Since the tidal tensor is a symmetric 3×3 matrix, its eigenvalues follow from the joint distribution of a Gaussian random matrix (the Wigner-Dyson distribution). For a Gaussian random field, the distribution of e at fixed δ is

$$p(e|\delta) \propto e^2 \exp\left(-\frac{5e^2\delta^2}{2\sigma^2} \right), \tag{120}$$

and similarly for p. Averaging over the most probable values of e and p conditional on the variance $S = \sigma^2(M)$, Sheth, Mo & Tormen arrived at an effective *moving collapse barrier*:

$$\delta_{\rm ec}(S) \approx \delta_c \left[1 + 0.47 \left(\frac{S}{\delta_c^2} \right)^{0.615} \right], \tag{121}$$

where $S = \sigma^2(M)$ is the mass variance defined in Eq. (85). This is the central result of the ellipsoidal collapse model as applied to Press-Schechter theory, for reference it can

be compared to the spherical one in Fig.7. Crucially, the barrier $\delta_{\rm ec}(S)$ is *not constant* but increases as S increases (i.e., as halo mass decreases). Less massive structures, which reside in regions with larger mass variance, require a higher overdensity to collapse than more massive ones, because their tidal environments are more asymmetric on average. This *moving barrier* is qualitatively distinct from the flat barrier δ_c of the spherical collapse model, and as we shall see, its introduction substantially improves the agreement of the analytic mass function with numerical simulations.

It is instructive to compare the two barriers visually as in Fig.7. At the mass scale $M = M_*$ where $S = \delta_c^2$, both barriers coincide: $\delta_{\rm ec} = \delta_c = 1.686$. At a scale one dex below M_*, where $S \sim 5\delta_c^2$, the ellipsoidal barrier gives $\delta_{\rm ec} \simeq 1.686 \times [1 + 0.47 \times 5^{0.615}] \simeq 2.3$, compared to 1.686 for the spherical case. This $\sim 30\%$ upward shift in the collapse threshold suppresses the abundance of low-mass haloes relative to the spherical collapse prediction, a correction that is in better agreement with N-body simulations.

3.9 Ellipsoidal collapse and the Sheth-Mo-Tormen mass function

The natural improvement is to replace the flat spherical-collapse barrier with the moving ellipsoidal-collapse barrier $\delta_{\rm ec}(S)$ of Eq. (121). For a moving barrier, the first-upcrossing statistics can no longer be computed analytically, and Sheth, Mo & Tormen (2001) [125] used Monte Carlo random-walk simulations to determine the multiplicity function. Their result is well approximated by

$$f_{\rm EC}(\nu) = 0.322 \left(1 + \frac{1}{\tilde{\nu}^{0.6}}\right) f_{\rm PS}(\tilde{\nu}), \qquad \tilde{\nu} = 0.84\,\nu, \tag{122}$$

where the normalisation 0.322 is set by requiring $\int_0^\infty f_{\rm EC}(\nu)\,{\rm d}\nu/\nu = 1$. The resulting *Sheth-Mo-Tormen (SMT) mass function* is

$$n_{\rm SMT}(M,t) = \frac{\bar{\rho}}{M^2} f_{\rm EC}(\nu) \left|\frac{{\rm d}\ln\nu}{{\rm d}\ln M}\right|, \tag{123}$$

which compared to the PS prediction suppresses the abundance at low masses and enhances it at high masses, in much better agreement with N-body simulations. Specifically, when compared to the Millennium Simulation data using a friends-of-friends halo finder, the SMT mass function reduces the fractional residuals from ~ 30–50% (PS) to $\lesssim 10$–20% (SMT) across four decades in halo mass.

A particularly important physical interpretation of the SMT improvement concerns the role of the tidal environment. The moving barrier encodes the fact that less massive objects need to overcome a higher effective collapse threshold because their environments are, on average, more asymmetric and subject to stronger tidal shear. In physical terms, small-scale overdensities are embedded in larger-scale structures (filaments, sheets) that resist collapse along some directions; it is the competition between self-gravity and tidal stretching that sets the effective barrier height and thus the mass function at low masses.

The SMT mass function, along with various further calibrations against simulations, form the basis of modern semi-analytic and semi-numerical models of structure formation. Crucially, it is the HMF that sets the initial conditions for the semi-analytic models of SMBH seeding and growth that we employ in this thesis.

3.10 Improving the Sheth-Mo-Tormen approach

With the two collapse models and the range of DM power spectra in hand, we can now discuss the technical improvements to the EPS formalism that are developed and exploited

throughout this thesis. These improvements concern two related issues: the accurate computation of conditional first-crossing probabilities for a moving barrier, and the choice of a smoothing filter that performs well across both CDM and non-standard DM models. Although the improvement seems small, it is conceptually very important for different DM models because it allows us to compute the evolution of the halos in non-standard DM models in a systematic and unified way.

3.10.1 Conditional first-crossing distribution in the ellipsoidal model

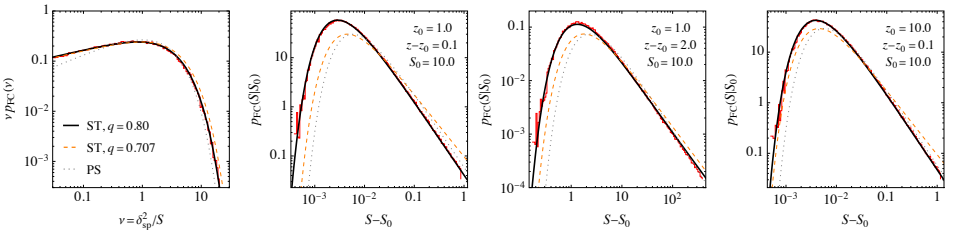


Figure 8: Left panel: *Unconditional first-crossing distribution* $\nu\, p_{\rm FC}(\nu)$ *for the spherical (PS) and ellipsoidal (SMT) barriers.* Other panels: *Conditional first-crossing distributions* $p_{\rm FC}(S|S_0)$ *for random walks starting at* $S_0 = 10$*, for different values of* z_0 *and* $z - z_0$*. The red histograms are Monte Carlo random walks; the black solid, orange dashed and grey dotted lines are the ellipsoidal-collapse estimates with* $q = 0.80$ *and* $q = 0.707$ *and the spherical-collapse (PS) result, respectively. Reproduced from Paper V [133],* © *ESO 2025.*

For the conditional mass function, we require first-crossing statistics for random walks that begin at a non-zero starting point (S_0, δ_0) with $\delta_0 = \delta_{\rm ell}(S_0, z_0)$. The simplest approach, used by Sheth, Mo & Tormen [125], is to shift the barrier and variance by $\delta_{\rm sp}(z) \to \delta_{\rm sp}(z) - \delta_{\rm sp}(z_0)$ and $S \to S - S_0$. However, as illustrated in Fig. 8, this ansatz does not accurately reproduce the distribution of first-crossings obtained from explicit Monte Carlo random-walk simulations, particularly when the redshift interval is large. A substantially better estimate is given by

$$p_{\rm FC}(S, z\,|\,S_0, z_0) = A' \left[1 + (a\nu)^{-p}\right] \sqrt{\frac{\nu'}{2\pi}}\, \frac{e^{-\nu'/2}}{S - S_0}\,, \tag{124}$$

where

$$\nu = \frac{\delta_{\rm sp}(z)^2}{S}\,, \qquad \nu' = \frac{\left[\delta_{\rm ell}(z, S_0) - \delta_{\rm ell}(z_0, S_0)\right]^2}{S - S_0}\,, \tag{125}$$

and the normalisation constant in the limit $z \to z_0$ is

$$A' = \left[1 + \left(a\, \frac{\delta_{\rm sp}(z_0)^2}{S_0}\right)^{-p}\right]^{-1}. \tag{126}$$

Here $a = 0.707$ and $p = 0.3$ as in the standard SMT barrier. This expression recovers the analytical estimate of [146] in the limit $z \to z_0$, and crucially also agrees well with numerical results for large redshift separations $z - z_0 \gg 0$, a regime not captured by the earlier approximation. From the first-crossing probabilities one constructs the probability density that a halo of mass M' at redshift z' ends up as part of a halo of mass $M > M'$ at a later time $z < z'$,

$$\frac{{\rm d}\tilde{P}(M, z\,|\,M', z')}{{\rm d}M} = p_{\rm FC}(S', z'\,|\,S, z)\, \frac{p_{\rm FC}(S, z)}{p_{\rm FC}(S', z')} \left|\frac{{\rm d}S}{{\rm d}M}\right|, \tag{127}$$

which in turn gives the mean halo mass accretion rate

$$\frac{\mathrm{d}M}{\mathrm{d}z} = \sqrt{\frac{2}{\pi}} \frac{\mathrm{d}\delta_{\mathrm{ell}}(S(M), z)}{\mathrm{d}z} \left| \frac{\mathrm{d}S}{\mathrm{d}M} \right|^{-1} \sqrt{S(M) - S(2M)} \,. \tag{128}$$

This estimate improves upon that of [41] in two respects: it incorporates the correction from ellipsoidal rather than spherical collapse, and it is based directly on the expected halo mass growth rather than the expected mass of the main progenitor. Both estimates agree to within a factor of two across a broad range of halo masses and redshifts.

3.10.2 A new smoothing filter

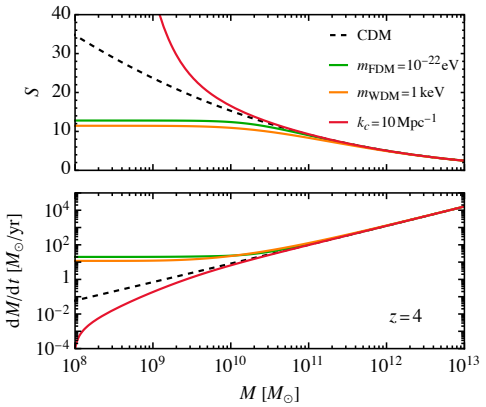


Figure 9: Variance $S = \sigma^2$ of matter perturbations (upper panel) and halo growth rate (lower panel) in the models considered in this work: CDM (dashed black line), FDM (green line), WDM (orange line), and an enhanced matter power spectrum (red line). Reproduced from Paper V [133], © ESO 2025.

A second improvement concerns the choice of window function used to define $\sigma^2(M)$. The standard real-space top-hat filter is natural for CDM, but it performs poorly for DM models with a modified power spectrum: its sharp boundary in configuration space gives rise to unphysical oscillations in $\sigma(M)$ when the power spectrum is truncated, which in turn distort the predicted HMF at masses near the suppression scale. We therefore adopt the smooth filter proposed by [89, 139],

$$\tilde{W}(k\,|\,R) = \frac{1}{1 + (c_1 k R)^{c_2}} \,, \tag{129}$$

where R is the filter scale related to the enclosed mass by $M = 4\pi\bar{\rho}R^3/3$. The two free parameters c_1 and c_2 determine the shape of the filter in Fourier space. In models with a suppressed power spectrum, the resulting HMF scales as $\mathrm{d}n/\mathrm{d}\ln M \propto M^{c_2/6}$ on scales well below the suppression mass, and the halo growth rate satisfies $\mathrm{d}M/\mathrm{d}z \propto M^{1-c_2/6}$. We fix $c_2 = 6$ throughout, which ensures that the halo growth rate increases with mass in all models considered, including WDM and FDM. The parameter $c_1 = 0.43$ is fixed by requiring that the variance $\sigma^2(M)$ computed with this filter agrees with that obtained from the standard top-hat filter in CDM, consistent with the simulation calibrations [120].

This choice also provides a good match to the HMF found in FDM simulations [120], though it slightly underestimates the halo mass at which the WDM HMF departs from the CDM prediction [91]. An alternative choice $c_1 = 1/3.3$, suggested by [89], recovers

the low-mass WDM behaviour but fails to match CDM predictions at high masses. As a practical resolution, we retain $c_1 = 0.43$ in the window function and compensate for the offset in the WDM case by rescaling the parameter α in the WDM transfer function by a factor of $3.3\,c_1 \approx 1.4$, which yields a consistent description across the full mass range.

The consequences of these choices for the variance $\sigma^2(M)$ and the halo mass accretion rate are illustrated in Fig. 9. In FDM and WDM, the variance asymptotes to a constant at low masses, reflecting the hard suppression of small-scale power; the transition is sharper in FDM, where the Jeans scale provides a more abrupt cutoff than the free-streaming length of WDM particles. In models with enhanced perturbations, the variance grows rapidly toward low masses, driving faster early growth of the smallest haloes. The resulting HMFs at various redshifts are shown in Fig. 10. For suppressed-power models, the HMF grows as $\mathrm{d}n/\mathrm{d}\ln M \propto M$ below the suppression scale, while for enhanced models the imprint of the white-noise contribution is progressively diluted at lower redshifts as mini-clusters merge into more massive host haloes.

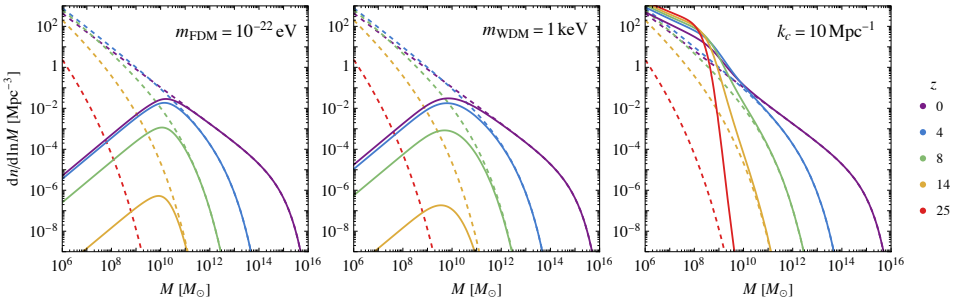


Figure 10: HMF at various redshifts z in the FDM (left) and WDM (middle) models, and in models where white noise dominates the matter perturbations at $k > k_c$ (right). The dashed contours show the HMF in the CDM model for comparison. Reproduced from Paper V [133], © ESO 2025.

3.11 The halo merger rate

The EPS formalism not only predicts the abundance of haloes as a function of mass and redshift, but, as we have shown, can be used to derive many more properties that we are going to explore, starting with the halo merger-rate. The classical derivation, due to Lacey & Cole [88], proceeds from the spherical-collapse conditional first-crossing probability and yields the following expression for the probability per unit time that a halo of mass M_1 accretes a companion of mass M_2 to form a halo of mass $M_f = M_1 + M_2$:

$$\frac{\mathrm{d}p(M_1, M_2, t)}{\mathrm{d}t\,\mathrm{d}M_2} = \frac{1}{M_f}\sqrt{\frac{2}{\pi}}\left|\frac{\dot{\delta}_c}{\delta_c}\right|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M_f}\left[1 - \frac{\sigma^2(M_f)}{\sigma^2(M_1)}\right]^{-3/2} \times \frac{\delta_c}{\sigma(M_f)}\exp\left[-\frac{\delta_c^2}{2}\left(\frac{1}{\sigma^2(M_f)} - \frac{1}{\sigma^2(M_1)}\right)\right]. \tag{130}$$

It is customary to absorb the ratio of this quantity to the HMF of the companion into a *merger rate kernel*,

$$Q(M_1, M_2, t) \equiv \frac{\mathrm{d}p(M_1, M_2, t)}{\mathrm{d}t\,\mathrm{d}M_2}\left[\frac{\mathrm{d}n(M_2, t)}{\mathrm{d}M_2}\right]^{-1}, \tag{131}$$

so that the differential halo merger rate per unit comoving volume takes the compact form

$$\frac{\mathrm{d}R_h}{\mathrm{d}M_1\,\mathrm{d}M_2} = \frac{\mathrm{d}n(M_1,t)}{\mathrm{d}M_1}\,\frac{\mathrm{d}n(M_2,t)}{\mathrm{d}M_2}\,Q(M_1,M_2,t)\,. \tag{132}$$

As noted by [23], the kernel Q is not symmetric under the exchange $M_1 \leftrightarrow M_2$, an artifact of the asymmetry inherent in the EPS conditional probability. Nevertheless, [59] demonstrated that restricting to $M_1 < M_2$ — i.e. always treating the lighter halo as the accreted object — closely reproduces the symmetrised result of the Benson-Kamionkowski-Hassani (BKH) theory [23], which derives the merger kernel from the Smoluchowski coagulation equation and explicitly preserves the EPS HMF under time evolution. We adopt this convention throughout.

While the Lacey-Cole result has proven to be a remarkably useful workhorse, it rests entirely on the spherical collapse barrier $\delta_c = \mathrm{const}$. As we have seen in the previous sections, the more realistic ellipsoidal collapse model leads to a mass-dependent, moving barrier $\delta_{\mathrm{ell}}(S,z)$, and the improvements to the conditional first-crossing distribution described in Section 3.10 make it possible to extend the merger rate to this setting in a fully consistent way. The derivation starts from the inverse conditional probability: given a halo of variance S at redshift z, what is the rate at which it transitions into a larger halo of variance $S_0 < S$ at a slightly later time? Applying Bayes' theorem to the conditional first-crossing distributions of Eqs. (114) and (124) yields

$$\frac{\mathrm{d}P_{\mathrm{FC}}(S_0,z_0\,|\,S,z)}{\mathrm{d}S_0} = \frac{\mathrm{d}P_{\mathrm{FC}}(S,z\,|\,S_0,z_0)}{\mathrm{d}S}\,\frac{\mathrm{d}P_{\mathrm{FC}}(S_0,z_0)/\mathrm{d}S_0}{\mathrm{d}P_{\mathrm{FC}}(S,z)/\mathrm{d}S} \equiv B\,\sqrt{\nu'}\,, \tag{133}$$

where

$$B = \frac{A'}{\sqrt{2\pi}}\,\frac{[1+(a\nu)^{-p}]\,[1+(q\nu_0)^{-p}]}{[1+(q\nu)^{-p}]}\,\frac{\sqrt{S}\,e^{-(\nu'+q\nu_0-q\nu)/2}}{S_0\,(S-S_0)}\,, \tag{134}$$

with $\nu_0 = \delta_{\mathrm{sp}}(z_0)^2/S_0$ and ν' as defined in Eq. (124). Expanding in the limit $z_0 \to z$ gives the instantaneous *transition rate*

$$\frac{\mathrm{d}P(S\to S_0\,|\,z)}{\mathrm{d}S_0\,\mathrm{d}z} = \frac{B(z_0=z)}{\sqrt{S-S_0}}\left|\frac{\mathrm{d}\delta_{\mathrm{ell}}(z,S_0)}{\mathrm{d}z}\right|, \tag{135}$$

which is proportional to the rate of change of the collapse barrier with redshift and is suppressed as $(S-S_0)^{-1/2}$ for small mass-ratio mergers. Converting to halo masses and cosmic time,

$$\frac{\mathrm{d}P(M\to M_0\,|\,z)}{\mathrm{d}M_0\,\mathrm{d}t} = (1+z)\,H(z)\left|\frac{\mathrm{d}S_0}{\mathrm{d}M_0}\right|\frac{\mathrm{d}P(S\to S_0\,|\,z)}{\mathrm{d}S_0\,\mathrm{d}z}\,. \tag{136}$$

The improved merger rate then follows directly, with the parent mass identified as $M_p = \min(M_1,M_2)$ and the final mass as $M_f = M_1 + M_2$:

$$\frac{\mathrm{d}R_h}{\mathrm{d}M_1\,\mathrm{d}M_2} = \frac{\mathrm{d}n(M_p,z)}{\mathrm{d}M_p}\,\frac{\mathrm{d}P(M_p\to M_f\,|\,z)}{\mathrm{d}M_f\,\mathrm{d}t}\,. \tag{137}$$

This expression reduces to the BKH-consistent Lacey-Cole result in the limit of a flat spherical barrier, and extends it in two respects: it incorporates the ellipsoidal collapse correction through δ_{ell}, and it uses the improved conditional first-crossing probability of Eq. (124), which as shown in Section 3.10 accurately reproduces Monte Carlo random-walk results even for large redshift separations. The dependence on the DM model enters entirely through $\sigma^2(M)$ and its logarithmic slope.

3.12 The halo dark matter accretion rate

Now we want to turn our attention to another important quantity, the rate at which a DM halo accretes DM. Within the EPS framework, this *mean* accretion rate can be derived directly from the conditional first-crossing probability without constructing explicit merger trees, by asking how much mass a halo of mass M is expected to accumulate in an infinitesimal redshift step $\mathrm{d}z$. The expected mass increase of a halo of mass M at redshift z over a step $\mathrm{d}z$ is obtained by integrating the inverse conditional probability Eq. (136) over all final masses $M_0 > M$:

$$\frac{\mathrm{d}M}{\mathrm{d}z} = \int_M^{2M} \mathrm{d}M_0 \, (M_0 - M) \, \frac{\mathrm{d}P(M \to M_0 \,|\, z)}{\mathrm{d}M_0 \, \mathrm{d}z} \,. \tag{138}$$

Evaluating this integral in the limit $\mathrm{d}z \to 0$, using the transition rate of Eq. (136), yields the closed-form expression

$$\frac{\mathrm{d}M}{\mathrm{d}z} = \sqrt{\frac{2}{\pi}} \, \frac{\mathrm{d}\delta_{\mathrm{ell}}(S(M), z)}{\mathrm{d}z} \left| \frac{\mathrm{d}S}{\mathrm{d}M} \right|^{-1} \sqrt{S(M) - S(2M)} \,. \tag{139}$$

Converting to physical time gives the halo mass accretion rate

$$\dot{M}_{\mathrm{halo}}(M, z) = (1+z) \, H(z) \sqrt{\frac{2}{\pi}} \, \frac{\mathrm{d}\delta_{\mathrm{ell}}(S(M), z)}{\mathrm{d}z} \left| \frac{\mathrm{d}S}{\mathrm{d}M} \right|^{-1} \sqrt{S(M) - S(2M)} \,. \tag{140}$$

This result generalises the estimate of [41] in two important respects. First, it uses the ellipsoidal barrier δ_{ell} rather than the spherical threshold $\delta_c = \mathrm{const}$, ensuring consistency with the SMT mass function used throughout. Second, it is derived from the *expected* mass increase of the halo, whereas the Correa et al. estimate is inferred from the growth of the most massive progenitor in Monte Carlo trees, which can differ systematically at high redshift.

The DM model dependence of Eq. (140) enters entirely through the variance $\sigma^2(M)$ and its mass derivative. In CDM, $S(M)$ is a monotonically decreasing function and $\dot{M}_{\mathrm{halo}}$ increases with M at fixed z. In WDM and FDM, $\sigma^2(M)$ asymptotes to a constant at masses below the free-streaming or Jeans scale, so that the factor $\sqrt{S(M) - S(2M)} \to 0$ and small haloes grow more slowly than their CDM counterparts. Conversely, in models with an enhanced small-scale power spectrum — such as axion miniclusters or scenarios with a significant PBH fraction — the variance rises rapidly at small M, driving accelerated early growth of low-mass structures.

3.13 Evolution of baryons and black holes through mergers

Our final derivation, and one of the most important aspects of this thesis, is the rate at which a DM halo accretes baryons, stars, or SMBHs, because the EPS conditional probability derived in Section 3.7 does not only describe how the DM HMF evolves; it also provides a self-consistent framework for tracking the mean baryonic and BH content of haloes as they merge hierarchically. The key observation is that the baryonic matter— stars, gas, and the central SMBH — shares the same merger history as the DM halo that hosts it.

Consider a population of haloes of mass M' at redshift z', each hosting a mean mass $M_J(M', z')$ of component $J \in \{\mathrm{BH}, *, \mathrm{gas}\}$. As structure formation proceeds from z' to a later time $z < z'$, these progenitor haloes are incorporated into larger haloes of mass M. The mean mass of component J in a halo of mass M at redshift z receives a contribution

from mergers given by

$$M_J(M,z)\big|_{\rm merg} = \left[\frac{{\rm d}n(z)}{{\rm d}M}\right]^{-1} \int_0^M {\rm d}M' \, \frac{{\rm d}n(z')}{{\rm d}M'} \, \frac{{\rm d}P(M,z\,|\,M',z')}{{\rm d}M} \, M_J(M',z') \,, \tag{141}$$

where ${\rm d}P(M,z\,|\,M',z')/{\rm d}M$ is the inverse conditional probability that a halo of mass M' at z' ends up as part of a halo of mass M at $z < z'$, derived in Eq. (115) for the case of spherical collapse. Substituting explicitly,

$$\begin{aligned} M_J(M,z)\big|_{\rm merg} = & \int_0^M {\rm d}M' \left|\frac{{\rm d}S}{{\rm d}M'}\right| \frac{M}{M'} \, M_J(M',z') \\ & \times \frac{\delta_c(z') - \delta_c(z)}{\sqrt{2\pi[S(M') - S(M)]^3}} \exp\left(-\frac{[\delta_c(z') - \delta_c(z)]^2}{2[S(M') - S(M)]}\right) , \end{aligned} \tag{142}$$

For a better estimate, the ellipsoidal-collapse conditional probability of Eq. (133) should be used instead; we do not write out the resulting expression explicitly. The instantaneous merger growth rate is obtained by taking the limit $z' \to z$:

$$\boxed{\dot{M}_J^{\rm merg}(M,z) = (1+z)\,H(z) \lim_{\Delta z \to 0} \frac{M_J(M,z)\big|_{\rm merg} - M_J(M,z)}{\Delta z}} \tag{143}$$

which is driven by the same transition rate that governs the halo merger rate of Eq. (136), evaluated at the baryonic content of the progenitors rather than their DM mass.

Several features of Eq. (142) deserve emphasis. The integral runs over all progenitor masses $M' < M$, weighted by the EPS probability that a halo of mass M' at z' is incorporated into a halo of mass M by redshift z. The prefactor M/M' accounts for the fact that each progenitor of mass M' contributes a fraction M'/M of the final halo mass, so the integral correctly conserves the total baryonic mass budget. Another way to understand it is that the new halo gains the properties of all the progenitors, which are lighter and are going to be accreted. The DM model dependence enters through the variance $\sigma^2(M)$: in WDM or FDM, the suppression of $S(M)$ at small masses narrows the progenitor distribution and delays the hierarchical assembly of baryonic structure, while enhanced small-scale power broadens it and accelerates early merging. In this way, the same modification to the matter power spectrum that shapes the HMF and merger rate propagates directly and self-consistently into the assembly histories of stars, gas and SMBHs.

4 Wave optics lensing of gravitational waves

We have started this thesis with an introduction to GWs, its phenomenology, and then we moved to the EPS formalism and structure formation. Now we will explore a consequence of the two. GWs interact with matter, but more specifically with DM in a very interesting way. Gravitational lensing is one of the most direct consequences of GR: the spacetime curvature produced by a massive object deflects the trajectories of all propagating fields, including GWs. For electromagnetic radiation, except for extreme cases, the wavelength is so short compared to the Schwarzschild radius of any astrophysical lens that ray optics provides a complete description; image positions, magnifications, and time delays are all encoded in the geometrical path.

For GWs the situation is qualitatively different. The wavelengths accessible to current and planned detectors range from fractions of a kilometre to fractions of a light-year, which are comparable to, or even larger than, the Schwarzschild radius of compact DM structures. In this regime, diffraction and interference become essential, and the full wave treatment must be applied. On the phenomenological side, electromagnetic sources in the universe are incoherent, whereas GW sources tend to be highly coherent. This allows us to also look for precision, tiny deviations in the waveform induced by DM structure. This section reviews the theoretical foundations of wave-optics gravitational lensing of GWs and its application to DM searches.

4.1 The amplification function

Let us consider a GW propagating in the weak gravitational field of a massive object. The background metric is

$$\mathrm{d}s^2 = -(1+2U)\,c^2\,\mathrm{d}t^2 + (1-2U)\,\mathrm{d}r^2\,, \tag{144}$$

where $U(r)$ ($|U| \ll 1$) is the Newtonian gravitational potential of the lens. The GW is again treated as a linear perturbation $h_{\mu\nu}$ on this background. Under the transverse-traceless Lorenz gauge ($h^{\mu}{}_{\mu} = 0$, $h^{\mu\nu}{}_{;\nu} = 0$), the perturbation satisfies

$$h_{\mu\nu;\alpha}{}^{;\alpha} + 2R^{\alpha}{}_{\mu}{}^{\beta}{}_{\nu}\,h_{\alpha\beta} = 0\,, \tag{145}$$

where the semicolons denote covariant derivatives with respect to the background metric $g^{(0)}_{\mu\nu}$, and $R^{\alpha}{}_{\mu}{}^{\beta}{}_{\nu}$ is the background Riemann tensor. In the limit where the GW wavelength λ is much smaller than the curvature radius of the background, the Riemann term is negligible, $R \approx 0$. Writing $h_{\mu\nu} = \phi\, e_{\mu\nu}$, where $e_{\mu\nu}$ is a polarisation tensor that is parallel-transported along the null geodesic [100], the change in polarisation due to lensing is of order $U \ll 1$ and can be neglected. This means that we will only consider the effect on the amplitude of the GW, and we will not need to work in the polarisation basis. The amplitude ϕ then satisfies the scalar wave equation

$$(\nabla^2 + \omega^2)\phi = 4\omega^2 U\phi\,, \tag{146}$$

where $\omega = 2\pi f$ is the angular frequency of the GW in the observer frame. This reduction to a scalar problem is valid whenever the fractional change in polarisation per wavelength is small compared to ϕ itself, which is amply satisfied for any astrophysical lens.

We solve Eq. (146) using the Kirchhoff diffraction formalism. The position vector of the source in the source plane is η; the impact parameter in the lens plane is ξ. We adopt the *thin-lens approximation*, in which the lens is characterised solely by its projected surface mass density $\Sigma(\xi)$ and the GW is deflected only at the lens plane. Defining the volume V

enclosed between the lens plane and a large sphere centred on the observer, and applying Green's theorem to the Helmholtz equation $(\nabla^2+\omega^2)\phi=0$ that holds within V (since $U=0$ there by assumption), when $\phi = Ae^{iS_P}$, the lensed amplitude at the observer is

$$\phi_{\rm obs}(\omega,\eta) = \frac{\omega A}{2\pi i D_L}\int \mathrm{d}^2\xi\, \exp[i\omega t_d(\xi,\eta)]\,, \tag{147}$$

where the *arrival-time surface* $t_d(\xi,\eta)$ is

$$t_d(\xi,\eta) = \frac{D_L D_S}{2c\,D_{LS}}\left(\frac{\xi}{D_L}-\frac{\eta}{D_S}\right)^2 - \frac{1}{c}\hat{\psi}(\xi)+t_{\rm ref}(\eta)\,. \tag{148}$$

The first term is the geometrical time delay arising from the longer path of the deflected ray; the second is the Shapiro delay produced by propagation through the lens potential. The deflection potential $\hat{\psi}(\xi)$ is related to the surface mass density by

$$\nabla^2_\xi \hat{\psi}(\xi) = \frac{8\pi G}{c^2}\Sigma(\xi)\,. \tag{149}$$

The reference time $t_{\rm ref}(\eta)$ is a constant chosen so that t_d is zero for the unlensed ray path. In a cosmological setting, the results above remain valid provided the angular-diameter distances are used, and the frequency ω is replaced by $\omega(1+z_L)$, where z_L is the redshift of the lens [19]. The lensed amplitude then reads

$$\phi_{\rm obs}(\omega,\eta) = \frac{\omega(1+z_L)D_S}{2\pi i D_L D_{LS}}\int \mathrm{d}^2\xi\, \exp[i\omega(1+z_L)t_d(\xi,\eta)]\,. \tag{150}$$

It is natural to define the *amplification factor* as the ratio of the lensed to the unlensed amplitude:

$$F(\omega,\eta) \equiv \frac{\phi^{\rm lensed}_{\rm obs}(\omega,\eta)}{\phi^{\rm unlensed}_{\rm obs}(\omega,\eta)}\,. \tag{151}$$

From Eq. (150) this is

$$\boxed{F(\omega,\eta) = \frac{\omega(1+z_L)D_S}{2\pi i D_L D_{LS}}\int \mathrm{d}^2\xi\, \exp[i\omega(1+z_L)t_d(\xi,\eta)]\,,} \tag{152}$$

normalised so that $|F|=1$ in the no-lens limit ($U=\hat{\psi}=0$). This is the amplification factor, and it is going to be the most important object in GW lensing, since the lensed waveform in the frequency domain is then simply

$$\tilde{h}(f) = F(f)\,\tilde{h}_0(f)\,, \tag{153}$$

where $\tilde{h}_0(f)$ is the unlensed waveform. It is convenient to introduce dimensionless variables. Choosing a reference length ξ_0 in the lens plane, we define the dimensionless impact parameter and source position:

$$x = \frac{\xi}{\xi_0}\,, \qquad y = \frac{D_L\eta}{\xi_0 D_S}\,, \tag{154}$$

a dimensionless time delay

$$T(x,y) = \frac{D_L D_S}{D_{LS}\,\xi_0^2}\,c\,t_d\big(\xi_0 x,\, \xi_0 D_S y/D_L\big) = \frac{1}{2}\,|x-y|^2 - \psi(x) + \phi_m(y)\,, \tag{155}$$

and a dimensionless frequency

$$w = \frac{D_S}{D_L D_{LS}} \frac{\xi_0^2\, \omega(1+z_L)}{c} = \frac{8\pi G M_{Lz} f}{c^3}\,, \tag{156}$$

where $M_{Lz} = M_L(1+z_L)$ is the redshifted lens mass and we have chosen $\xi_0 = \xi_E = (4GM_{Lz}D_L D_{LS}/D_S)^{1/2}$ (the Einstein radius) as the reference length. The dimensionless lensing potential $\psi(x)$ satisfies

$$\nabla_x^2 \psi = \Sigma(|\xi_0 x|)/\Sigma_{\rm cr}\,, \tag{157}$$

$$\Sigma_{\rm cr} = D_S/(4\pi G D_L D_{LS})\,, \tag{158}$$

where $\Sigma_{\rm cr}$ is the critical surface density. The constant $\phi_m(y)$ is chosen so that $\min_x T = 0$. In these units the amplification factor takes the elegant form

$$F(w,y) = \frac{w}{2\pi i}\int \mathrm{d}^2x\, \exp[iw\,T(x,y)]\,. \tag{159}$$

The single parameter w controls the relative importance of wave vs. geometric effects: for $w \gg 1$, the phase varies rapidly, and the integral is dominated by the stationary points, recovering geometric optics; for $w \ll 1$ the phase is nearly uniform and the integral converges to unity, meaning the lens starts to become transparent to long-wavelength GWs. For axially symmetric lenses, $\psi(x) = \psi(|x|)$, and the azimuthal integral can be performed analytically, reducing Eq. (159) to a single integral over the radial coordinate $x = |x|$:

$$F(w,y) = -iw\, e^{iw(y^2/2+\phi_m(y))}\int_0^\infty \mathrm{d}x\, x J_0(wxy)\, \exp\left[iw\left(\tfrac{1}{2}x^2 - \psi(x)\right)\right], \tag{160}$$

where J_0 is the Bessel function of zeroth order and $y = |y|$. The integrand is a rapidly oscillating function for large w, which requires careful numerical treatment. This oscillatory behaviour is the key to separating wave-optics from geometric-optics.

4.1.1 Geometrical-optics limit

In the limit $w \gg 1$, the stationary-phase approximation to Eq. (159) is valid. The stationary points of $T(x,y)$ satisfy

$$\nabla_x T = 0 \quad \Longleftrightarrow \quad y = x - \nabla_x \psi(x)\,, \tag{161}$$

which is the *lens equation* of standard gravitational lensing; it is nothing but Fermat's principle applied to the time-delay surface. Each solution x_j corresponds to one image. Expanding $T(x,y)$ to second order around x_j and performing the resulting Gaussian integral gives

$$F_{\rm geo}(w,y) = \sum_j |\mu_j|^{1/2} \exp\left(iwT_j - in_j\frac{\pi}{2}\right), \tag{162}$$

where $T_j = T(x_j,y)$ is the time delay of the j-th image, $\mu_j = 1/\det(\partial y/\partial x_j)$ is its magnification, and $n_j \in \{0, 1/2, 1\}$ is the Morse index, equal to 0 if x_j is a minimum, $1/2$ if it is a saddle, and 1 if it is a maximum of the time-delay surface. In the time domain, Eq. (162) becomes

$$\phi^*(t,r) = \sum_j |\mu_j|^{1/2}\, \phi_0(t - t_{d,j},\, r)\,, \tag{163}$$

i.e. the sum of time-delayed, magnified copies of the unlensed waveform. The oscillatory behaviour of $|F(w,y)|$ at high w arises entirely from the interference between these copies, and its period encodes the time delay $\Delta t = (D_S\xi_0^2/D_L D_{LS}) \times \Delta T$, where ΔT is the difference in dimensionless time delays between the dominant pair of images. Now that we understand the amplification function, we can turn to understanding different lens models.

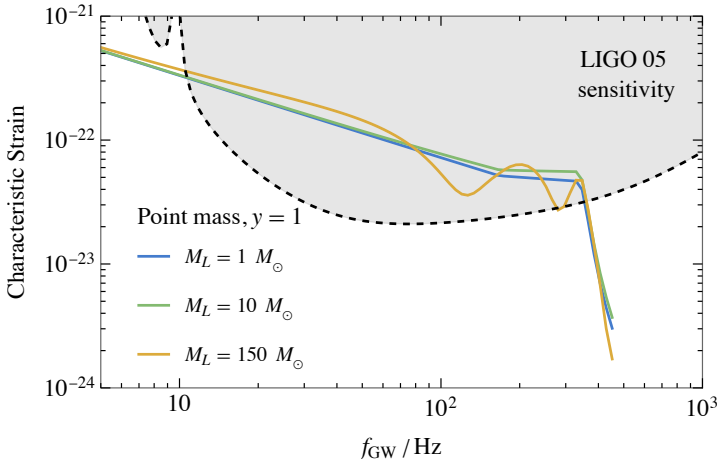


Figure 11: Illustrative example on the same binary as in Fig.4, but now lensed by point masses of different masses at impact parameter $y = 1$. We can see how the lens produces a diffractive increase in the amplitude, a wave optics oscillation and finally the geometric optics interference between paths, the case of a $1\,M_\odot$ lens is indistinguishable by eye from the unlensed waveform.

4.1.2 Point-mass lens

The easiest lens profile is the size-less point-mass lens. The point-mass lens has surface density $\Sigma(\xi) = M_L\,\delta^{(2)}(\xi)$ and dimensionless lensing potential $\psi(x) = \ln x$. This physically means that the mass distribution can be parameterised by a Dirac delta. The normalisation length is the Einstein radius $\xi_0 = \xi_E$. In this case the radial integral in Eq. (160) can be evaluated analytically [129]

$$F(w,y) = \exp\left[\frac{\pi w}{4} + \frac{iw}{2}\left(\ln\frac{w}{2} - 2\phi_m(y)\right)\right]\Gamma\left(1-\frac{iw}{2}\right){}_1F_1\left(\frac{iw}{2};1;\frac{iwy^2}{2}\right), \tag{164}$$

where Γ is Euler's gamma function, ${}_1F_1$ is the confluent hypergeometric (Kummer) function, and

$$\phi_m(y) = \frac{(y+x_m)^2}{2} - \ln x_m\,, \qquad x_m = \frac{y+\sqrt{y^2+4}}{2} \tag{165}$$

is the value of the dimensionless time delay at the image of minimum delay. The amplification factor depends on only two parameters, which should not be surprising given that the lens has infinitesimal size: the dimensionless frequency w and the dimensionless source position y. In the geometric-optics limit ($w \gg 1$), Eq. (164) reduces to Eq. (162) with the two image magnifications

$$\mu_\pm = \pm\frac{1}{2} + \frac{y^2+2}{2y\sqrt{y^2+4}}\,, \tag{166}$$

and the dimensionless time delay

$$\Delta T = \frac{y\sqrt{y^2+4}}{2} + \ln\frac{\sqrt{y^2+4}+y}{\sqrt{y^2+4}-y}\,. \tag{167}$$

The point-mass lens model applies to any compact object whose physical size is much smaller than the Einstein radius, including stellar-mass BHs, neutron stars, and PBHs. Now,

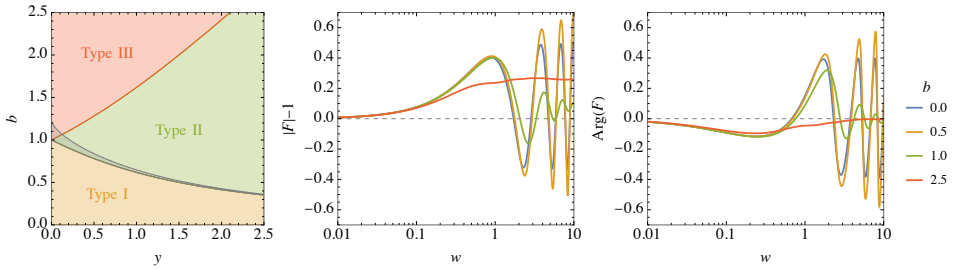


Figure 12: The left panel illustrates the three lensing regimes of a uniform sphere described in the main text. The modulus and argument of the amplification function F are shown in the middle and right panels, respectively, as a function of the dimensionless frequency w for $y = 1$ and various lens sizes b. The limiting case $b = 0$ recovers the point mass lens, while $b = 0.5$, $b = 1$, and $b = 2.5$ correspond to type I, type II, and type III lensing, respectively, as defined in the left panel. Reproduced from Paper I [61], © IOP Publishing Ltd and Sissa Medialab 2023.

we will turn to a constant-density sphere, which will help us understand how lensing behaves for extended lens objects.

4.1.3 Uniform density sphere

The simplest extended lens model beyond a point mass is a sphere of uniform density. Such an object is fully specified by its total mass M_l and physical radius R, with a density profile of the form

$$\rho(r) = \frac{3M_l}{4\pi R^3}\,\theta(R-r)\,, \tag{168}$$

where θ denotes the Heaviside step function. Projecting this density onto the lens plane and solving Eq. (149), one obtains the dimensionless deflection potential

$$\psi(x) = \begin{cases} \frac{1}{3}\sqrt{1-\frac{x^2}{b^2}}\left(\frac{x^2}{b^2}-4\right) + \ln\left(b+\sqrt{b^2-x^2}\right)\,, & x \le b\,, \\ \ln x\,, & x > b\,. \end{cases} \tag{169}$$

Here $b \equiv R/\xi_0$ is the dimensionless sphere radius and $\xi_0 = \xi_E$ is chosen as the Einstein radius. The potential naturally reduces to the point-mass result $\psi = \ln x$ outside the sphere, as expected from Gauss's theorem applied to the projected mass distribution.

Since the potential outside the sphere is identical to that of a point mass, the amplification function can be computed by combining the known analytical expression for the exterior region with a numerical evaluation of the integral only over the interior $x \le b$. The dimensionless frequency retains the same form as in the point-mass case, $w = 8\pi M_{lz} f$.

The structure of the amplification function is governed by the number and nature of the extremal paths of the time-delay surface. Paths crossing the lens plane at $x > b$ behave exactly as in the point-mass case, while additional extremal paths threading through the interior of the sphere may or may not exist depending on the geometry. A systematic analysis reveals three distinct regimes, illustrated in Fig. 12:

- **Type I**: if

$$0 < b \le 1 \quad \text{and} \quad y \le \frac{1-b^2}{b} \tag{170}$$

  two extremal paths cross the lens plane outside the sphere and one additional path passes through its interior. This configuration supports interference between three images.

- **Type II**: if

$$b > 0 \quad \text{and} \quad y > \frac{|1-b^2|}{b} \tag{171}$$

a single exterior extremal path exists. In a narrow region just above the lower bound on y, two further paths cross the interior of the sphere, but they produce only a subdominant contribution to the interference pattern.

- **Type III**: if

$$b > 1 \quad \text{and} \quad y \leq \frac{b^2-1}{b} \tag{172}$$

there is one interior extremal path. For $b < \sqrt{3/2}$ two additional interior paths appear in a small sub-region of this regime.

The observational consequences of these three regimes are illustrated in the middle and right panels of Fig. 12. For compact spheres falling in the Type I regime, the amplification function closely resembles that of a point mass, with the interior path capable of producing an additional boost in $|F|$, as visible when comparing the $b = 0$ and $b = 0.5$ curves. In the Type II regime the interference effect is weaker, since only a single exterior path contributes oscillatory behaviour, and $|F|$ saturates to a constant deep in the geometric-optics limit. In the Type III regime, all extremal paths are interior and the lens causes only a mild, smoothly varying amplification with no pronounced oscillatory pattern. Nevertheless, even this featureless frequency dependence could in principle permit detection through matched filtering. The Type III behaviour marks a natural bridge between the analytically tractable point-mass model and the more physically motivated NFW profiles considered next.

4.1.4 Navarro–Frenk–White halo

The density profile proposed by Navarro, Frenk and White [104] provides a more realistic description of DM halos than the uniform sphere. It is characterised by a scale radius r_s that separates a central cusp scaling as $\rho \propto r^{-1}$ from an outer envelope falling as $\rho \propto r^{-3}$:

$$\rho(r) = \frac{r_s^3 \rho_s}{r\,(r_s + r)^2}\,. \tag{173}$$

The cylindrical mass enclosed within projected radius r in the lens plane takes the form

$$M_{\rm cyl} = 4\pi\rho_s r_s^3 \left[\ln\left(\frac{r}{r_s}\right) + \frac{r_s}{\sqrt{r_s^2 - r^2}}\,\operatorname{arcoth}\left(\frac{r_s}{\sqrt{r_s^2 - r^2}}\right)\right], \tag{174}$$

and yields the dimensionless deflection potential

$$\psi(x) = \kappa\left[\ln^2\left(\frac{x}{2b}\right) - \operatorname{artanh}^2\sqrt{1 - \frac{x^2}{b^2}}\right], \tag{175}$$

where $\kappa \equiv 2\pi\rho_s r_s^3/M_{\rm v}$ encodes the normalised central concentration and $b \equiv r_s/\xi_0$ is the dimensionless scale radius. Throughout, distances are scaled by $\xi_0 = R_E(M_v)$, the Einstein radius evaluated at the halo virial mass M_v, and the corresponding dimensionless frequency is $w = 8\pi(1+z_l)M_{\rm v} f$. The two NFW parameters r_s and ρ_s are not independent: given the virial mass $M_{\rm v}$, both are fixed by the halo concentration relation, so κ is determined by $M_{\rm v}$ alone, while b carries an additional dependence on the source and lens distances through ξ_0.

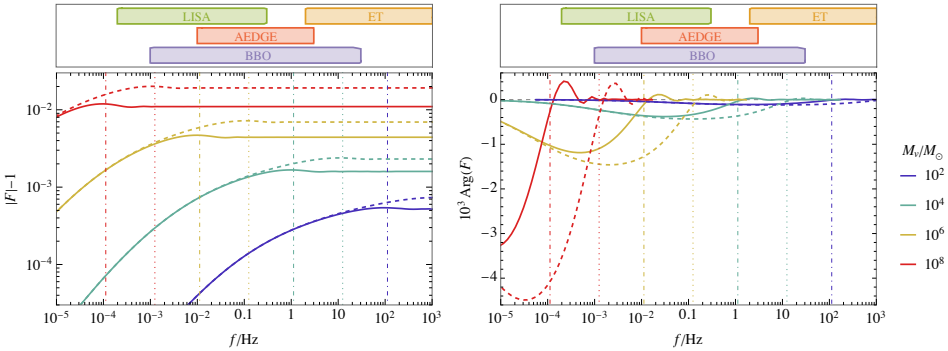


Figure 13: The modulus and argument of the amplification function F for an NFW halo lens are shown as a function of frequency for various halo virial masses. Solid and dashed curves correspond to $y = 1$ and $y = 0.3$, respectively, with source and lens luminosity distances fixed to $D_{\rm sL} = 5$ Gpc and $D_{\rm lL} = 2.5$ Gpc. Vertical lines mark the frequency at which $|F|$ is maximised. The frequency bands relevant to different experiments are indicated for reference. Reproduced from Paper I [61], © IOP Publishing Ltd and Sissa Medialab 2023.

The amplification function is evaluated numerically using the strategy of [128], with results displayed for a range of virial masses in Fig. 13. For $M_{\rm v} < 10^{10} M_\odot$ one typically finds $\kappa \lesssim 0.3$ and $b > 2$, a regime in which the time-delay surface admits only a single extremal path, located approximately at $x \approx y$. The resulting amplification function mirrors the Type III behaviour of the uniform-sphere model: the modulus $|F|$ rises smoothly with frequency and shows no clear interference oscillations. In the geometric-optics limit, this single-image configuration means that the NFW halo acts primarily as a magnifier, shifting the apparent luminosity distance to the source without introducing detectable fringe patterns. The transition frequency from the wave-optics to the geometric-optics regime occurs at $w_{\rm geo} \approx 2/y^2$, corresponding to a physical frequency

$$f_{\rm geo} \approx \frac{1}{4\pi y^2 (1+z_{\rm l}) M_{\rm v}}\,, \tag{176}$$

which also locates the peak of $|F|$ visible as the vertical lines in Fig. 13. This peak defines the characteristic scale of the lensing signature. Ground-based detectors such as LIGO and the Einstein Telescope, which are sensitive above ~ 1 Hz, cannot resolve the frequency-dependent modulation of the NFW lensing signal for halos with $M_{\rm v} < 10^4 M_\odot$, except in the unlikely case of a near-perfect alignment with $y \ll 1$. At still lower halo masses the induced magnification, of order $\mathcal{O}(10^{-3})$ for realistic impact parameters, falls well below the detection threshold of any current or near-future ground-based instrument. The natural arena for NFW halo lensing therefore shifts to space-based and atom-interferometry detectors operating below 1 Hz, namely LISA, AEDGE, and BBO, which are sensitive to the wave-optics regime for halos with $M_{\rm v} \gtrsim 10^4 M_\odot$. We will return to this point in the second half of the lensing discussion, in which we turn to the cumulative effects of the lenses instead of the single lens effects.

4.2 Single lens effects: Computing the cross section

Having established the theoretical framework for wave-optics lensing and characterised the amplification factor for different lens models, we now turn to the question of detectability: what is the probability that a given GW event is lensed in a way that leaves an

observable imprint? Since all GWs are subject to some degree of lensing as they propagate across cosmological distances, we restrict our attention to configurations where the lensing signature is both significant and measurable. Assuming a Poisson distribution of lenses along the line of sight, the probability that a given event is lensed is

$$P_l = 1 - e^{-\tau}, \tag{177}$$

where τ is the optical depth associated with the lensing cross-section, defined as

$$\tau(z_s) = \int_0^{D_s} \mathrm{d}V\, n(D_l) \frac{\sigma(D_l)}{4\pi D_l^2} = \int_0^{z_s} \mathrm{d}z_l \frac{n(z_l)\sigma(z_l)}{(1+z_l)H(z_l)}, \tag{178}$$

where $\mathrm{d}V = 4\pi D_l^2 \mathrm{d}D_l$ is the differential comoving volume element, D_l and z_l are the angular diameter distance and redshift of the lens respectively, n is the number density of lenses, and H is the Hubble expansion rate. The key ingredient encoding detectability is the lensing cross-section,

$$\sigma = \pi \xi_{\rm max}^2 = \pi \xi_0^2 y_{\rm max}^2, \tag{179}$$

where $\xi_{\rm max} = \xi_0 y_{\rm max}$ is the maximum impact parameter for which the lensing effect remains detectable. The threshold $y_{\rm max}$ can be determined either via the mismatch method or through a Fisher matrix analysis, as introduced in the preceding section.

The first approach for computing the maximal cross-section is via the mismatch method that we introduced in the first section. The idea is to find the maximal value of y as a function of the lens parameters so that we can separate the waveform from the unlensed one up to some confidence level. To quantify this more precisely, we consider the lensed GW signal $\tilde{\phi}_{\rm opt}(f) = F(f;\theta_{l,\rm opt})A(f;\theta_{\rm opt})e^{-i\Psi(f;\theta_{\rm opt})}$, where $F(f;\theta_{l,\rm opt})$ is the lensing amplification factor evaluated at the optimal lens parameters $\theta_{l,\rm opt}$. We then assess how well this signal can be recovered by an unlensed template by evaluating Eq. (57), minimising over the four unlensed waveform parameters. In practice, we find that the minimum is always attained at $\mathscr{M} \approx \mathscr{M}_{\rm opt}$ and $t_c \approx t_{c,\rm opt}$, so the scan reduces to two parameters: $\lambda \equiv (D_L/D_{L,\rm opt})^{-1}$ and $\delta\phi_c \equiv \phi_c - \phi_{c,\rm opt}$. The discriminating statistic then takes the form

$$\Delta\chi^2 = 4 \min_{\lambda,\,\delta\phi_c} \int \mathrm{d}f \frac{A(f;\theta_{\rm opt})^2 \left|F(f;\theta_{l,\rm opt}) - \lambda e^{-i\delta\phi_c}\right|^2}{S_{\rm n}(f)}. \tag{180}$$

Note that when the merger falls outside the observatory's sensitivity window, $\Delta\chi^2 \propto \mathscr{M}^{5/3}/D_L^2$. Another approach, which we also introduced at the beginning, is to do a Fisher forecast. Now, the idea is that we compute the Fisher matrix including the lens parameters. For the point-mass lenses in the geometric optics limit, approximate analytical expressions for the measurement errors were derived in [129]:

$$\begin{aligned} \frac{\Delta M_{lz}}{M_{lz}} &= \frac{1}{\rm SNR} \frac{\sqrt{y(y^2+2)}\,(y^2+4)^{\frac{5}{4}}}{2\tau}, \\ \frac{\Delta y}{y} &= \frac{1}{\rm SNR} \frac{\sqrt{y^2+2}\,(y^2+4)^{\frac{3}{4}}}{2\sqrt{y}}, \end{aligned} \tag{181}$$

where $\tau = \Delta t_d/(4M_{lz})$. A conclusion emerges from both approaches, although it is clearer with the second method that the lens parameters can be measured inversely proportional to the SNR of the signal.

Now that we have explored the importance of the cross section, we can move to the lens distribution and connect it to the structure formation. If the lens is the fundamental element of DM, if for instance DM is made of BH or other compact objects, then

$$n(z_l) = (1+z_l)^3 f_{\rm PBH}\rho_{\rm DM}/m_{\rm PBH}\,, \tag{182}$$

where $\rho_{\rm DM}$ is the density of DM, $m_{\rm PBH}$ is the mass of the lens and $f_{\rm PBH}$ is the fraction of DM in the form of these compact objects; as we will see, this is an interesting scenario to explore with ground detectors like LIGO. In the case that the lenses you are interested in finding are the DM halos, then the optical depth becomes

$$\tau(z_s) = \int_0^{z_s} {\rm d}z_l \int {\rm dlog}_{10}M_{\rm v}\,\frac{{\rm d}n}{{\rm dlog}_{10}M_{\rm v}}\,\frac{\sigma(z_l,M_{\rm v})}{(1+z_l)H(z_l)}\,, \tag{183}$$

where one integrates over the HMF instead of the constant-density DM distribution. Now that we have the probability for lensing, one can take two different approaches. One can take an existing catalogue of binaries for which we know that no lensing effect has been detected and compute bounds on the DM abundance, or make prospects for future GW detectors. For the latter case, we can integrate over the expected merger rate:

$$\bar{N}_l(m_{\rm PBH}, f_{\rm PBH}) = \mathscr{T}\int {\rm d}\lambda\, P_l(\theta_s)\, p_{\rm det}\left(\frac{{\rm SNR}_c}{{\rm SNR}(\theta_s)}\right)\,, \tag{184}$$

where $\mathscr{T}$ is the observation time, $p_{\rm det}$ is the detection probability [67], SNR is the signal-to-noise ratio [63], and

$${\rm d}\lambda = \frac{1}{1+z_s}\frac{{\rm d}R}{{\rm d}m_1{\rm d}m_2}\frac{{\rm d}V_c}{{\rm d}z_s}{\rm d}m_1{\rm d}m_2{\rm d}z_s \tag{185}$$

is the differential merger rate that we introduced before. If we instead want to get constraints from existing observations, we should compute how many of those observations should have been lensed, and by assuming that the distribution of the lensing events should be Poisson, we can get bounds. The expected number of lensed events from an existing set of observations is given by

$$\bar{N}_l(m_{\rm PBH}, f_{\rm PBH}) = \sum_{j=1}^{N} P_l(\theta_j)\,, \tag{186}$$

where N is the number of observed events. This is what I will use to constrain DM models using existing observations from LIGO.

5 Massive black holes in our universe

As we have established in the introduction, BHs have evolved from a mathematical curiosity into one of the central pillars of observational cosmology. The centres of virtually all massive galaxies harbour SMBHs, objects with masses in the range 10^6–$10^{10}\,M_\odot$ [86], and the correlations that the SMBH properties have with the host galaxy have a clear implication: whatever mechanism assembled galaxies also, inextricably, assembled their central BHs.

Understanding these observations requires a coherent theoretical picture connecting the cosmological growth of DM halos, the star-formation history of their baryonic contents, and the co-evolution of the central BH. We therefore structure this section around two complementary regimes. In the local universe, the SMBH population is largely assembled, and the scaling relations between BH mass, stellar mass, and halo mass are well established; these relations, combined with the EPS halo merger rate, provide the astrophysical input to the GW background relevant for PTAs and future high-frequency GW observatories. The local existence and properties of SMBH are tightly connected to the observations by the Hubble Space Telescope (HST). In the following section, we turn to the high-redshift universe, a far more dynamic epoch in which mergers cannot be disentangled from the very formation and growth of SMBHs, and whose exploration has been transformed by a newer telescope, the JWST.

5.1 Local Universe

While theoretical arguments and the energetics of quasars and AGNs had long pointed to the existence of compact, accreting objects with masses $M_{\rm BH} \sim 10^6$–$10^{10}\,M_\odot$, direct dynamical confirmation required resolving the gravitational sphere of influence of the BH, defined as

$$r_{\rm infl} = \frac{GM_{\rm BH}}{\sigma^2}, \tag{187}$$

where σ is the stellar velocity dispersion of the host bulge, the central galactic region that surrounds the BH. For typical SMBH masses in nearby galaxies at distances $D \sim 1$–20 Mpc, $r_{\rm infl}$ subtends only $\sim 0.1''$–$1.0''$ on the sky [86]. This angular scale placed the earliest dynamical searches firmly at the resolution limit of ground-based optical spectroscopy, making the first detections possible only from the mid-1980s, when charge-coupled devices became routinely available on high-dispersion spectrographs.

The tight correlations between the stellar mass and BH mass that we review in this section are derived from a sample of galaxies in which both the BH mass and the bulge mass can be measured independently with high precision. For active galaxies hosting broad-line AGN, the BH mass can be estimated through *virial techniques*: the broad-line region (BLR) is assumed to be virialised under the gravity of the BH, so that $M_{\rm BH} \propto R_{\rm BLR}\,\Delta V^2/G$, where ΔV is inferred from the width of a broad emission line (typically Hα or Hβ) and the BLR radius is calibrated through reverberation-mapping campaigns via the radius–luminosity relation [24]. Dynamical measurements, on the other hand, can be obtained via *stellar dynamical modelling*: spectroscopic measurements of the absorption-line-of-sight velocity distribution in the nuclear region are fit with axisymmetric or triaxial Jeans models or orbit-superposition Schwarzschild models to infer the central mass concentration. They can also rely on *gas dynamical modelling*, in which the rotation curve of a nuclear emission-line disc is fit to a Keplerian model dominated by the SMBH potential. The third and most precise method, where applicable, is *megamaser astrometry*: water masers orbiting in a thin, edge-on, Keplerian disc around the central BH—as in the prototypical case of NGC 4258—can be monitored with very-long-baseline interferometry (VLBI) to measure the orbital

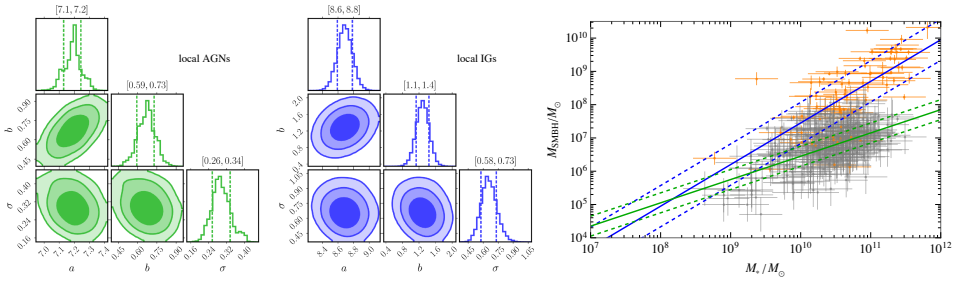


Figure 14: The left panel shows the posteriors of the fit of Eq. (188) *to the local AGNs of Ref. [118], and the central panel the posteriors of the fit to the local dynamically measured SMBHs in inactive galaxies (IGs) [86]. In the right panel we show in orange the dynamically measured SMBHs and in grey the local AGNs, compared with their respective fits. Adapted from Paper IX [55], © ESO 2024.*

acceleration directly, yielding $M_{\rm BH}$ with sub-percent precision [86].

In the landmark study exploiting $\sim 67{,}000$ emission-line galaxies from the Sloan Digital Sky Survey (SDSS) at $z < 0.055$, Reines & Volonteri [118] assembled a sample of 244 broad-line Seyferts selected via narrow-line diagnostic diagrams (BPT classification), together with 79 galaxies with dynamical BH masses and a supplementary set of dwarf galaxies extending the sample down to $M_\star \lesssim 10^{9.5}\,M_\odot$. We can see these observations summarised in Fig. 14. We will fit the probability distribution of the SMBH mass, m, for a given stellar mass, M_*, by the magnitude, a, and logarithmic slope, b, of the mean and the width, σ, of the distribution:

$$\frac{{\rm d}P(m|M_*,\theta)}{{\rm d}\log_{10}m} = \mathcal{N}\left(\log_{10}\frac{m}{M_\odot}\middle| a + b\log_{10}\frac{M_*}{10^{11}M_\odot}, \sigma\right), \tag{188}$$

where $\theta = (a, b, \sigma)$ and $\mathcal{N}(x|\bar{x}, \sigma)$ denotes the probability density of a Gaussian distribution with mean $\bar{x}$ and variance σ^2. As we anticipated, the heaviest SMBH are the dynamically measured ones, while the AGNs are the lightest and have a much shallower relation with respect to stellar mass. The posterior on the fit, as well as a comparison to the data, can be seen in Fig. 14. This is going to be a key step for being able to predict the merger rate of the local SMBHs. In order to do that, we should connect the stellar mass to the DM halo mass, since, as we have laid out, the fundamental merger rate comes from the merger of the DM halos, which triggers the subsequent merger of the galaxies.

5.1.1 PTA merger rate model

The connection between a galaxy's stellar mass $M_\star$ and the mass M_h of its host DM halo is encoded in the *stellar-to-halo mass relation* (SHMR), which provides a practical and physically motivated bridge between the outputs of dark matter N-body simulations and observable galaxy statistics. The foundational method for deriving the relation statistically is *halo abundance matching*: under the assumption that stellar mass increases monotonically with halo mass, a one-to-one correspondence between halos and galaxies is established by rank-ordering the cumulative HMF $n(> M_h)$ and the cumulative galaxy stellar mass function $\Phi(> M_\star)$, and equating the two at each number density [69].

Once the SHMR has been derived in each redshift bin through this matching procedure, it is convenient to describe it with an analytic fitting function. Following Moster et

al. [101], Girelli et al. [69] adopt a double power-law parameterisation of the form

$$\frac{M_\star}{M_h} = 2N \left[\left(\frac{M_h}{M_1}\right)^{-\beta} + \left(\frac{M_h}{M_1}\right)^{\gamma} \right]^{-1}, \tag{189}$$

where M_1 is a characteristic halo mass scale, N sets the overall normalisation (the peak efficiency), β governs the slope at halo masses below M_1, and γ controls the decline at halo masses above M_1 [69]. The ratio $M_\star/M_h$ therefore has a bell-shaped dependence on M_h: it rises as a power law at low masses, peaks at $M_h \sim M_1$, and falls off again toward cluster scales. Across all redshifts from $z = 0$ to $z \sim 4$ the peak efficiency $M_\star/M_h$ is always well below the cosmic baryon fraction, reaching at most $\sim 5\%$ at the halo mass $M_h \sim 10^{12}\, M_\odot$ where star formation is most efficient [69].

As we have discussed, the SHMR in the form of Eq. (189) serves as the key ingredient connecting the halo merger rate, computed from the extended EPS formalism, to the galaxy and SMBH merger rate. The scatter in the stellar mass halo mass relation, log-normal with a dispersion of $\sigma_{\log M_\star} \sim 0.15$ dex at fixed halo mass, propagates into the predicted SMBH mass function and ultimately into the amplitude and shape of the predicted stochastic GW background, a source of systematic uncertainty. So given a halo of virial mass M_v at redshift z, the probability distribution of hosting a central SMBH of mass $m_{\rm BH}$ is encoded in the occupation fraction $p_{\rm occ}(m_{\rm BH}|M_v, z)$. Combining the halo–stellar mass relation $M_*(M_v)$ with the empirical stellar-mass–BH-mass relation, and modelling the intrinsic scatter as a log-normal distribution, yields

$$p_{\rm occ}(m|M,z) = \frac{1}{\sqrt{2\pi}\,\sigma\, m} \exp\left[-\frac{\left(\ln m - \ln \bar{m}(M_*)\right)^2}{2\sigma^2} \right]. \tag{190}$$

Once we have this we can compute the SMBH merger rate by changing the variables of the halo merger rate in the following way

$$\frac{\mathrm{d}R_{\rm BH}}{\mathrm{d}m_1\,\mathrm{d}m_2} = p_{\rm BH}(m_1, m_2, z) \int \mathrm{d}M_1\,\mathrm{d}M_2\, \frac{\mathrm{d}R_h}{\mathrm{d}M_1\,\mathrm{d}M_2}\, p_{\rm occ}(m_1|M_1,z)\, p_{\rm occ}(m_2|M_2,z), \tag{191}$$

where we have added $p_{\rm BH}$, which is the probability that a halo merger results in the eventual coalescence of the two central SMBHs. This parameter encodes a great deal of complex microphysics: following the merger of their host halos, the two SMBHs must first sink to the centre of the new potential well through dynamical friction on the stellar and gas background, form a bound binary at separations of order a parsec, and harden further by three-body scattering against the stellar loss cone, and finally coalesce by emitting GWs. Each of these stages can be affected by environmental processes, and the overall efficiency of the sequence is captured by $p_{\rm BH}$. High merger efficiency, $p_{\rm BH} \simeq 1$, might be a good approximation if the SMBH is surrounded by a nuclear star cluster, which is very effective at lowering the orbital timescale [102]. However, it might be too optimistic for a more general case and should be considered an effective upper value. Studies of the efficiency of mergers induced by BH or three-body effects indicate a reduced BH merger efficiency $p_{\rm BH} \sim \mathcal{O}(0.1)$ [147], although with a big uncertainty.

Now that we have the merger rate, we can turn to computing a stochastic background. The stochastic GW background at nanohertz frequencies detected by PTAs arises from the incoherent superposition of GW emission by all SMBH binaries across cosmic time. The GW energy density per logarithmic frequency interval contributed by such a population is

$$\Omega_{\rm GW}(f) = \frac{1}{\rho_c} \int \frac{\mathrm{d}\lambda}{1+z}\, \frac{4\pi D_L^2}{c}\, \frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}\ln f_r}, \tag{192}$$

where D_L is the luminosity distance, $f_r = (1+z)f$ is the rest-frame frequency, and $\mathrm{d}\lambda$ is the differential merger rate in comoving volume. For binaries driven purely by GW emission, the characteristic strain spectrum follows a power law $h_c \propto f^{-2/3}$, corresponding to $\Omega_{\rm GW} \propto f^{2/3}$ in the frequency range probed by PTAs. As we saw in the first section, where we discussed PTA results, the NANOGrav 15-year data set [10] shows evidence for a spectral tilt shallower than $f^{2/3}$ at low frequencies, which can be partially explained by environmental coupling: at separations larger than the GW-dominated regime, the binary loses energy to dynamical friction with gas, stars, and DM, modifying the effective hardening timescale and thus the spectral shape. This is natural since the timescale from GWs can easily reach the Hubble time. What this implies physically is that the two halos merge, the galaxies can also slowly merge, but the final binary never gets to the separation that is small enough for the GW emission to become efficient, and the merger never happens. Now we will review the processes that make the binary lose energy at different separations.

5.1.2 The binary timescale

As the two host halos merge and the central SMBHs sink toward the common potential well, their orbital energy is dissipated through a sequence of physically distinct mechanisms, each dominant over a characteristic range of binary separation [82], as illustrated in Fig. 15. The dynamics at each stage can be captured through a single effective timescale, defined as $t_{\rm eff} \equiv |E|/\dot{E}$, where E is the total orbital energy of the binary. Because the energy loss from GW emission and from the surrounding environment proceed simultaneously, the effective timescale satisfies

$$t_{\rm eff}^{-1} = t_{\rm GW}^{-1} + t_{\rm env}^{-1}, \tag{193}$$

where $t_{\rm GW}$ and $t_{\rm env}$ denote the characteristic timescales of GW-driven and environmentally-driven orbital decay, respectively. For a circular orbit, the binding energy is $E = -m_1 m_2/d$, where d is the binary separation, so the separation evolves as

$$\dot{d} = -d\, t_{\rm eff}^{-1}. \tag{194}$$

At separations of order kiloparsecs, the binary is not yet gravitationally bound and its dynamics are governed by *dynamical friction*: the SMBHs experience a gravitational drag as they plough through the surrounding stellar and gaseous medium, transferring angular momentum to the background and causing the orbit to shrink. The associated timescale is [26]

$$t_{\rm dyn} \simeq \frac{20\,\mathrm{Myr}}{\ln\Lambda}\,\frac{\sigma}{200\,\mathrm{km\,s^{-1}}}\left[\frac{M}{10^9\,M_\odot}\right]^{-1}\left[\frac{d}{\mathrm{kpc}}\right]^{2}, \tag{195}$$

where $\ln\Lambda$ is the Coulomb logarithm and σ is the stellar velocity dispersion of the host galaxy. Once the binary becomes sufficiently hard, that is, once its binding energy per unit mass exceeds the typical kinetic energy of the surrounding stars, dynamical friction becomes inefficient, because the binary sweeps out the reservoir of stars on intersecting orbits faster than it can be replenished. The orbital decay then transitions to the regime of *loss-cone scattering*, in which stars on nearly radial orbits pass close to the binary and are ejected, carrying away energy and angular momentum. The associated hardening timescale is [113]

$$t_{\rm sh} \simeq \frac{\sigma}{H\rho\, d} \approx 300\,\mathrm{Myr}\,\frac{\sigma}{200\,\mathrm{km\,s^{-1}}}\left[\frac{\rho}{10\,M_\odot\,\mathrm{pc^{-3}}}\right]^{-1}\left[\frac{d}{\mathrm{pc}}\right]^{-1}, \tag{196}$$

where $H \approx 15$–20 is the dimensionless hardening rate, and ρ and σ describe the local stellar mass density and velocity dispersion in the vicinity of the binary. It is the efficiency of this loss-cone replenishment — through collisional relaxation, non-spherical perturbations, or gas-driven inflows — that largely determines whether a binary can survive the so-called *final parsec problem* and progress to separations at which GW emission becomes dominant.

At small separations, GW emission takes over as the primary driver of orbital decay. The corresponding timescale,

$$t_{\rm GW} = \frac{5\,d^4}{1024\,\eta\,M^3} \approx \frac{14\,{\rm Myr}}{\eta}\left[\frac{M}{10^9\,M_\odot}\right]^{-3}\left[\frac{d}{0.1\,{\rm pc}}\right]^4, \tag{197}$$

decreases extremely steeply with separation and becomes shorter than a Hubble time only at sub-parsec scales. For a binary with chirp mass $\mathscr{M}_c \sim 10^8\,M_\odot$, the GW-only coalescence time from a separation of $0.04\,{\rm pc}$ is of order the age of the universe. This underlines a fundamental challenge: without an efficient environmental channel to harden the binary well below the parsec scale, the vast majority of SMBH binaries would never merge within a Hubble time. At intermediate separations, we parametrise the combined environmental effects with

$$t_{\rm int} = t_{\rm GW}\left[\frac{d}{d_{\rm ref}}\right]^{-3\alpha/2}, \tag{198}$$

where $d_{\rm ref}$ is a reference separation related to the reference frequency by

$$d_{\rm ref}^3 = 8M/(\pi f_{\rm ref}(\mathscr{M}_c/10^9\,M_\odot)^{-\beta})^2\,. \tag{199}$$

The full picture of the timescales is shown in Fig. 15: the PTA band corresponds to the regime where environmental effects, most plausibly through a circumbinary gas disc, become active and can substantially shorten the effective orbital decay time with respect to the GW-only prediction. To see how this affects the GW emission, one proceeds as follows. Since the binary binding energy is $E = -(\pi f_r)^{2/3}\mathscr{M}_c^{5/3}/2$, where f_r is the rest-frame GW frequency, differentiation with respect to time gives the frequency evolution

$$\frac{{\rm d}\ln f_r}{{\rm d}t} = \frac{3}{2}\,t_{\rm GW}^{-1}\left(1+\frac{t_{\rm GW}}{t_{\rm env}}\right). \tag{200}$$

Combining this with the GW power ${\rm d}E_{\rm GW}/{\rm d}t$ yields the energy spectrum per logarithmic frequency interval,

$$\frac{{\rm d}E_{\rm GW}}{{\rm d}\ln f_r} = \frac{1}{3}\,\frac{(\pi f_r)^{2/3}\mathscr{M}_c^{5/3}}{1+t_{\rm GW}/t_{\rm env}}\,. \tag{201}$$

In the limit $t_{\rm GW} \ll t_{\rm env}$, one recovers the standard GW-driven spectrum. Environmental effects suppress the emission at frequencies below the GW-dominated regime, precisely the range probed by PTAs. Since PTAs observe a narrow frequency window, it is practical to expand $\ln(t_{\rm env}/t_{\rm GW})$ around a reference frequency $f_{\rm GW}$:

$$\frac{t_{\rm env}}{t_{\rm GW}} = \left(\frac{f_r}{f_{\rm GW}}\right)^{\alpha_1+\alpha_2\ln(f_r/f_{\rm GW})+\ldots}, \tag{202}$$

where the coefficients are defined by

$$\alpha_n \equiv \frac{1}{n!}\left(\frac{\rm d}{{\rm d}\ln f_r}\right)^n \ln\frac{t_{\rm env}}{t_{\rm GW}}\bigg|_{f_r=f_{\rm GW}}. \tag{203}$$

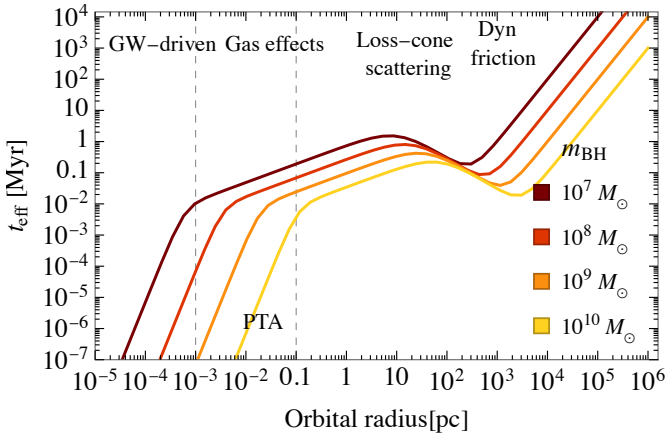


Figure 15: Effective orbital decay timescale $t_{\rm eff}$ of SMBH binaries of different masses as a function of the orbital separation, taking into account dynamical friction, loss-cone scattering, gas effects from a circumbinary disc and, finally, GW emission from a circular orbit. The separations at which the binaries emit GWs in the PTA band are indicated. Adapted from Paper VII [53], © 2024 American Physical Society.

Retaining only the leading-order term $\alpha_1 \equiv \alpha$ results in a power-law parametrisation of the environmental contribution, which has been widely adopted in the literature [70, 123, 82, 87]:

$$\begin{aligned} \frac{t_{\rm env}}{t_{\rm GW}} &= \left(\frac{f_r}{f_{\rm GW}}\right)^{\alpha}, \\ f_{\rm GW}(\mathscr{M}_c,\eta,z) &= f_{\rm ref}\left(\frac{\mathscr{M}_c}{10^9\,M_\odot}\right)^{-\beta}, \\ t_{\rm env} &= t_{\rm GW}(\mathscr{M}_{c,*},f_*)\left(\frac{\mathscr{M}_c}{\mathscr{M}_{c,*}}\right)^{\beta}\left(\frac{f_r}{f_*}\right)^{\alpha}. \end{aligned} \tag{204}$$

Here $f_{\rm ref}$ and $\alpha > 0$ are phenomenological parameters to be constrained by the NG15 data (noting that $t_{\rm env}$ grows with f_r whenever $\alpha > 8/3$, implying that the environmental timescale is longer at higher frequencies). The mass slope β reflects the underlying physical mechanism: studies in the literature consider $\beta = 0.2$–0.8 depending on whether the dominant effect is stellar hardening or gas-driven dissipation [123, 87, 70]; following Ref. [87], we adopt $\beta = 0.4$, a choice whose impact on the predicted background is expected to be modest given that the PTA signal originates from a relatively narrow range of BH masses. Through Eq. (201), the environmental parametrisation directly shapes the spectral form of the stochastic GW background in the PTA band. We also see in Fig.15, how the relevant masses that merge close to the PTA band are the ones that correspond to the dynamically measured local SMBHs that are inactive today. This is an important result of our analysis of the nanohertz stochastic GW background.

5.1.3 Imprints of the seeding process at low redshifts

As we have stated, one of the central open questions in SMBH research concerns their origin. The massive, locally observed SMBHs believed to drive the nanohertz stochastic background have grown through accretion by many orders of magnitude since their formation, effectively erasing any memory of their initial seeding conditions. The situation is

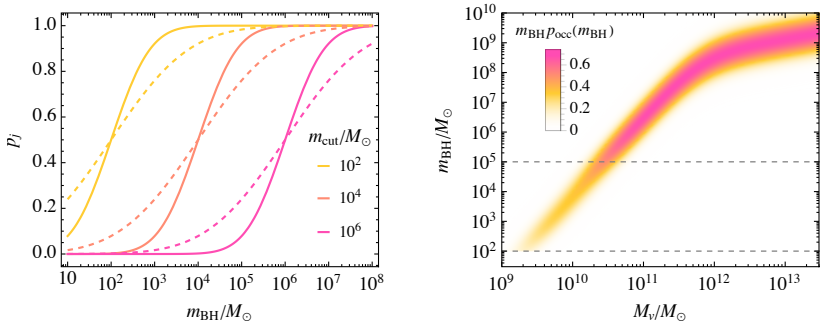


Figure 16: Left panel: *The occupation probability p_j as a function of BH mass for selected values of the low-mass cut $m_{\rm cut}$, with width parameters $w = 1$ (solid) and $w = 2$ (dashed). The error-function ansatz provides a smooth transition between zero and the asymptotic fraction f_j, capturing the crossover from halos too shallow to retain a seed to those hosting a fully formed BH.* Right panel: *The halo mass–BH mass plane, colour-coded by the occupation fraction $p_{\rm occ}$ for a two-component seeding model with $f_1 = f_2 = 0.5$, $w_1 = w_2 = 1$, $m_{\rm cut,1} = 100\,M_\odot$, and $m_{\rm cut,2} = 10^5\,M_\odot$. The figure illustrates how the coexistence of light and heavy seed populations gives rise to occupation in distinct regions of the M_v–$m_{\rm BH}$ plane. Reproduced from Paper VIII [57], © AAS 2024.*

markedly different at high redshift, where the seeding mechanism leaves a direct imprint on the predicted SMBH population. Among the local observables, perhaps the most informative probe of SMBH origins is the occupation fraction of the lightest galaxies, which host the lowest-mass BHs closest in mass to the original seeds.

The principal seeding channels that have been proposed in the literature can reproduce the properties of the observed SMBH population for appropriate choices of assembly parameters such as accretion rates, yet they differ substantially in their predictions for the spectrum of intermediate-mass BH mergers across cosmic time; see [143] for a comprehensive overview. These mechanisms generically predict either light seeds with masses $\sim 10^2$–$10^3\,M_\odot$, associated with Pop III remnants, or heavy seeds with masses $\sim 10^4$–$10^5\,M_\odot$, associated with direct collapse, at $z \lesssim 10$. At low redshifts, $z \lesssim 7$, the imprint of the seeding scenario manifests primarily in the low-mass tail of the BH occupation fraction. This is captured through the functions $p_j(m_{\rm BH}, \theta_j)$, for which we adopt the parametric form

$$p_j(m_{\rm BH}, \theta_j) = \frac{f_j}{2}\left(1 + \mathrm{erf}\left[\frac{\log_{10}(m_{\rm BH}/m_{{\rm cut},j})}{w_j}\right]\right), \tag{205}$$

where $\theta_j = (f_j, m_{{\rm cut},j}, w_j)$. Here f_j is the fraction of SMBHs produced by seeding mechanism j, $m_{{\rm cut},j}$ characterises the minimal SMBH mass associated with that channel, and w_j controls the sharpness of the transition. The condition $\sum_j f_j = 1$ is imposed. The function p_j is shown in the left panel of Fig. 16 for different values of $m_{\rm cut}$ and w. The occupation fraction entering the merger rate is then obtained by convolving p_j with a log-normal distribution over halo mass,

$$p_{\rm occ}(m_{\rm BH}|M_v, z, \theta) = \tag{206}$$

$$\sum_j \frac{p_j(m_{\rm BH}, \theta_j)}{\sqrt{2\pi}\, m_{\rm BH}\, \sigma} \exp\left[-\frac{\ln(m_{\rm BH}/\bar{m})^2}{2\sigma^2}\right], \tag{207}$$

which allows for scatter in the mass relation between DM halos and their central BHs, as illustrated in the right panel of Fig. 16. The occupation fraction enters the BH merger rate

directly. Following the formalism introduced earlier, the differential merger rate generalised to include seeding-dependent occupation is

$$\frac{\mathrm{d}R_{\mathrm{BH}}(p_{\mathrm{BH}},\theta)}{\mathrm{d}m_1\,\mathrm{d}m_2} \approx p_{\mathrm{BH}} \int \mathrm{d}M_1\,\mathrm{d}M_2\,\frac{\mathrm{d}R_h}{\mathrm{d}M_1\,\mathrm{d}M_2} \times \prod_{j=1,2} p_{\mathrm{occ}}(m_j|M_j,z,\theta)\,, \tag{208}$$

where p_{BH} is the probability that a halo merger leads to the coalescence of the central BHs. The expected number of binaries detectable by a given gravitational-wave observatory over an observational period is

$$\bar{N}(\theta) = \int_0^\infty \mathrm{d}\tau \int \mathrm{d}\lambda(\theta,x)\,p_{\mathrm{det}}[\mathrm{SNR}_c/\mathrm{SNR}(x,\tau)]\,, \tag{209}$$

where $\mathrm{SNR}(x)$ is the signal-to-noise ratio of the gravitational-wave signal from a binary with parameters x, p_{det} is the detection probability that accounts for source sky location and inclination [64, 67], and $\mathrm{SNR}_c = 8$ is the threshold for detection. The differential event rate per unit comoving volume is

$$\mathrm{d}\lambda(\theta,x) \equiv \frac{1}{1+z}\,\frac{\mathrm{d}V_c}{\mathrm{d}z}\,\frac{\mathrm{d}R_{\mathrm{BH}}(p_{\mathrm{BH}},\theta)}{\mathrm{d}m_1\,\mathrm{d}m_2}\,\mathrm{d}m_1\,\mathrm{d}m_2\,\mathrm{d}z\,. \tag{210}$$

This framework provides a direct link between the low-redshift SMBH population and the expected event rates at higher-frequency gravitational-wave detectors such as LISA, thereby connecting the seeding physics accessible at high redshift to future observational prospects at lower redshifts.

5.2 Supermassive Black Holes in the first billion years of the universe

In this section, we will elaborate on how to model the BH at high redshifts, which is much more complicated, as the situation is more dynamic. We have more mergers and bigger overall accretion and star formation. The question of how much or how is not yet settled. Given a seed BH of mass m_{seed}, the maximum rate at which it can grow through radiatively efficient accretion is set by the Eddington limit: the luminosity at which radiation pressure balances gravity, $L_{\mathrm{Edd}} = 4\pi G M_{\mathrm{BH}} m_p c/\sigma_T$, where m_p is the proton mass and σ_T is the Thomson cross-section. Assuming a radiative efficiency $\varepsilon_{\mathrm{rad}}$, the corresponding accretion rate is $\dot{M} = L_{\mathrm{Edd}}/(\varepsilon_{\mathrm{rad}}c^2)$, which leads to exponential growth on the Salpeter timescale of $\sim 45\,\mathrm{Myr}$ for $\varepsilon_{\mathrm{rad}} = 0.1$. Starting from a stellar-mass remnant of $m_{\mathrm{seed}} \sim 10^2\,M_\odot$ at redshift $z \sim 20$, reaching $\sim 10^9\,M_\odot$ by $z \sim 7$ requires uninterrupted Eddington-limited accretion, leaving no margin for periods of sub-Eddington growth. Quasars observed at $z \gtrsim 7$, however, prove that $\sim 10^9\,M_\odot$ SMBHs do exist at lookback times exceeding 13 Gyr, pressing hard against this limit [80, 45]. The problem is further sharpened by the recent JWST findings, which have revealed an abundant population of high-redshift AGN and candidate SMBHs at $z \gtrsim 4$–10 that appear systematically over-massive relative to their host stellar masses when compared to local scaling relations [73, 96]. Understanding the origins and early growth of these objects requires a framework that simultaneously tracks the assembly of DM halos, the formation of stars within them, and the feeding and feedback of the central BH, all against a background cosmology that may or may not be the standard cold-DM (ΛCDM) paradigm. It is precisely this kind of unified semi-analytic framework that motivates the two works on which this section draws.

The universe at $z \gtrsim 5$ is in the midst of reionisation: the first sources of UV photons are ionising the intergalactic medium, and galaxies are assembling rapidly in a bottom-up hierarchical sequence driven by the gravitational collapse of DM density fluctuations. At

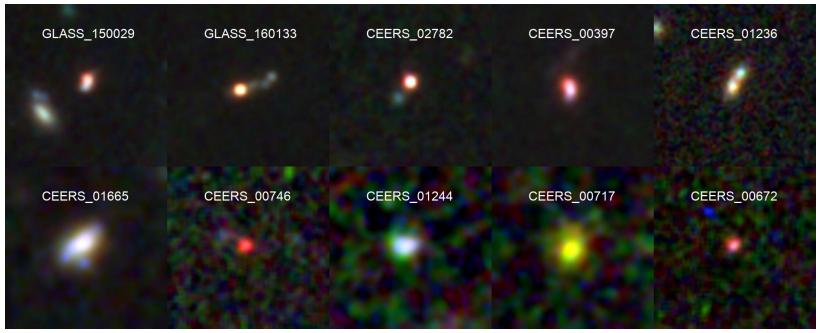


Figure 17: JWST/NIRCam images of broad-line AGNs at $z = 4$–7. Reproduced from Harikane et al. [73], © AAS 2023.

these redshifts, the HMF is dominated by objects far below the non-linear mass scale at $z = 0$, and the observational windows on this population remained very limited before the advent of JWST. The telescope's near-infrared sensitivity allows it to reach rest-frame UV wavelengths for sources at $z \sim 5$–25, making the UV luminosity function (UVLF) — the comoving number density of galaxies per unit UV absolute magnitude — the primary statistical observable for constraining both the galaxy population and its host DM haloes at these epochs. In addition, the detection of AGN at $z \gtrsim 4$ by JWST provides a complementary constraint on the BH mass function, enabling, for the first time, a joint analysis of stellar and BH growth histories in the high-redshift universe. The rest of this section builds up the theoretical ingredients needed to exploit these two observational handles.

We start by describing the minimal physics of a high-z universe and what the physics of the feedback and BH accretion are. We will then show how we can use the UV luminosity function as a tool to fix this feedback. Finally, we will show how the feedback fitted from the UV luminosity function, plus the method for accretion via mergers we have developed in the first section, can be used to evolve the SMBH seeds and compare to the JWST observations, while considering the DM model consistent along all of the processes.

5.2.1 Minimal high-redshift galaxy model

We have started by describing the DM halo abundance, their accretion and how they merge in different DM models, using analytical estimates from the EPS formalism. Galaxy formation and evolution are much more complicated to describe than the DM halo that hosts it. It involves the interplay between gas, star formation, and the energetic processes through which newly formed stars and accreting BHs regulate their own growth. These energetic processes are generically termed *feedback*, and their correct modelling is arguably the single most important challenge in galaxy formation theory. The rate at which these processes happen is assumed to be proportional to the overall accretion of gas by the halo that hosts the galaxy. There is a fundamental difference between stars and BHs; star formation requires cold gas, while BHs can also accrete hot gas. The gas is heated in a galaxy because of the feedback from the accreting SMBH, and it is expelled from the supernova (SN) feedback. Here, we outline the minimal set of feedback processes that must be incorporated into any realistic model of early galaxy formation. We first briefly review what are the physics behind both supernova and BH accretion feedback to motivate the parametric form for star formation, which we will fit against the UV luminosity function. The shape of the fit for star formation is expressed as $\dot{M}_* = f_*(M) f_{\rm cold}(M) f_B \dot{M}$ where $f_B = \Omega_B/\Omega_M \approx 0.16$ denotes the baryon fraction and $f_*(M)$ is the fraction of gas

that is converted into stars in a halo of mass M and $f_{\rm cold}(M)$ is the fraction of the gas that remains cold. We express this product as a broken power-law with an exponential suppression

$$f_* f_{\rm cold} = \varepsilon \frac{\alpha+\beta}{\beta(M/M_c)^{-\alpha}+\alpha(M/M_c)^{\beta}} e^{-M_t/M}, \tag{211}$$

in the next sections, we will see how we can learn a great deal by interpreting this parametric function. The first and most fundamental feedback channel at high redshift in light DM halos is driven by core-collapse supernovae (Type II SNe). When massive stars ($M_* \gtrsim 8\,M_\odot$) exhaust their nuclear fuel after a few million years, they explode and impart kinetic energy to the surrounding gas. The physical mechanism of SN feedback is as follows. When a stellar mass $M_*(z)$ is formed at redshift z, the total SN kinetic energy imparted to the interstellar medium is

$$E_{\rm SN} = f_w E_{51}\, \nu\, M_*(z) \equiv f_w\, v_s^2\, M_*(z)\,. \tag{212}$$

For this energy to eject all the remaining gas from the halo, it must exceed the binding energy of the ISM in the DM potential well. The escape velocity from the halo is $v_e = \sqrt{2}\,v_c$, where v_c is the circular velocity, so the energy required to unbind and expel the gas not converted into stars is

$$E_{\rm ej} = \frac{1}{2}\left[M_{g,i}(z) - M_*(z)\right] v_e^2\,, \tag{213}$$

where $M_{g,i}(z)$ is the initial gas mass available at that time step. The *ejection efficiency* $f_{\rm ej}$, the minimum star-formation efficiency required to blow away all remaining gas, is found by setting $E_{\rm SN} = E_{\rm ej}$, which yields [43]

$$f_{\rm ej} = \frac{v_c^2(z)}{v_c^2(z) + f_w\, v_s^2}\,. \tag{214}$$

This expression encodes a crucial physical insight: because $v_c \propto M_h^{1/3}(1+z)^{1/2}$ for virialized halos, the ejection efficiency is smaller for low-mass halos and decreases as a function of redshift at fixed mass. In other words, the shallower the potential well, the easier it is for SN energy to expel the gas. To follow this model exactly does not produce a good fit to the UV luminosity function but it provides a toy model for the relevant physics. We instead parametrise the feedback from the fitted star formation rate as a function of the halo mass

$$f_{\rm cold} \equiv \begin{cases} 1\,, & M < M_c \\ \dot{M}_*/(\varepsilon f_B \dot{M})\,, & M \geq M_c \end{cases}, \tag{215}$$

and the SFR efficiency

$$f_* \equiv \varepsilon \begin{cases} \dot{M}_*/(\varepsilon f_B \dot{M})\,, & M < M_c \\ 1\,, & M \geq M_c \end{cases}. \tag{216}$$

The suppression of f_* at $M < M_c$ arises because, as we have described, SN feedback depletes the cold gas, halting further star formation, the remaining gas after it has been expelled by the SN feedback can be expressed as

$$f_{\rm ej} \equiv (1-f_*)(1-f_{\rm SN}) \begin{cases} 1\,, & M < M_c \\ f_{\rm cold}^2\,, & M \geq M_c \end{cases}. \tag{217}$$

The quantity $f_{\rm SN}$ is the fraction of the gas left after star formation that is not expelled by the SN feedback. The motivation for introducing this parameter is to allow for some inefficiency in SN feedback and left gas to fuel the SMBH in light DM halos.

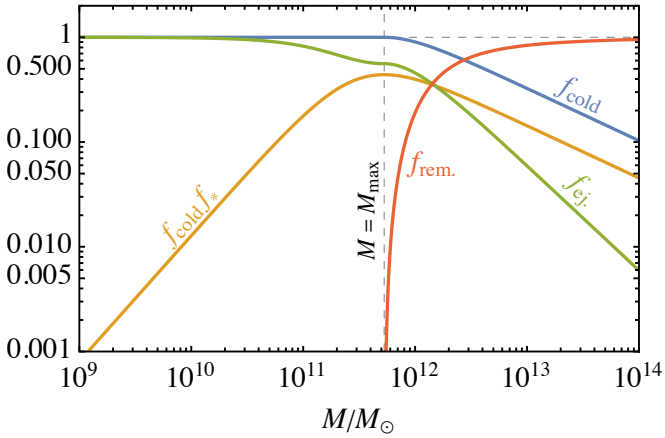


Figure 18: We show the relevant functions in our minimal high-z model of galaxy formation and evolution, for this case $f_{\rm SN} = 0$, so the fraction of available gas below M_c vanishes, which implies that all the gas below that halo mass is either used for star formation or expelled by SN feedback. In practice, we fit these parameters as a function of redshift to the UV luminosity function; this figure is only for illustrative purposes. Reproduced from Paper XI [58], © 2026 American Physical Society.

The second major feedback channel is associated with accretion onto the central SMBH, collectively referred to as active galactic nuclei (AGN) feedback. While SN feedback is primarily relevant for regulating star formation in low-mass haloes, where the shallow potential well makes gas retention difficult, AGN feedback operates predominantly in massive haloes where the SMBH has grown sufficiently large to couple its energy output to the galaxy-scale gas reservoir. We parametrise, in terms of the star formation rate, the combined fraction of hot and cold gas available for the SMBH to accrete, $f_{\rm rem}$,

$$f_{\rm rem} = 1 - (f_{\rm ej} + f_*) f_{\rm cold} \,. \tag{218}$$

We will return later in more detail to the SMBH accretion model. Notice that if $f_{\rm SN} = 0$, it implies that below $M < M_c$ no halo can have an accreting SMBH. Physically, this is motivated because winds from SN feedback eject a large portion of the available gas in light DM halos but also because efficient accretion requires the SMBH to reside stably at the centre of the gravitational potential, where the gas density is highest. In massive, well-relaxed systems this condition is naturally satisfied: the BH sinks rapidly to the centre through dynamical friction, and random gravitational perturbations are small compared to its binding energy. In low-mass halos, however, the situation is qualitatively different. When the BH mass is comparable to or smaller than the characteristic stellar mass within the influence radius, the restoring gravitational force is weak and stochastic kicks from the surrounding mass distribution impart a random walk to the BH trajectory [18]. The resulting Brownian motion displaces the seed from the density cusp and therefore suppresses accretion. This wandering regime constitutes an additional, purely gravitational barrier to early SMBH growth that is absent in the more massive haloes.

5.2.2 The UV Luminosity Function at high redshift

The UV luminosity function is going to be the main tool to learn about the co-evolution of SMBHs and stars. In its rest frame, a galaxy emits over a wide range of frequencies, and over cosmological distances this light is redshifted to lower frequencies. For a galaxy

at, say, $z \sim 10$ to be observed in the near infrared, it needs to emit in the UV range. The main sources of UV light are the heaviest stars in the universe, which are hotter and have short lifetimes. In conclusion, the UVLF is so powerful because it is the primary statistical observable for the galaxy population at high redshift and encodes, in a single function of two variables, the combined effects of DM structure formation, gas physics, stellar populations, and feedback.

The advent of JWST has pushed direct measurements of the UV LF to $z \sim 25$ [107, 38], complementing the more precise lower-redshift legacy data from HST [31, 72]. Together, these datasets span more than twelve billion years of cosmic history and provide the most direct observational window available on the evolution of star formation in the first billion years. In this subsection, we explain how the feedback model developed in the previous section translates the DM halo distribution into a prediction for $\Phi(M_{\rm UV}, z)$, and what the shape and high-redshift evolution of the observed luminosity function imply about the physics of early star formation and the nature of DM.

The theoretical UVLF is constructed by connecting the HMF to the star formation rate and thence to the UV luminosity of each galaxy. The observed UVLF receives two systematic corrections with respect to the intrinsic luminosity distribution: gravitational lensing by intervening large-scale structure, which boosts the apparent brightness of distant sources, and dust attenuation, which suppresses the observed UV flux. Including both effects, the observed luminosity function is written as

$$\Phi_{\rm UV}(L_{\rm obs}) = \int \mathrm{d}L \int \mathrm{d}\ln M \; \frac{\mathrm{d}P(L_{\rm obs}|L)}{\mathrm{d}L_{\rm obs}} \frac{\mathrm{d}P(L|M)}{\mathrm{d}L} \frac{\mathrm{d}n}{\mathrm{d}\ln M}\,, \tag{219}$$

where $\mathrm{d}P(L|M)/\mathrm{d}L$ is the distribution of intrinsic luminosities emitted by galaxies in halos of mass M, and $\mathrm{d}P(L_{\rm obs}|L)/\mathrm{d}L_{\rm obs}$ is the distribution of observed luminosities given an emitted luminosity L, which encapsulates lensing magnification and dust attenuation:

$$\frac{\mathrm{d}P(L_{\rm obs}|L)}{\mathrm{d}L_{\rm obs}} \approx \int \mathrm{d}\mu\, \mathrm{d}B \; \frac{\mathrm{d}P(\mu)}{\mathrm{d}\mu} \frac{\mathrm{d}P(B)}{\mathrm{d}B} \; \delta\!\left(L_{\rm obs} - \frac{\mu}{B} L\right), \tag{220}$$

where μ is the lensing magnification and B is the dust attenuation factor. Following standard practice, the variance in dust attenuation is neglected and $\mathrm{d}P(B)/\mathrm{d}B \approx \delta(B - \bar{B})$ is adopted, so that the UV LF simplifies to

$$\Phi_{\rm UV}(L_{\rm obs}) = \int \mathrm{d}\mu\, \mathrm{d}\ln M \; \frac{\bar{B}}{\mu} \frac{\mathrm{d}P(\mu)}{\mathrm{d}\mu} \frac{\mathrm{d}P(L|M)}{\mathrm{d}L} \frac{\mathrm{d}n}{\mathrm{d}\ln M}\,. \tag{221}$$

It is convenient to express the UV LF in terms of the absolute AB magnitude, defined through [105]

$$\log_{10}\!\left(\frac{L_{\rm UV}}{\mathrm{erg\,s^{-1}}}\right) = 0.4\,(51.63 - M_{\rm UV})\,, \tag{222}$$

from which $\Phi_{\rm UV}(M_{\rm obs})$ follows by the appropriate Jacobian, with the effective magnitude shift $M_{\rm UV} = M_{\rm obs} - A_{\rm UV} + 1.086 \ln \mu$ encoding both dust attenuation and lensing [133]. Gravitational lensing by intervening DM halos magnifies the apparent flux of distant sources, boosting the bright end of the observed UV LF with respect to the intrinsic distribution. The magnification μ is related to the line-of-sight convergence κ by [130]

$$\mu \approx (1 - \kappa)^{-2}\,. \tag{223}$$

The total convergence is built up as a sum over all intervening lenses j,

$$\kappa = \kappa_E + \sum_j \kappa^{(1)}(\theta_j)\,, \qquad \kappa^{(1)}(\theta_j) = \frac{8\pi D_j D_{j,s}}{D_s}\, \Sigma_j(r_j)\,, \tag{224}$$

where $\Sigma_j(r_j)$ is the projected surface mass density of the j-th lens at angular diameter distance D_j and transverse separation r_j from the line of sight, D_s is the angular diameter distance of the source, $D_{j,s}$ the distance between lens and source, and $\kappa_E = -\langle \sum_j \kappa^{(1)} \rangle$ is the empty-beam convergence introduced to ensure $\langle \kappa \rangle = 0$. The distribution of lens halos along a line of sight to a source at redshift z_s is

$$\frac{\mathrm{d}N(\theta)}{\mathrm{d}\ln M} = \int_0^{z_s} \mathrm{d}z_l \int_0^{r_{\max}} \mathrm{d}r \, \frac{2\pi(1+z_l)^2 r}{H(z_l)} \frac{\mathrm{d}n(z_l)}{\mathrm{d}\ln M}\,, \tag{225}$$

where the upper limit $r_{\max}(M, z_l)$ is set by the condition $\kappa^{(1)} > \kappa_{\rm thr}$ for a sufficiently small threshold $\kappa_{\rm thr}$. In practice, lensing magnification at these redshifts is dominated by halos of mass $M \gtrsim 10^{10} M_\odot$, and is therefore largely insensitive to the modifications of small-scale structure that distinguish different DM models. Dust attenuation suppresses the observed UV flux, most significantly at low redshift and bright magnitudes where the dust-to-gas ratio in galaxies is highest. The mean attenuation magnitude $A_{\rm UV} = 2.5 \log_{10} \bar{B}$ is estimated through the empirical relation

$$A_{\rm UV} = \frac{1}{s} \ln\left[e^{s\left(C_0 + 0.2\ln 10\,(C_1\sigma_\beta)^2 + C_1\bar{\beta}_{\rm UV}\right)} + 1 \right], \tag{226}$$

where $s = 3$, $C_0 = 4.4$, $C_1 = 2.0$ and $\sigma_\beta = 0.34$ are fixed parameters, and the mean UV slope is

$$\bar{\beta}_{\rm UV} = -\exp\left(\frac{0.17(19.5 + M_{\rm UV})}{1.54 + 0.075z} \right) (1.54 + 0.075z)\,, \tag{227}$$

obtained by fitting the observational results of [30] with an exponential scaling in $M_{\rm UV}$. The softplus form of Eq. (226) replaces a piecewise function, with the sharpness parameter $s = 3$ chosen to reproduce numerical sampling results from the literature [141]. At $z \gtrsim 15$, where the dust-to-gas ratio of early galaxies is expected to be negligible, dust attenuation is switched off.

The UV luminosity function probes the star-formation rate (SFR) because the rest-frame ultraviolet continuum is dominated by massive, short-lived stars with lifetimes $\tau \lesssim 10\,\mathrm{Myr}$, which act as near-instantaneous tracers of ongoing star formation. This connection is quantified through the UV conversion factor $\kappa_{\rm UV}$, defined by

$$\bar{L}_{\rm UV} = \frac{\dot{M}_*}{\kappa_{\rm UV}}, \tag{228}$$

so that a higher SFR $\dot{M}_*$ produces a proportionally higher UV luminosity. The value of $\kappa_{\rm UV}$ is set by the shape of the stellar initial mass function (IMF): more top-heavy IMFs, in which massive stars contribute a larger share of the total stellar mass, yield a lower $\kappa_{\rm UV}$ and thus a higher UV luminosity per unit SFR [93, 65]. For the standard Salpeter IMF, appropriate for the Pop I and Pop II stellar populations observed in the local and intermediate-redshift Universe, one obtains $\kappa_{\rm UV} = \kappa_0 = 1.15 \times 10^{-22} M_\odot\,\mathrm{s\,erg^{-1}\,Myr^{-1}}$ [93]. Population III (Pop III) stars, forming from pristine, metal-free gas in the first generation of halos at $z \gtrsim 10$, are instead expected to be characterised by a distinctly top-heavy IMF, with typical masses of order $\mathcal{O}(10\text{–}10^3)\,M_\odot$ [32, 83]. This shift in the mass spectrum results in $\kappa_{\rm UV} < \kappa_0$, meaning that a Pop III-dominated stellar population produces more UV photons per unit of stellar mass formed than its metal-enriched counterpart, and can therefore account for an observed excess in the UV luminosity function without necessarily requiring a higher SFR or a more abundant halo population. This will play a very important role when fitting

the JWST observations at the highest redshift. The emitted magnitudes $M_{\rm UV}$ in a given halo are assumed to follow a Gaussian

$$\frac{\mathrm{d}P(M_{\rm UV}|M)}{\mathrm{d}M_{\rm UV}} = \frac{1}{\sqrt{2\pi}\,\sigma_{\rm UV}} \exp\!\left(-\frac{(M_{\rm UV}-\bar{M}_{\rm UV})^2}{2\sigma_{\rm UV}^2}\right). \tag{229}$$

The SFR is assumed to be proportional to the halo mass growth rate computed from the EPS formalism,

$$\dot{M}_* = f_B\, f_*(M)\, \dot{M}_h\,, \tag{230}$$

where $f_B = \Omega_B/\Omega_M \approx 0.16$ is the cosmic baryon fraction. The proportionality coefficient $f_*(M)$ encodes the reduction of the star-formation efficiency by SN and AGN feedback as a function of halo mass. The shape and properties of Eq. (230) have been introduced and motivated in the previous section.

5.2.3 Black hole co-evolution

Having established the UV luminosity function as a probe of the star-formation rate and the HMF, we can now turn to the co-evolution of SMBHs and their host galaxies. The key insight behind the approach is that the same EPS formalism developed in Section 3.13 to track the baryonic assembly of halos can be applied directly to the BH mass, without the computational overhead of explicit merger-tree realisations. Because an SMBH residing in a halo of mass M at redshift z has, on average, followed the same hierarchical assembly history as any mass element of that halo, the EPS conditional probabilities describe the stochastic history of the BH just as faithfully as they describe the DM distribution. The result is a set of coupled ordinary differential equations for the mean BH mass $M_{\rm BH}(M,z)$ and the mean stellar mass $M_*(M,z)$ in a halo of mass M at redshift z, which can be solved numerically at a computational cost orders of magnitude lower than hydrodynamic simulations or Monte Carlo merger trees. To track the growth of the SMBHs, we need to introduce the growth via mergers, which is subdominant for stars but very relevant for the BHs. The total rate of change of each component is written as the sum of a merger contribution and a growth contribution from baryonic processes

$$\dot{M}_{\rm BH}(M) = \dot{M}_{\rm BH}^{\rm merg}(M) + \dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M)\,, \tag{231}$$

$$\dot{M}_*(M) = \dot{M}_*^{\rm merg}(M) + \dot{M}_*^{\rm sf}(M)\,, \tag{232}$$

where $\dot{M}_J^{\rm merg}$ denotes the growth rate driven by halo mergers, $\dot{M}_{\rm BH}^{\rm acc}$ is the BH growth by gas accretion, and $\dot{M}_*^{\rm sf}$ is the stellar mass growth by star formation. The two components $M_{\rm BH}$ and M_* are coupled through the gas budget: the same baryonic gas reservoir that feeds star formation also supplies the material for BH accretion, and both channels are regulated by the SN and AGN feedback processes discussed in Section 5.2.1. The merger contribution to the BH and stellar mass growth rates follows directly from the EPS framework developed in Section 3.13. The mean mass of component $J \in \{\mathrm{BH}, *\}$ in a halo of mass M at redshift z that has been inherited from progenitor halos at redshift $z' > z$ is given by Eq. (141), and the instantaneous merger growth rate is obtained by taking the limit $\Delta z \equiv z' - z \to 0$:

$$\Delta M_J^{\rm merg}(M,z,z') = \int_0^M \mathrm{d}M' \,\frac{\mathrm{d}N}{\mathrm{d}M'} M_J(M',z') - M_J(M,z')\,. \tag{233}$$

The merger growth rate $\dot{M}_J^{\rm merg}(M,z)$ is obtained from Eq. (233) via the relation in Eq. (143). Physically, the integral sums the contribution of all progenitor halos of mass $M' < M$

whose BH or stellar content $M_J(M', z')$ is incorporated into the halo of mass M as structure builds up hierarchically.

A key difference between stars and SMBH growth via mergers is that not every halo merger results in the prompt coalescence of the two central SMBHs. Following the dynamical processes we have previously described, the two BHs must sink to the centre of the merged potential through dynamical friction, form a bound binary at parsec separations, and harden via stellar loss-cone scattering before GW emission can drive the final coalescence. Each of these stages may stall or proceed at different rates depending on the environment. This complexity is captured by a single efficiency parameter $0 \leq p_{\rm BH} \leq 1$, which multiplies the BH merger growth rate:

$$\dot{M}_{\rm BH}^{\rm merg}(M) \;\rightarrow\; p_{\rm BH}\,\dot{M}_{\rm BH}^{\rm merg}(M)\,. \tag{234}$$

The stellar mass merger rate, $\dot{M}_*^{\rm merg}$, proceeds without such a penalty since stars participate in the galactic merger regardless of what the central BHs do.

The accretion channel couples BH growth directly to the gas available in the host halo after star formation and internal feedback have acted on the baryonic reservoir. The BH accretion rate is the minimum of the available gas supply and a fraction of the Eddington rate:

$$\dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M) = \min\left[f_{\rm Edd}\,\dot{M}_{\rm Edd}(M_{\rm BH}),\; f_{\rm rem}\,f_B\left(f_1^{\rm acc}\,\dot{M} + f_2^{\rm acc}\,M\right)\right], \tag{235}$$

where the Eddington rate is

$$\dot{M}_{\rm Edd} = \frac{L_{\rm Edd}}{\varepsilon_{\rm rad}\,c^2} = \frac{4\pi G M_{\rm BH} m_p}{\varepsilon_{\rm rad}\,c\,\sigma_T}\,, \tag{236}$$

with $\varepsilon_{\rm rad}$ the radiative efficiency, and $f_{\rm Edd}$ is a dimensionless parameter that allows for accretion up to or beyond the classical Eddington limit. The second argument of the minimum contains two physically distinct contributions. The term $f_1^{\rm acc}\,\dot{M}$ accounts for the fraction of freshly accreted baryons that reach the central region after star formation and feedback, with $f_1^{\rm acc}$ a dimensionless efficiency; the term $f_2^{\rm acc}\,M$ describes accretion from the ambient gas already present in the halo, with $f_2^{\rm acc}$ carrying units of inverse time. Both contributions are modulated by $f_{\rm rem}$, which encodes the feedback history.

This parametrisation incorporates the AGN feedback loop self-consistently. Once the BH has grown to the point where $f_{\rm Edd}\,\dot{M}_{\rm Edd}$ dominates the accretion rate, the luminosity $L_{\rm Edd}$ heats and removes the surrounding gas, reducing $f_{\rm cold}$ and hence $f_{\rm rem}$. Above M_c, this heating is the primary driver of the cold-gas suppression in Eq. (214), so that the accretion and star-formation histories are tightly coupled through the shared gas reservoir. The competition between merger-driven and accretion-driven growth then shapes the $M_{\rm BH}$–M_* relation across halo masses: in halos below M_c, where $f_{\rm rem} \approx f_{\rm SN} \ll 1$, BH growth is dominated by the merger channel, while in massive halos, the Eddington-limited accretion term takes over.

The coupled equations from Eq.(235) must be supplemented with an initial condition: the mass of the first BH seeds, $m_{\rm seed}$, planted in every halo above a minimum mass $M_{\rm seed}$ at a formation redshift $z_{\rm seed}$. This initial condition encodes the unknown physics of the very first generation of massive BHs, which cannot be derived from the EPS formalism alone. As we have discussed, two broad families of seeding scenarios are considered in the literature: *light seeds* with $m_{\rm seed} \sim 10^2$–$10^3\,M_\odot$ planted in relatively low-mass halos ($M_{\rm seed} \sim 10^5\,M_\odot$), corresponding to Pop III stellar remnants; and *heavy seeds* with $m_{\rm seed} \sim 10^4$–$10^5\,M_\odot$ planted in more massive halos ($M_{\rm seed} \sim 10^6$–$10^8\,M_\odot$), corresponding to direct-collapse BHs or the products of dynamical processes in dense stellar systems.

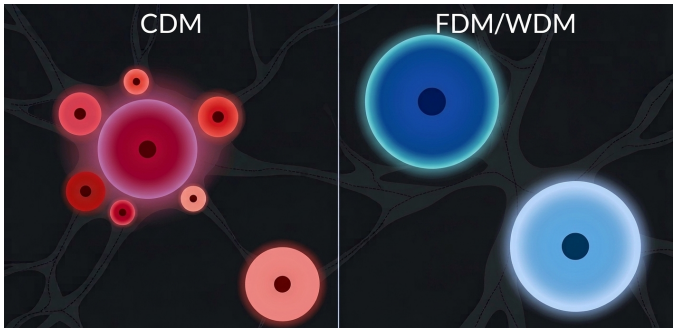


Figure 19: In this sketch, we illustrate the difference in growth through mergers in CDM versus DM models for which there is suppressed small-scale structure like in FDM or WDM.

The DM model enters the BH co-evolution equations through the same channel as in the UVLF: the variance $\sigma^2(M)$ and its derivative with respect to mass, which control both the halo growth rate and the merger rate in the EPS formalism. In CDM, $\sigma^2(M)$ is a monotonically decreasing function of mass and the merger rate is a smooth function at all relevant mass scales. In FDM and WDM, the suppression of $\sigma^2(M)$ below the Jeans or free-streaming scale reduces both the abundance of low-mass halos and the rate at which they merge into more massive systems. This has a twofold effect on the predicted $M_{\rm BH}$–M_* relation. First, the reduced number of merger events at the low end of the halo mass function means that BHs in low-mass galaxies have fewer merger partners, suppressing the merger-driven growth channel that is dominant for $M_* \lesssim 10^{10}\,M_\odot$. Second, the reduced halo growth rate at small masses translates into a smaller gas supply and hence a lower star-formation rate, which reduces the stellar mass at fixed halo mass relative to CDM.

The magnitude of this shift depends on the DM particle mass. For FDM with $m_{\rm FDM} \sim 10^{-22}$ eV, the suppression of structure begins at halo masses $M_h \sim 10^9$–$10^{10}\,M_\odot$, precisely the range that dominates the merger history of the low-M_* AGN population detected by JWST. For WDM with $m_{\rm WDM} \sim$ keV, the corresponding suppression mass is similar, as the free-streaming length and the FDM Jeans scale are comparable in the observationally relevant parameter space. The comparison between the CDM predictions, which reproduce the observed $M_{\rm BH}$–M_* relation at all redshifts, and the FDM or WDM predictions, which produce a systematic deficit of BH mass at high redshift, is thus a direct probe of the nature of DM on sub-galactic scales. This sensitivity makes the stellar mass–BH mass relation at high redshift a particularly powerful cosmological probe. Unlike the UVLF, which depends on the *number density* of halos and is therefore sensitive to the DM model primarily through the HMF, the BH mass at fixed stellar mass depends on the *merger history* of halos, which is encoded in the conditional EPS probabilities. The merger history is a more direct probe of the shape and amplitude of $\sigma^2(M)$ at small scales, because it depends on how many and how massive the progenitors were at each step in the hierarchical assembly. In WDM and FDM, the suppression of progenitors below the Jeans or free-streaming scale truncates this merger history in a way that is visible in the $M_{\rm BH}$–M_* relation even when the UV LF shows only a mild deviation from CDM. As we will show in the papers, this constitutes one of the main results of this thesis.

6 Summary and conclusions

The introductory part of this thesis has reviewed the generation and detection of GWs, the analytic description of structure formation in different DM models, the wave-optics lensing of GWs and the astrophysics of SMBHs in the local and in the high-redshift universe. This section summarises the original results of the eleven publications on which the thesis is based, organised according to the four research objectives formulated in Section 1.1, and states the main conclusions of the thesis.

6.1 Wave-optics lensing of gravitational waves by dark matter (O1)

Paper I [61] presented the first systematic treatment of GW lensing by extended DM structures in the wave-optics regime. The amplification factor was computed numerically for uniform spheres, NFW halos and FDM halos with solitonic cores, none of which admits a closed analytical expression, and the lensing signatures were classified according to the number and nature of the extremal paths of the time-delay surface, as reviewed in Section 4. The detectability of these signatures was quantified for LIGO, ET, LISA, AEDGE and BBO through the waveform mismatch and the lensing optical depth. A central, and initially surprising, result was that more sensitive detectors do not substantially improve the prospects of detecting individual DM halos.

Paper II [138] extended the microlensing analysis of Ref. [136] to PBHs dressed with halos of particle DM. It provided the first fully consistent treatment of such dressed lenses, and its methodology was subsequently adopted by the LIGO–Virgo–KAGRA collaboration in its search for microlensing signatures [4]. Constraints on the PBH abundance were derived from the absence of lensing signatures in the LIGO–Virgo catalogue, and projections were made for ET. An important implication is that ET will observe the stellar-mass BH population twice, through the mergers themselves and through the microlensing of GWs by the same population, which opens the possibility of testing the DM profile around these objects and of discriminating between a stellar and a primordial origin even for PBH fractions well below the total DM abundance.

Paper III [137] explained why individual DM halos remain undetectable with LISA regardless of its sensitivity: the accessible lens mass is bounded by the Fresnel volume of the GW, a fundamental wave-optics limitation rather than a technical one. The same paper showed that the cumulative effect of the many light halos contained in the Fresnel volume, the "dark timbre" of the GW, is a guaranteed signal in CDM and constitutes a new class of DM observable.

6.2 Structure formation beyond cold dark matter (O2)

The analytic framework of Section 3 was developed across Papers IV, V, X and XI. Paper IV [77] computed the variance of the matter perturbations and the HMF in models with a white-noise enhancement of the small-scale power, as arises from axion miniclusters or heavy PBHs. Paper X [134] introduced the key technical advance of following the evolution of the halo population and of its baryonic and BH content with differential equations derived from the EPS conditional probabilities, rather than with Monte Carlo merger trees; this makes the computation orders of magnitude cheaper and allows the HMF to be followed down to arbitrarily small masses. Paper V [133] extended the formalism to ellipsoidal collapse, providing a conditional first-crossing distribution that reproduces Monte Carlo random walks even for large redshift separations, and introduced a smooth window function that yields a consistent HMF and halo growth rate in CDM, WDM, FDM and enhanced-power models. Paper XI [58] combined these ingredients into a single model

in which the dependence on the DM model enters only through the variance $\sigma^2(M)$ and its derivative.

6.3 Supermassive black hole binaries and pulsar timing arrays (O3)

Paper VI [54] built the SMBH binary merger-rate model of Section 5.1, anchored to the local SMBH–stellar mass relations and to the EPS halo merger rate, fitted its amplitude to the NANOGrav 12.5-year data and extended it to lower masses to predict the event rates for LISA and atom interferometers. It also established that the distribution of $\Omega_{\rm GW}$ is more efficiently computed from the merger rate and then convolved than by simulating individual binaries, and that this distribution is strongly non-Gaussian, with a tail towards large $\Omega_{\rm GW}$ and a mode well below the mean, so that individual realisations of the background can be markedly spiky. Paper VII [53] confronted an improved version of the model with the NANOGrav 15-year data, which for the first time provided information on the spectral shape. The spectrum requires environmental hardening of the binaries in the PTA band, and the population responsible for the signal was identified with the inactive galaxies hosting the heaviest, dynamically measured SMBHs. Paper VIII [57] extended the model self-consistently to lower masses and showed that the minimum mass of the SMBH occupation fraction, a low-redshift tracer of the seeding mechanism, can be recovered with LISA and with atom interferometers. Paper IX [55] connected the merger rates and environmental effects inferred from the PTA data to the JWST observations of little red dots and dual AGN, interpreted as merging SMBHs.

6.4 Supermassive black holes and galaxies in the early universe (O4)

Paper IV [77] showed that the first, photometric JWST galaxy candidates could not be explained in CDM even with maximal star-formation efficiency, and explored an enhancement of small-scale structure as a remedy. Paper V [133] revisited this question with the much larger and spectroscopically calibrated UVLF data set that is now available up to $z \sim 25$, and with the full model of Section 5.2, including feedback, dust attenuation and lensing magnification. Enhanced small-scale power does not improve the fit; instead, the data require a rapid transition to Pop III stars and reduced SN and AGN feedback at the highest redshifts, and they constrain FDM and WDM. Paper IX [55] showed that the overmassive SMBHs observed by JWST are a natural continuation of the high-mass end of the local relation of dynamically measured SMBHs, and that it is the local AGN sample that is the outlier. Paper X [134] evolved light and heavy seeds with the differential EPS model and found that light seeds can reproduce the JWST SMBHs only if the merger efficiency is high, whereas heavy seeds provide a more natural fit. Finally, Paper XI [58] fitted the feedback parameters to the UVLF and the remaining astrophysical parameters to the SMBH observations, and found that the SMBH population constrains deviations from CDM significantly more strongly than the UVLF alone, the constraints becoming stronger for lighter seeds.

6.5 Main conclusions

The main conclusions of this thesis are the following.

1. Wave-optics lensing of GWs by individual DM halos is limited by the Fresnel volume of the wave and not by the detector sensitivity. Individual NFW or FDM halos will not be detected with LISA or with ground-based detectors, whereas the cumulative lensing by the many light halos within the Fresnel volume is a guaranteed signal in CDM (Papers I and III).

2. The absence of microlensing signatures in the LIGO–Virgo catalogue constrains the abundance of PBHs, including PBHs dressed with particle DM halos. ET will observe the stellar-mass BH population both through mergers and through microlensing, which will allow the DM profile around these objects to be probed and a primordial origin to be distinguished from a stellar one (Paper II).

3. The EPS formalism, extended to ellipsoidal collapse and equipped with a smooth window function, provides an accurate and computationally efficient description of the HMF, the halo merger rate and the accretion of DM, baryons and BHs in CDM, WDM, FDM and enhanced-power models, without the need for merger trees (Papers V, X and XI).

4. The nanohertz GW background measured by PTAs is consistent with a population of SMBH binaries whose merger rate is inherited from the EPS halo merger rate and whose masses follow the relation of the inactive, dynamically measured local SMBHs. Its spectral shape requires environmental hardening of the binaries in the PTA band, and the background is strongly non-Gaussian (Papers VI and VII).

5. The same population model, extrapolated to lower masses, predicts the event rates for LISA and atom interferometers and shows that the low-mass cut-off of the SMBH occupation fraction, an imprint of the seeding mechanism, is recoverable with these detectors (Paper VIII).

6. The JWST UV luminosity function up to $z \sim 25$ does not require an enhancement of small-scale structure. It requires instead a rapid transition to Pop III stars and reduced feedback at the highest redshifts, and it constrains FDM and WDM (Papers IV and V).

7. The over-massive SMBHs observed by JWST at high redshift continue the local relation of dynamically measured SMBHs, the local AGN sample being the outlier. Light seeds can reproduce them only with a high merger efficiency, whereas heavy seeds provide a more natural fit (Papers IX and X).

8. Because SMBHs retain the imprint of the mergers that took place in light halos at high redshift, the SMBH population constrains deviations from CDM more strongly than the galaxy population does, and the lighter the seeds, the stronger the constraint (Paper XI).

6.6 Outlook

The tools developed in this thesis are directly applicable to the data that will arrive in the coming years. On the GW side, the growing PTA data sets and, eventually, LISA and the atom interferometers will test the SMBH population model of Section 5 at the level of individual binaries, and the dark timbre of the GWs detected by LISA will offer a first look at the population of light DM halos. On the electromagnetic side, spectroscopic JWST samples of SMBHs at $z \gtrsim 5$, and in particular a clarification of the nature of the little red dots, will sharpen the constraints on the seeding mechanism and, through it, on the DM model. The EPS-based framework of this thesis, in which the DM model enters only through the variance of the matter perturbations, is well suited to confront all of these observations consistently.

List of Figures

1 First black hole sketch from 1784 12
2 LIGO Hanford 13
3 Bullet Cluster 14
4 LIGO binary examples 25
5 NANOGrav 2023 posteriors and power-law fit 31
6 GW detector landscape 32
7 EPS random path illustration 41
8 First-crossing distributions 45
9 Variance and accretion in different DM models 46
10 Halo mass function in different DM models 47
11 Point mass effect on LIGO binaries 54
12 Amplification function for a constant sphere 55
13 Amplification function for a NFW halo 57
14 SMBH local observations 61
15 Supermassive BH binary timescales 65
16 Occupation fraction and low-mass cut of the SMBH population 66
17 Broad-line AGNs from JWST 68
18 Feedback in the minimal high-z model 70
19 Sketch of the difference in growth of the BH 75

References

[1] B. P. Abbott et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116(6):061102, 2016.

[2] B. P. Abbott et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116(6):061102, 2016.

[3] B. P. Abbott et al. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev. X*, 9(3):031040, 2019.

[4] R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. X. Adhikari, V. B. Adya, C. Affeldt, D. Agarwal, M. Agathos, K. Agatsuma, N. Aggarwal, O. D. Aguiar, L. Aiello, A. Ain, P. Ajith, K. M. Aleman, G. Allen, A. Allocca, P. A. Altin, A. Amato, S. Anand, A. Ananyeva, S. B. Anderson, W. G. Anderson, S. V. Angelova, S. Ansoldi, J. M. Antelis, S. Antier, S. Appert, K. Arai, M. C. Araya, J. S. Areeda, M. Arène, N. Arnaud, S. M. Aronson, K. G. Arun, Y. Asali, G. Ashton, S. M. Aston, P. Astone, F. Aubin, P. Aufmuth, K. AultONeal, C. Austin, S. Babak, F. Badaracco, M. K. M. Bader, S. Bae, A. M. Baer, S. Bagnasco, Y. Bai, J. Baird, M. Ball, G. Ballardin, S. W. Ballmer, M. Bals, A. Balsamo, G. Baltus, S. Banagiri, D. Bankar, R. S. Bankar, J. C. Barayoga, C. Barbieri, B. C. Barish, D. Barker, P. Barneo, F. Barone, B. Barr, L. Barsotti, M. Barsuglia, D. Barta, J. Bartlett, M. A. Barton, I. Bartos, R. Bassiri, A. Basti, M. Bawaj, J. C. Bayley, A. C. Baylor, M. Bazzan, B. Bécsy, V. M. Bedakihale, M. Bejger, I. Belahcene, V. Benedetto, D. Beniwal, M. G. Benjamin, T. F. Bennett, J. D. Bentley, M. BenYaala, F. Bergamin, B. K. Berger, S. Bernuzzi, C. P. L. Berry, D. Bersanetti, A. Bertolini, J. Betzwieser, R. Bhandare, A. V. Bhandari, D. Bhattacharjee, S. Bhaumik, J. Bidler, I. A. Bilenko, G. Billingsley, I. A. Birney, O. Birnholtz, S. Biscans, M. Bischi, S. Biscoveanu, A. Bisht, B. Biswas, M. Bitossi, M.-A. Bizouard, J. K. Blackburn, J. Blackman, C. D. Blair, D. G. Blair, R. M. Blair, F. Bobba, N. Bode, M. Boer, G. Bogaert, M. Boldrini, F. Bondu, E. Bonilla, R. Bonnand, P. Booker, B. A. Boom, R. Bork, V. Boschi, N. Bose, S. Bose, V. Bossilkov, V. Boudart, Y. Bouffanais, A. Bozzi, C. Bradaschia, P. R. Brady, A. Bramley, A. Branch, M. Branchesi, J. E. Brau, M. Breschi, T. Briant, J. H. Briggs, A. Brillet, M. Brinkmann, P. Brockill, A. F. Brooks, J. Brooks, D. D. Brown, S. Brunett, G. Bruno, R. Bruntz, J. Bryant, A. Buikema, T. Bulik, H. J. Bulten, A. Buonanno, R. Buscicchio, D. Buskulic, R. L. Byer, L. Cadonati, M. Caesar, G. Cagnoli, C. Cahillane, H. W. C. , III, J. C. Bustillo, J. D. Callaghan, T. A. Callister, E. Calloni, J. B. Camp, M. Canepa, M. Cannavacciuolo, K. C. Cannon, H. Cao, J. Cao, E. Capote, G. Carapella, F. Carbognani, J. B. Carlin, M. F. Carney, M. Carpinelli, G. Carullo, T. L. Carver, J. Casanueva Diaz, C. Casentini, G. Castaldi, S. Caudill, M. Cavaglià, F. Cavalier, R. Cavalieri, G. Cella, P. Cerdá-Durán, E. Cesarini, W. Chaibi, and K. Chakravarti. Search for Lensing Signatures in the Gravitational-Wave Observations from the First Half of LIGO–Virgo's Third Observing Run. *Astrophys. J.*, 923(1):14, Dec. 2021.

[5] R. Abbott et al. GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *Phys. Rev. X*, 11:021053, 2021.

[6] R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run. 11 2021.

[7] R. Abbott et al. The population of merging compact binaries inferred using gravitational waves through GWTC-3. 11 2021.

[8] A. Abdalla et al. Terrestrial Very-Long-Baseline Atom Interferometry: summary of the second workshop. *EPJ Quant. Technol.*, 12(1):42, 2025.

[9] G. Agazie et al. Comparing recent PTA results on the nanohertz stochastic gravitational wave background. 9 2023.

[10] G. Agazie et al. The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background. *Astrophys. J. Lett.*, 951(1):L8, 2023.

[11] G. Agazie et al. The NANOGrav 15 yr Data Set: Observations and Timing of 68 Millisecond Pulsars. *Astrophys. J. Lett.*, 951(1):L9, 2023.

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

[13] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020. [Erratum: Astron.Astrophys. 652, C4 (2021)].

[14] P. Ajith et al. A Template bank for gravitational waveforms from coalescing binary black holes. I. Non-spinning binaries. *Phys. Rev. D*, 77:104017, 2008. [Erratum: Phys.Rev.D 79, 129901 (2009)].

[15] P. Amaro-Seoane et al. Laser Interferometer Space Antenna. 2017.

[16] J. Antoniadis et al. The second data release from the European Pulsar Timing Array - III. Search for gravitational wave signals. *Astron. Astrophys.*, 678:A50, 2023.

[17] J. Antoniadis et al. The second data release from the European Pulsar Timing Array I. The dataset and timing analysis. 6 2023.

[18] J. N. Bahcall and R. A. Wolf. Star distribution around a massive black hole in a globular cluster. *Astrophys. J.*, 209:214–232, Oct. 1976.

[19] C. Baraldo, A. Hosoya, and T. T. Nakamura. Gravitationally induced interference of gravitational waves by a rotating massive object. *Phys. Rev. D*, 59:083001, 1999.

[20] M. C. Begelman. Super-Eddington Fluxes from Thin Accretion Disks? *Astrophys. J. Lett.*, 568:L97–L100, 2002.

[21] K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy. The first gravitational-wave source from the isolated evolution of two 40-100 Msun stars. *Nature*, 534:512, 2016.

[22] M. Benito, G. Hütsi, K. Müürsepp, J. S. Almeida, J. Urrutia, V. Vaskonen, and H. Veermäe. Fuzzy dark matter fails to explain dark matter cores. *Phys. Dark Univ.*, 49:102010, 2025.

[23] A. J. Benson, M. Kamionkowski, and S. H. Hassani. Self-consistent theory of halo mergers. *Mon. Not. Roy. Astron. Soc.*, 357:847–858, 2005.

[24] M. C. Bentz, K. D. Denney, C. J. Grier, A. J. Barth, B. M. Peterson, M. Vestergaard, V. N. Bennert, G. Canalizo, G. De Rosa, A. V. Filippenko, E. L. Gates, J. E. Greene, W. Li, M. A. Malkan, R. W. Pogge, D. Stern, T. Treu, and J.-H. Woo. The Low-luminosity End of the Radius-Luminosity Relationship for Active Galactic Nuclei. *Astrophys. J.*, 767(2):149, Apr. 2013.

[25] G. Bertone, D. Hooper, and J. Silk. Particle dark matter: Evidence, candidates and constraints. *Phys. Rept.*, 405:279–390, 2005.

[26] J. Binney and S. Tremaine. *Galactic Dynamics: Second Edition*. 2008.

[27] P. Bode, J. P. Ostriker, and N. Turok. Halo formation in warm dark matter models. *Astrophys. J.*, 556:93–107, 2001.

[28] A. Bogdan et al. Evidence for Heavy-Seed Origin of Early Supermassive Black Holes from a $z \approx 10$ X-Ray Quasar. *Nature Astron.*, 8:126–133, 2024.

[29] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion set mass functions for hierarchical Gaussian fluctuations. *Astrophys. J.*, 379:440, 1991.

[30] R. J. Bouwens, G. D. Illingworth, P. A. Oesch, I. Labbé, P. G. van Dokkum, M. Trenti, M. Franx, R. Smit, V. Gonzalez, and D. Magee. UV-Continuum Slopes of $>$4000 $z \sim$ 4-8 Galaxies from the HUDF/XDF, HUDF09, ERS, CANDELS-South, and CANDELS-North Fields. *Astrophys. J.*, 793(2):115, 2014.

[31] R. J. Bouwens, P. A. Oesch, M. Stefanon, G. Illingworth, I. Labbé, N. Reddy, H. Atek, M. Montes, R. Naidu, T. Nanayakkara, E. Nelson, and S. Wilkins. New Determinations of the UV Luminosity Functions from z 9 to 2 Show a Remarkable Consistency with Halo Growth and a Constant Star Formation Efficiency. *Astron. J.*, 162(2):47, Aug. 2021.

[32] V. Bromm and N. Yoshida. The First Galaxies. *Ann. Rev. Astron. Astrophys.*, 49:373–407, 2011.

[33] Caltech/MIT/LIGO Lab. LIGO Hanford, May 2008. Photograph, ID: ligo20150731e. Accessed: 2026-03-24.

[34] N. Cappelluti, G. Hasinger, and P. Natarajan. Exploring the High-redshift PBH-ΛCDM Universe: Early Black Hole Seeding, the First Stars and Cosmic Radiation Backgrounds. *Astrophys. J.*, 926(2):205, 2022.

[35] B. Carr, F. Kuhnel, and M. Sandstad. Primordial Black Holes as Dark Matter. *Phys. Rev. D*, 94(8):083504, 2016.

[36] B. Carr and J. Silk. Primordial Black Holes as Generators of Cosmic Structures. *Mon. Not. Roy. Astron. Soc.*, 478(3):3756–3775, 2018.

[37] B. J. Carr and J. Silk. Can graininess in the early universe make galaxies? *Astrophys. J.*, 268:1–16, May 1983.

[38] M. Castellano et al. Pushing JWST to the extremes: search and scrutiny of bright galaxy candidates at z$\simeq$15-30. 4 2025.

[39] S. Chatrchyan et al. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012.

[40] D. Clowe, M. Bradac, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky. A direct empirical proof of the existence of dark matter. *Astrophys. J. Lett.*, 648:L109–L113, 2006.

[41] C. A. Correa, J. S. B. Wyithe, J. Schaye, and A. R. Duffy. The accretion history of dark matter haloes – I. The physical origin of the universal function. *Mon. Not. Roy. Astron. Soc.*, 450(2):1514–1520, 2015.

[42] C. Cutler, E. Poisson, G. J. Sussman, and L. S. Finn. Gravitational radiation from a particle in circular orbit around a black hole. 2: Numerical results for the nonrotating case. *Phys. Rev. D*, 47:1511–1518, 1993.

[43] P. Dayal, A. Ferrara, J. Dunlop, and F. Pacucci. Essential physics of early galaxy formation. *Mon. Not. Roy. Astron. Soc.*, 445(3):2545–2557, 2014.

[44] V. De Luca, V. Desjacques, G. Franciolini, and A. Riotto. The clustering evolution of primordial black holes. *JCAP*, 11:028, 2020.

[45] X. Ding, M. Onoue, J. D. Silverman, Y. Matsuoka, T. Izumi, et al. Detection of stellar light from quasar host galaxies at redshifts above 6. *Nature*, 621(7977):51–55, Sept. 2023.

[46] V. Domcke. Discovery Opportunities with Gravitational Waves – TASI 2024 Lecture Notes. 9 2024.

[47] A. Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)*, 1916:688–696, 1916.

[48] A. Einstein. Über Gravitationswellen. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)*, 1918:154–167, 1918.

[49] D. J. Eisenstein and W. Hu. Baryonic features in the matter transfer function. *Astrophys. J.*, 496:605, 1998.

[50] D. Ellis, D. J. E. Marsh, B. Eggemeier, J. Niemeyer, J. Redondo, and K. Dolag. Structure of axion miniclusters. *Phys. Rev. D*, 106(10):103514, 2022.

[51] J. Ellis. Gravitational Waves: Echoes of the Biggest Bangs Since the Big Bang and/or BSM Physics? *Universe*, 11(7):213, 2025.

[52] J. Ellis, M. Fairbairn, G. Franciolini, G. Hütsi, A. Iovino, M. Lewicki, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. What is the source of the PTA GW signal? *Phys. Rev. D*, 109(2):023522, 2024.

[53] J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data. *Phys. Rev. D*, 109(2):L021302, 2024.

[54] J. Ellis, M. Fairbairn, G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Prospects for future binary black hole gravitational wave studies in light of PTA measurements. *Astron. Astrophys.*, 676:A38, 2023.

[55] J. Ellis, M. Fairbairn, G. Hütsi, J. Urrutia, V. Vaskonen, and H. Veermäe. Consistency of JWST black hole observations with NANOGrav gravitational wave measurements. *Astron. Astrophys.*, 691:A270, 2024.

[56] J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. In search of the biggest bangs since the big bang. *Int. J. Mod. Phys. D*, 33(15):2440004, 2024.

[57] J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Probing Supermassive Black Hole Seed Scenarios with Gravitational-wave Measurements. *Astrophys. J.*, 964(1):11, 2024.

[58] J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Constraints on Dark Matter Models from Supermassive Black Hole Evolution. *Phys. Rev. Lett.*, 136(13):131001, 2026.

[59] A. L. Erickcek, M. Kamionkowski, and A. J. Benson. Supermassive Black Hole Merger Rates: Uncertainties from Halo Merger Theory. *Mon. Not. Roy. Astron. Soc.*, 371:1992–2000, 2006.

[60] M. Fairbairn, D. J. E. Marsh, J. Quevillon, and S. Rozier. Structure formation and microlensing with axion miniclusters. *Phys. Rev. D*, 97(8):083502, 2018.

[61] M. Fairbairn, J. Urrutia, and V. Vaskonen. Microlensing of gravitational waves by dark matter structures. *JCAP*, 07:007, 2023.

[62] M. Feix, J. Frank, A. Pargner, R. Reischke, B. M. Schäfer, and T. Schwetz. Isocurvature bounds on axion-like particle dark matter in the post-inflationary scenario. *JCAP*, 05:021, 2019.

[63] L. S. Finn. Detection, measurement and gravitational radiation. *Phys. Rev. D*, 46:5236–5249, 1992.

[64] L. S. Finn and D. F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev. D*, 47:2198–2219, 1993.

[65] K. Freese, R. Price, and D. N. Schramm. Formation of population III stars and galaxies with primordial planetary-mass black holes. *Astrophys. J.*, 275:405–412, Dec. 1983.

[66] J. P. Gardner et al. The James Webb Space Telescope Mission. *Publ. Astron. Soc. Pac.*, 135:068001, 2023.

[67] D. Gerosa, S. Ma, K. W. K. Wong, E. Berti, R. O'Shaughnessy, Y. Chen, and K. Belczynski. Multiband gravitational-wave event rates and stellar physics. *Phys. Rev. D*, 99(10):103004, 2019.

[68] I. D. Gialamas, G. Hütsi, M. Raidal, J. Urrutia, M. Vasar, and H. Veermäe. Quintessence and phantoms in light of DESI 2025. *Phys. Rev. D*, 112(6):063551, 2025.

[69] G. Girelli, L. Pozzetti, M. Bolzonella, C. Giocoli, F. Marulli, and M. Baldi. The stellar-to-halo mass relation over the past 12 Gyr: I. Standard ΛCDM model. *Astron. Astrophys.*, 634:A135, 2020.

[70] Z. Haiman, Z. Haiman, B. Kocsis, B. Kocsis, K. Menou, and K. Menou. The Population of Viscosity- and Gravitational Wave-Driven Supermassive Black Hole Binaries Among Luminous AGN. *Astrophys. J.*, 700:1952–1969, 2009. [Erratum: Astrophys.J. 937, 129 (2022)].

[71] S. H. Hansen, J. Lesgourgues, S. Pastor, and J. Silk. Constraining the window on sterile neutrinos as warm dark matter. *Mon. Not. Roy. Astron. Soc.*, 333:544–546, 2002.

[72] Y. Harikane, M. Ouchi, M. Oguri, Y. Ono, K. Nakajima, Y. Isobe, H. Umeda, K. Mawatari, and Y. Zhang. A Comprehensive Study of Galaxies at z $\sim$ 9–16 Found in the Early JWST Data: Ultraviolet Luminosity Functions and Cosmic Star Formation History at the Pre-reionization Epoch. *Astrophys. J. Suppl.*, 265(1), 2023.

[73] Y. Harikane, Y. Zhang, K. Nakajima, M. Ouchi, Y. Isobe, Y. Ono, S. Hatano, Y. Xu, and H. Umeda. A JWST/NIRSpec First Census of Broad-line AGNs at z = 4-7: Detection of 10 Faint AGNs with M $_{BH}$ 10^6-10^8 M $_\odot$ and Their Host Galaxy Properties. *Astrophys. J.*, 959(1):39, Dec. 2023.

[74] R. W. Hellings and G. S. Downs. UPPER LIMITS ON THE ISOTROPIC GRAVITATIONAL RADIATION BACKGROUND FROM PULSAR TIMING ANALYSIS. *Astrophys. J. Lett.*, 265:L39–L42, 1983.

[75] W. Hu, R. Barkana, and A. Gruzinov. Cold and fuzzy dark matter. *Phys. Rev. Lett.*, 85:1158–1161, 2000.

[76] L. Hui, J. P. Ostriker, S. Tremaine, and E. Witten. Ultralight scalars as cosmological dark matter. *Phys. Rev. D*, 95(4):043541, 2017.

[77] G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Did JWST observe imprints of axion miniclusters or primordial black holes? *Phys. Rev. D*, 107(4):043502, 2023.

[78] G. Hütsi, M. Raidal, V. Vaskonen, and H. Veermäe. Two populations of LIGO-Virgo black holes. *JCAP*, 03:068, 2021.

[79] D. Inman and Y. Ali-Haïmoud. Early structure formation in primordial black hole cosmologies. *Phys. Rev. D*, 100(8):083528, 2019.

[80] T. Izumi, Y. Matsuoka, S. Fujimoto, M. Onoue, M. A. Strauss, et al. Subaru High-z Exploration of Low-luminosity Quasars (SHELLQs). XIII. Large-scale Feedback and Star Formation in a Low-luminosity Quasar at z = 7.07 on the Local Black Hole to Host Mass Relation. *Astrophys. J.*, 914(1):36, June 2021.

[81] P. Jaranowski and A. Krolak. Gravitational-Wave Data Analysis. Formalism and Sample Applications: The Gaussian Case. *Living Rev. Rel.*, 8:3, 2005.

[82] L. Z. Kelley, L. Blecha, and L. Hernquist. Massive Black Hole Binary Mergers in Dynamical Galactic Environments. *Mon. Not. Roy. Astron. Soc.*, 464(3):3131–3157, 2017.

[83] R. S. Klessen and S. C. O. Glover. The First Stars: Formation, Properties, and Impact. *Ann. Rev. Astron. Astrophys.*, 61:65–130, 2023.

[84] A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada. Where Are the Missing Galactic Satellites? *Astrophys. J.*, 522:82–92, 1999.

[85] E. W. Kolb and I. I. Tkachev. Axion miniclusters and Bose stars. *Phys. Rev. Lett.*, 71:3051–3054, 1993.

[86] J. Kormendy and L. C. Ho. Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies. *Ann. Rev. Astron. Astrophys.*, 51:511–653, 2013.

[87] M. M. Kozhikkal, S. Chen, G. Theureau, M. Habouzit, and A. Sesana. Mass-redshift dependency of Supermassive Black Hole Binaries for the Gravitational Wave Background. 5 2023.

[88] C. G. Lacey and S. Cole. Merger rates in hierarchical models of galaxy formation. *Mon. Not. Roy. Astron. Soc.*, 262:627–649, 1993.

[89] M. Leo, C. M. Baugh, B. Li, and S. Pascoli. A new smooth-k space filter approach to calculate halo abundances. *JCAP*, 04:010, 2018.

[90] V. M. Lipunov and I. E. Panchenko. Pulsars revived by gravitational waves. *Astron. Astrophys.*, 312:937–940, Aug. 1996.

[91] M. R. Lovell, C. S. Frenk, V. R. Eke, A. Jenkins, L. Gao, and T. Theuns. The properties of warm dark matter haloes. *Mon. Not. Roy. Astron. Soc.*, 439:300–317, 2014.

[92] D. Lynden-Bell. Galactic Nuclei as Collapsed Old Quasars. *Nature*, 223:690–694, 1969.

[93] P. Madau and M. Dickinson. Cosmic Star Formation History. *Ann. Rev. Astron. Astrophys.*, 52:415–486, 2014.

[94] P. Madau and M. J. Rees. Massive black holes as Population III remnants. *Astrophys. J. Lett.*, 551:L27–L30, 2001.

[95] R. Maiolino et al. A small and vigorous black hole in the early Universe. *Nature*, 627:59–63, 2024.

[96] R. Maiolino et al. JADES - The diverse population of infant black holes at $4 < z < 11$: Merging, tiny, poor, but mighty. *Astron. Astrophys.*, 691:A145, 2024.

[97] D. J. E. Marsh. Axion Cosmology. *Phys. Rept.*, 643:1–79, 2016.

[98] J. Matthee et al. Little Red Dots: An Abundant Population of Faint Active Galactic Nuclei at $z \sim 5$ Revealed by the JWST Advanced Deep Extragalactic Survey. *Astrophys. J.*, 963:129, 2024.

[99] J. Michell. On the Means of Discovering the Distance, Magnitude, &c. of the Fixed Stars, in Consequence of the Diminution of the Speed of Their Light. *Phil. Trans. R. Soc. Lond.*, 74:35–57, 1784.

[100] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. 1973.

[101] B. P. Moster, R. S. Somerville, C. Maulbetsch, F. C. v. d. Bosch, A. V. Maccio', T. Naab, and L. Oser. Constraints on the relationship between stellar mass and halo mass at low and high redshift. *Astrophys. J.*, 710:903–923, 2010.

[102] D. Mukherjee, Y. Zhou, N. Chen, U. N. Di Carlo, and T. Di Matteo. MAGICS. III. Seeds Sink Swiftly: Nuclear Star Clusters Dramatically Accelerate Seed Black Hole Mergers. *Astrophys. J.*, 981(2):203, 2025.

[103] NASA, ESA, CSA, STScI, and CXC. Bullet Cluster (Webb and Chandra Image), June 2025. Object: Bullet Cluster (1E 0657-56). Instrument: NIRCam. Exposure date: 20 January 2025. Science: J. Jee (Yonsei University/UC Davis), S. Cha (Yonsei University), K. Finner (Caltech/IPAC). Image processing: J. DePasquale (STScI). Page last updated: 12 November 2025. Accessed: 2026-03-24.

[104] J. F. Navarro, C. S. Frenk, and S. D. M. White. A Universal Density Profile from Hierarchical Clustering. *Astrophys. J.*, 490(2):493–508, Dec. 1997.

[105] J. B. Oke and J. E. Gunn. Secondary standard stars for absolute spectrophotometry. *Astrophys. J.*, 266:713, 1983.

[106] S. Passaglia and W. Hu. Accurate effective fluid approximation for ultralight axions. *Phys. Rev. D*, 105(12):123529, 2022.

[107] P. G. Pérez-González et al. The Rise of the Galactic Empire: Ultraviolet Luminosity Functions at z $\sim$ 17 and z $\sim$ 25 Estimated with the MIDIS+NGDEEP Ultra-deep JWST/NIRCam Data Set. *Astrophys. J.*, 991(2):179, 2025.

[108] P. C. Peters. Gravitational Radiation and the Motion of Two Point Masses. *Phys. Rev.*, 136:B1224–B1232, 1964.

[109] P. C. Peters and J. Mathews. Gravitational radiation from point masses in a Keplerian orbit. *Phys. Rev.*, 131:435–439, 1963.

[110] E. S. Phinney. A Practical theorem on gravitational wave backgrounds. 7 2001.

[111] E. Poisson and C. M. Will. Gravitational waves from inspiraling compact binaries: Parameter estimation using second postNewtonian wave forms. *Phys. Rev. D*, 52:848–855, 1995.

[112] W. H. Press and P. Schechter. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *Astrophys. J.*, 187:425–438, 1974.

[113] G. D. Quinlan. The dynamical evolution of massive black hole binaries - I. hardening in a fixed stellar background. *New Astron.*, 1:35–56, 1996.

[114] J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Eccentricity effects on the supermassive black hole gravitational wave background. *Astron. Astrophys.*, 691:A212, 2024.

[115] J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Statistics of supermassive black hole gravitational wave background anisotropy. *Astron. Astrophys.*, 706:A159, 2026.

[116] D. J. Reardon et al. Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array. *Astrophys. J. Lett.*, 951(1):L6, 2023.

[117] A. E. Reines and M. Volonteri. Relations between Central Black Hole Mass and Total Galaxy Stellar Mass in the Local Universe. *Astrophys. J.*, 813:82, 2015.

[118] A. E. Reines and M. Volonteri. Relations between Central Black Hole Mass and Total Galaxy Stellar Mass in the Local Universe. *Astrophys. J.*, 813(2):82, Nov. 2015.

[119] V. C. Rubin, W. K. Ford, Jr., and N. Thonnard. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc). *Astrophys. J.*, 238:471–487, June 1980.

[120] H.-Y. Schive, T. Chiueh, T. Broadhurst, and K.-W. Huang. Contrasting Galaxy Formation from Quantum Wave Dark Matter, ψDM, with ΛCDM, using Planck and Hubble Data. *Astrophys. J.*, 818(1):89, 2016.

[121] B. F. Schutz. Determining the Hubble constant from gravitational wave observations. *Nature*, 323(6086):310–311, Sept. 1986.

[122] K. Schwarzschild. On the gravitational field of a mass point according to Einstein's theory. *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)*, 1916:189–196, 1916.

[123] A. Sesana. Insights into the astrophysics of supermassive black hole binaries from pulsar timing observations. *Class. Quant. Grav.*, 30:224014, 2013.

[124] A. Sesana, A. Vecchio, and C. N. Colacino. The stochastic gravitational-wave background from massive black hole binary systems: implications for observations with Pulsar Timing Arrays. *Mon. Not. Roy. Astron. Soc.*, 390:192, 2008.

[125] R. K. Sheth and G. Tormen. An Excursion Set Model of Hierarchical Clustering : Ellipsoidal Collapse and the Moving Barrier. *Mon. Not. Roy. Astron. Soc.*, 329:61, 2002.

[126] V. Springel et al. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435:629–636, 2005.

[127] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, June 2005.

[128] R. Takahashi. *Wave Effects in the Gravitational Lensing of Gravitational Waves from Chirping Binaries*. PhD thesis, Department of Physics, Kyoto University, 2004.

[129] R. Takahashi and T. Nakamura. Wave effects in gravitational lensing of gravitational waves from chirping binaries. *Astrophys. J.*, 595:1039–1051, 2003.

[130] R. Takahashi, M. Oguri, M. Sato, and T. Hamana. Probability Distribution Functions of Cosmological Lensing: Convergence, Shear, and Magnification. *Astrophys. J.*, 742:15, 2011.

[131] A. Toubiana, L. Sberna, M. Volonteri, E. Barausse, S. Babak, R. Enficiaud, D. I. Villalba, J. R. Gair, J. E. Greene, and H. Quelquejay Leclere. Reconciling PTA and JWST, and preparing for LISA with POMPOCO: a Parametrisation Of the Massive black hole POpulation for Comparison to Observations. *Astron. Astrophys.*, 700:A135, 2025.

[132] J. Urrutia. Using gravitational waves to detect axion miniclusters and axion stars. *PoS*, COSMICWISPers:046, 2024.

[133] J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. Starlight from JWST: Implications for star formation and dark matter models. *Astron. Astrophys.*, 702:A109, 2025.

[134] J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. The origin of the JWST supermassive black holes. *Astron. Astrophys.*, 703:A138, 2025.

[135] J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. The origin of the JWST supermassive black holes. *Astron. Astrophys.*, 703:A138, 2025.

[136] J. Urrutia and V. Vaskonen. Lensing of gravitational waves as a probe of compact dark matter. *Mon. Not. Roy. Astron. Soc.*, 509(1):1358–1365, 2021.

[137] J. Urrutia and V. Vaskonen. Dark timbre of gravitational waves. *Phys. Rev. D*, 111(12):123047, 2025.

[138] J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational wave microlensing by dressed primordial black holes. *Phys. Rev. D*, 108(2):023507, 2023.

[139] J. Verwohlt, C. A. Mason, J. B. Muñoz, F.-Y. Cyr-Racine, M. Vogelsberger, and J. Zavala. Separating dark acoustic oscillations from astrophysics at cosmic dawn. *Phys. Rev. D*, 110(10):103533, 2024.

[140] M. Viel, J. Lesgourgues, M. G. Haehnelt, S. Matarrese, and A. Riotto. Constraining warm dark matter candidates including sterile neutrinos and light gravitinos with WMAP and the Lyman-alpha forest. *Phys. Rev. D*, 71:063534, 2005.

[141] M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. F. Snyder, S. Bird, D. Nelson, and L. Hernquist. Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509:177–182, 2014.

[142] M. Volonteri. Formation of Supermassive Black Holes. *Astron. Astrophys. Rev.*, 18:279–315, 2010.

[143] M. Volonteri, M. Habouzit, and M. Colpi. The origins of massive black holes. *Nature Rev. Phys.*, 3(11):732–743, 2021.

[144] H. Xu et al. Searching for the Nano-Hertz Stochastic Gravitational Wave Background with the Chinese Pulsar Timing Array Data Release I. *Res. Astron. Astrophys.*, 23(7):075024, 2023.

[145] Y. B. Zel'dovich. Gravitational instability: An approximate theory for large density perturbations. *Astron. Astrophys.*, 5:84–89, Mar. 1970.

[146] J. Zhang, C.-P. Ma, and O. Fakhouri. Conditional Mass Functions and Merger Rates of Dark Matter Halos in the Ellipsoidal Collapse Model. *Mon. Not. Roy. Astron. Soc.*, 387:L13–L17, 2008.

[147] Y. Zhou, D. Mukherjee, N. Chen, T. Di Matteo, P. H. Johansson, A. Rantala, C. Partmann, U. N. Di Carlo, S. Bird, and Y. Ni. MAGICS. II. Seed Black Holes Stripped of Their Surrounding Stars Do Not Sink. *Astrophys. J.*, 980(1):79, Feb. 2025.

Acknowledgements

I would like to thank my family, and especially my grandparents, for their continuous interest and their support throughout this journey. I have been fortunate to carry out this thesis surrounded by remarkable people, and I am grateful to everyone in the KBFI High Energy Group for pushing me to work hard and discover what I am capable of. In particular, I would like to thank Malcolm, Hardi, John, and Ville, colleagues who have become friends, and whose ability and love for physics have been a constant source of inspiration and example. Finally, I owe my deepest gratitude to my wife Yaquelin, who has kept me both humane and happy through these years of hard work. Whatever success this work may have, it belongs to us both.

This work was supported by European Regional Development Fund through the CoE program grant TK133 and by the Estonian Research Council grant PRG803.

Abstract
Black Holes and Dark Matter in light of new infrared and gravitational wave observations

The aim of this thesis is to use gravitational waves and black holes as precision probes of the nature of dark matter. The thesis is organised around four interconnected research programmes, each addressing a distinct direction in the observational and theoretical study of supermassive black holes, dark matter and gravitational waves, and corresponding to the four research objectives formulated in the introduction: wave-optics lensing of gravitational waves by dark matter structures; an analytic description of structure formation valid beyond cold dark matter; the supermassive black hole binary population and the nanohertz gravitational wave background; and the formation of supermassive black holes and galaxies in the first billion years as a test of dark matter.

The first program concerns the gravitational lensing of gravitational waves as a probe of dark matter. In Paper I [61], we studied the detectability of NFW halos and fuzzy dark matter cores via GW microlensing, presenting the first systematic treatment of NFW profiles, which lack a closed analytical expression for the amplification factor F, and the first inclusion of fuzzy dark matter halos with their characteristic cores. The techniques developed in that work have since become standard in the community. A notable and initially surprising finding was that more sensitive future detectors did not substantially improve detectability prospects; as we later understood, this reflected a fundamental physical limitation rather than a technical one. In Paper II [138], we extended the results from [136] to primordial black holes, deriving constraints from LIGO and the Einstein Telescope. We incorporated waveform mismatch, the effect of particle dark matter halos around PBHs, and the possibility of detecting such halos directly. Our treatment of particle dark matter halos around PBHs was the first fully consistent one in the literature, and the methodology was subsequently cited by the LIGO collaboration [4] in their search for microlensing signatures. An important implication of this work is that the Einstein Telescope will observe the binary black hole population twice, once through mergers and once through microlensing of gravitational waves from the same population, enabling independent tests of the dark matter profile around these objects and, in principle, a discrimination between stellar and primordial origin even for primordial black hole fractions well below the dark matter abundance. Finally, in Paper III [137] we returned to the question of why dark matter halos are undetectable with LISA, tracing the bottleneck to a fundamental wave-optics effect: the Fresnel volume limits the accessible dark matter regardless of detector sensitivity. We then showed that the cumulative lensing signal from many lighter dark matter halos within this Fresnel volume is, in contrast, guaranteed in the CDM framework, and constitutes a new class of detectable signal.

The second direction of research concerns the high-redshift galaxy and black hole observations from JWST and their implications for fundamental physics. In Paper IV [77], we computed the variance of matter perturbations in the presence of axion minicluster enhancement, showing that even maximal star formation efficiency, all baryons converted to stars, was insufficient to account for the observed galaxy masses, and that an enhancement of small-scale structure was required. This paper captures the attitude of the early post-JWST community, in which non-standard physics appeared necessary to explain the observations. Over the following years, spectroscopic confirmation of galaxy samples and improved photometric calibration revised some of those early results. In Paper V [133], we carried out a considerably more sophisticated analysis incorporating realistic star formation, dust attenuation, and weak lensing, using the substantially enlarged dataset of UV

luminosity function measurements now available up to $z \sim 25$. The result was perhaps counterintuitive: the inclusion of primordial black hole or axion minicluster structure enhancements did not improve the fit, and significant deviations from CDM, including fuzzy dark matter and warm dark matter scenarios, were constrained. What the data required instead was a rapid transition to Pop III stars and reduced supernova and AGN feedback at very high redshifts, with which an excellent fit to the full dataset was obtained.

The third program addresses gravitational wave predictions from supermassive black hole binaries in the local universe. In Paper VI [54], we fitted the amplitude of the stochastic gravitational wave background from SMBH mergers to the NANOGrav 12.5-year dataset, and extended the model to lower masses to produce predictions for LISA and for atom interferometers. Although the observational constraints at the time were limited to a measured amplitude, this work established an important methodological advance: it is significantly more efficient to compute the distribution of $\Omega_{\rm GW}$ from the merger rate and simulate the total signal by convolving this distribution, rather than generating individual binaries directly. This approach also revealed that the true distribution of the background is highly non-Gaussian, with a tail towards large $\Omega_{\rm GW}$ and a mode displaced well below the mean, implying that individual realisations of the background can be markedly spiky. In Paper VII [53], we applied an improved version of this framework to the NANOGrav 15-year dataset, which for the first time provided spectral shape information in addition to amplitude. This new observable revealed that environmental effects are necessary to reproduce the observed spectral shape, and by identifying the population responsible for the signal with inactive galaxies, we obtained an excellent fit to the data while simultaneously constraining the environmental hardening mechanisms. In Paper VIII [57], we extended the PTA model self-consistently to lower masses and demonstrated that the minimum mass of the SMBH occupation fraction, a low-redshift tracer of the seeding mechanism, is recoverable with both LISA and atom interferometers.

The fourth and final program concerns the formation and early evolution of SMBHs from their seeds in the first billion years of cosmic history. In Paper IX [55], we investigated the SMBH–galaxy scaling relations observed by JWST, arguing that the apparent tension with local AGN scaling relations is resolved once one recognises that the high-redshift observations are a natural continuation of the high-mass end of local SMBH scaling relations, and that it is the local AGN sample that is the outlier. We also connected the merger rates and environmental effects inferred from PTAs to JWST observations of little red dots and dual AGNs, interpreted as merging SMBHs. In Paper X [134], we developed a semi-analytical model to evolve SMBH seeds, introducing the key improvement of tracking halo mergers via differential equations rather than merger trees, substantially increasing computational efficiency and allowing the halo mass function to be followed down to arbitrarily small masses, which is critical for constraining the light seed scenario. A likelihood scan over seeding parameters showed that light seeds can reproduce the observations provided merger efficiency is high, while heavy seeds provide a more natural fit. The final paper of this thesis, Paper XI [58], synthesises and extends many of these threads. It combines an improved elliptical-collapse EPS formalism with feedback parameters fitted directly to the UV luminosity function, rather than fixed, and fits all remaining astrophysical parameters directly to the SMBH observations. This constitutes the most parameter-complete model to date fitted directly to data. Its central finding is that SMBH observations place significantly more stringent constraints on deviations from CDM than those derived from the UV luminosity function alone. The physical reason is clear: whereas the UV luminosity function probes structure formation at the epoch of observation, SMBHs, which live practically forever, carry imprints of mergers reaching back to

their formation in light dark matter halos at high redshift, precisely the regime most sensitive to departures from the cold dark matter paradigm. A further consequence is that the lighter the seeding mass, the more stringent the resulting constraints on non-standard dark matter models.

Kokkuvõte
Mustad augud ja tumeaine uute infrapuna- ja gravitatsioonilainete vaatluste valguses

Käesoleva doktoritöö eesmärk on kasutada gravitatsioonilaineid ja musti auke tumeaine täppisuuringuteks. Töö põhineb neljal omavahel seotud uurimissuunal, mis vastavad sissejuhatuses sõnastatud neljale uurimiseesmärgile: gravitatsioonilainete laineoptiline murdumine tumeaine struktuuridel; struktuuritekke analüütiline kirjeldus, mis kehtib ka väljaspool külma tumeaine (CDM) mudelit; ülimassiivsete mustade aukude kaksiksüsteemide populatsioon ja gravitatsioonilainete foon nanohertsi sagedusalas; ning tumeaine omaduste analüüs ülimassiivsete mustade aukude ja galaktikate tekke abil universumi esimese miljardi aasta jooksul.

Esimene uurimissuund hõlmab tumeaine uuringuid gravitatsioonilainete gravitatsiooniläätsedes murdumise abil. Artiklis I [61] uurisime Navarro-Frenk-White ehk NFW-halode ja häguse tumeaine (FDM) tuumade avastatavust gravitatsioonilainete mikrolääatseefekti abil, esitades esimese süstemaatilise käsitluse NFW-profiilidest, mille võimendustegurile F puudub analüütiline avaldis, ning esmakordselt analüüsides häguse tumeaine halosid koos neile iseloomulike tuumadega. Selles töös välja töötatud meetodid on vahepeal saanud kogukonnas standardiks. Tähelepanuväärne ja esialgu üllatav tulemus oli, et tulevased tundlikumad detektorid ei paranda avastamisväljavaateid oluliselt; nagu hiljem mõistsime, peegeldab see pigem fundamentaalset füüsikalist kui tehnilist piirangut. Artiklis II [138] laiendasime töö [136] tulemusi ürgsetele mustadele aukudele (PBH), tuletades piirangud Laserinterferomeetrilise gravitatsioonilainete observatooriumi (LIGO) ja Einsteini teleskoobi jaoks. Võtsime arvesse lainekujude ebakõla, PBH-sid ümbritsevate elementaarosakestest koosneva tumeaine halode mõju ning võimaluse selliseid halosid otse avastada. Meie käsitlus PBH-sid ümbritsevatest tumeainehalodest oli kirjanduses esimene täielikult kooskõlaline käsitlus ning LIGO kollaboratsioon viitas sellele metoodikale hiljem oma mikrolääatseefekti signatuuride otsingus [4]. Selle töö oluline järeldus on, et Einsteini teleskoop näeb mustade aukude kaksiksüsteemide populatsiooni kaks korda: üks kord mustade aukude ühinemiste kaudu ja teine kord sama populatsiooni gravitatsioonilainete mikrolääatseefekti kaudu. See võimaldab sõltumatult testida tumeaine profiili nende objektide ümber ning põhimõtteliselt eristada musta augu tähelist või ürgset päritolu isegi siis, kui ürgsete mustade aukude osakaal tumeaines on väike. Lõpuks pöördusime artiklis III [137] tagasi küsimuse juurde, miks tumeainehalod jäävad Laserinterferomeetrilisele kosmoseantennile (LISA) nähtamatuks, ja näitasime, et kitsaskoht tuleneb fundamentaalsest laineoptika efektist: Fresneli ruumala piirab ligipääsetavat tumeaine hulka sõltumata detektori tundlikkusest. Seejärel näitasime, et paljude kergemate tumeainehalode kumulatiivne gravitatsioonilainete murdumis-signaal Fresneli ruumalas on CDM-i raamistikus seevastu garanteeritud ning moodustab uue avastatavate signaalide klassi.

Teine uurimissuund hõlmab James Webbi kosmoseteleskoobi (JWST) suure punanihkega galaktikate ja mustade aukude vaatlusi ning nende tähendust fundamentaalfüüsika jaoks. Artiklis IV [77] arvutasime ainehäirituste dispersiooni aksioni-miniparvede põhjustatud võimenduse korral ja näitasime, et isegi maksimaalne tähetekke efektiivsus, mille puhul kõik barüonid muutuvad tähtedeks, ei olnud vaadeldud galaktikamasside selgitamiseks piisav ning vajalik oli väikesemastaabilise struktuuritekke võimendus. See artikkel peegeldab teadlaste hoiakuid pärast esimesi JWST vaatlusi, mille kohaselt näis vaatlustulemuste seletamiseks olevat vaja mittestandardset füüsikat. Järgnevatel aastatel korrigeerisid galaktikavalimite spektroskoopiline valideerimine ja täpsem fotomeetriline kalibreerimine osa neist varastest tulemustest. Artiklis V [133] viisime läbi oluliselt põhjalikuma

analüüsi, mis hõlmas realistlikku tähetekke kirjeldust, tolmu põhjustatud neeldumist ja nõrka gravitatsioonilist läätseefekti ning kasutas UV-heledusfunktsiooni mõõtmiste märkimisväärselt laienenud andmestikku kuni punanihkeni $z \sim 25$. Tulemus oli mõnevõrra ootamatu: ürgsete mustade aukude või aksioni-miniparvede põhjustatud struktuuritekke võimenduse lisamine sobivust ei parandanud ning ei esinenud olulisi kõrvalekaldeid CDM-ist, kuna häguse ja sooja tumeaine stsenaariumidele seati piirangud. Selle asemel nõudsid andmed kiiret üleminekut III populatsiooni tähtedele ning nõrgemat supernoovade ja aktiivsete galaktikatuumade (AGN) tagasisidet väga suurtel punanihetel; nendel eeldustel saavutati kogu andmestikuga suurepärane kooskõla.

Kolmas uurimissuund hõlmab ülimassiivsete mustade aukude kaksiksüsteemidest pärinevate gravitatsioonilainetest järelduvaid ennustusi lähiuniversumi kohta. Artiklis VI [54] sobitasime ülimassiivsete mustade aukude ühinemistest tekkiva stohhastilise gravitatsioonilainete fooni amplituudi NANOGravi 12,5 aasta andmestikuga ning laiendasime mudelit väiksematele massidele, et teha ennustusi LISA ja aatomi-interferomeetrite jaoks. Kuigi tollased vaatluspiirangud piirdusid mõõdetud amplituudiga, tõi see töö olulise metodoloogilise edasimineku: on märksa tõhusam arvutada Ω_{GW} jaotus mustade aukude ühinemissagedusest ja simuleerida kogusignaali selle jaotuse konvolutsiooni abil kui genereerida üksikuid kaksiksüsteeme otse. See lähenemine näitas ühtlasi, et fooni tegelik jaotus on tugevalt erinev Gaussi jaotusest, pika sabaga suurte Ω_{GW} väärtuste suunas ja moodiga, mis jääb keskväärtusest tunduvalt allapoole, mistõttu fooni üksikud realisatsioonid võivad olla väga teravate tippudega. Artiklis VII [53] rakendasime selle raamistiku täiustatud versiooni NANOGravi 15 aasta andmestikule, mis andis esimest korda lisaks amplituudile ka teavet spektri kuju kohta. See uus vaadeldav suurus näitas, et vaadeldud spektrikuju taasesitamiseks on vajalikud ümbritseva keskkonna efektid, ning oletades, et signaali põhjustab mitteaktiivsete galaktikate populatsioon, saime andmetega suurepärase kooskõla, piirates samal ajal kaksiksüsteemi orbiidi kahanemist põhjustavaid keskkonnamehhanisme. Artiklis VIII [57] laiendasime PTA-mudelit kooskõlaliselt väiksematele massidele ja näitasime, et ülimassiivsete mustade aukude hõive tõenäosuse minimaalne mass, mis on jälg mustade aukude algete tekkemehhanismist väikesel punanihkel, on leitav nii LISA kui ka aatomi-interferomeetritega.

Neljas ja viimane uurimissuund hõlmab ülimassiivsete mustade aukude teket ja varast arengut nende algetest kosmilise ajaloo esimese miljardi aasta jooksul. Artiklis IX [55] uurisime JWST vaadeldud ülimassiivsete mustade aukude ja galaktikate mastaabiseoseid ning väitsime, et näiline vastuolu lokaalsete AGN-ide mastaabiseostega laheneb, kui mõista, et suure punanihke vaatlused on lokaalsete ülimassiivsete mustade aukude mastaabiseoste massiivse otsa loomulik jätk ning et erandiks on hoopis lokaalne AGN-ide valim. Samuti seostasime PTA-dest tuletatud ühinemissagedused ja keskkonnaefektid JWST väikeste punaste täppide vaatlustega ja kaksik-AGN-idest, mida tõlgendatakse ühinevate ülimassiivsete mustade aukudena. Artiklis X [134] töötasime välja semianalüütilise mudeli ülimassiivsete mustade aukude algete arengu jälgimiseks, mille peamine uuendus on halode ühinemiste kirjeldamine diferentsiaalvõrrandite, mitte ühinemispuude abil. See suurendab oluliselt arvutustõhusust ja võimaldab jälgida halode massifunktsiooni kuitahes väikeste massideni, mis on kergete algete stsenaariumi piiramiseks kriitilise tähtsusega. Tõepärafunktsiooni skaneerimine üle algparameetrite näitas, et kerged alged suudavad vaatlusi reprodutseerida vaid siis, kui ühinemiste efektiivsus on suur, samas kui rasked alged sobivad andmetega loomulikumalt. Doktoritöö viimane artikkel XI [58] sünteesib ja laiendab paljusid neist suundadest. See ühendab täiustatud, elliptilisel kollapsil põhineva laiendatud Press–Schechteri formalismi tagasisideparameetritega, mis on sobitatud otse UV-heledusfunktsiooniga, mitte fikseeritud, ning sobitab kõik ülejäänud astrofüüsi-

kalised parameetrid otse ülimassiivsete mustade aukude vaatlustega. Tegemist on seni kõige täielikumalt parametriseeritud mudeliga, mis on sobitatud otse andmetega. Selle keskne tulemus on, et ülimassiivsete mustade aukude vaatlused seavad CDM-ist kõrvalekalletele oluliselt rangemad piirangud kui ainuüksi UV-heledusfunktsioon. Füüsikaline põhjus on selge: kui UV-heledusfunktsioon testib struktuuri teket vaatlushetkel, siis ülimassiivsed mustad augud, mis elavad praktiliselt igavesti, kannavad endas jälgi ühinemistest, mis ulatuvad tagasi nende tekkeni kergetes tumeainehalodes suurel punanihkel, st just sellesse režiimi, mis on külma tumeaine paradigmast kõrvalekallete suhtes kõige tundlikum. Sellest järeldub ühtlasi, et mida kergem on alge mass, seda rangemad on piirangud mittestandardsetele tumeaine mudelitele.

Appendix 1

I

M. Fairbairn, J. Urrutia, and V. Vaskonen. Microlensing of gravitational waves by dark matter structures. *JCAP*, 07:007, 2023

Journal of Cosmology
and Astroparticle
Physics

SISSA

PAPER

Microlensing of gravitational waves by dark matter structures

To cite this article: Malcolm Fairbairn *et al* JCAP07(2023)007

View the article online for updates and enhancements.

You may also like

- Constraints on Compact Dark Matter from Gravitational Wave Microlensing
S. Basak, A. Ganguly, K. Haris et al.
- Gravitational wave lensing: probing Fuzzy Dark Matter with LISA
Shashwat Singh, Guilherme Brando, Stefano Savastano et al.
- First Constraints on Compact Dark Matter from Fast Radio Burst Microstructure
Mawson W. Sammons, Jean-Pierre Macquart, Ron D. Ekers et al.

This content was downloaded from IP address 194.12.167.239 on 02/09/2026 at 13:02

Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Microlensing of gravitational waves by dark matter structures

Malcolm Fairbairn,[a] **Juan Urrutia**[b,c] **and Ville Vaskonen**[b,d,e]

[a]Department of Physics, King's College London,
Strand, WC2R 2LS London, U.K.
[b]National Institute of Chemical Physics and Biophysics,
Rävala 10, Tallinn, Estonia
[c]Division of Physics, Tallinn University of Technology,
Akadeemia tee 21, 12618 Tallinn, Estonia
[d]Dipartimento di Fisica e Astronomia, Università degli Studi di Padova,
Via Marzolo 8, 35131 Padova, Italy
[e]Istituto Nazionale di Fisica Nucleare, Sezione di Padova,
Via Marzolo 8, 35131 Padova, Italy

E-mail: malcolm.fairbairn@kcl.ac.uk, juan.urrutia@kbfi.ee, ville.vaskonen@kbfi.ee

Received November 7, 2022
Revised April 25, 2023
Accepted June 16, 2023
Published July 4, 2023

Abstract. Gravitational lensing of gravitational waves provides a potential new probe of dark matter structures. In this work, we consider the microlensing effect on gravitational wave signals from black hole binaries induced by low-mass dark matter halos that do not retain enough baryonic matter to hold stars. We clarify systematically when this microlensing effect is relevant and study in detail its detectability by future gravitational wave observatories. We consider lensing by cold dark matter halos and by solitonic cores that reside in fuzzy dark matter halos. Our results show that although the effect can be detectable at relatively large impact parameters, the probability of detecting such lensed events is low. In particular, we find that the expected number of events lensed by cold dark matter halos is $\mathcal{O}(0.01)$ per year for BBO and the expected number of events lensed by solitonic cores inside fuzzy dark matter halos is $\mathcal{O}(0.01)$ per year for ET. In the case that a significant fraction of dark matter consists of $\mathcal{O}(100 M_\odot)$ objects that are relatively compact, $R < \mathcal{O}(0.1\,\mathrm{pc})$, we show that the expected number of lensed events per year for ET can be very large, $\mathcal{O}(1000)$.

Keywords: gravitational lensing, dark matter theory, gravitational waves / theory

ArXiv ePrint: 2210.13436

© 2023 IOP Publishing Ltd and Sissa Medialab
https://doi.org/10.1088/1475-7516/2023/07/007

JCAP07(2023)007

Contents

1 Introduction **1**

2 Lensing by dark matter structures **3**
- 2.1 Point mass 4
- 2.2 Uniform density sphere 6
- 2.3 Navarro-Frenk-White halo 7

3 Differentiating the lensed signal from unlensed one **9**
- 3.1 Cold dark matter halos 10
- 3.2 Fuzzy dark matter halos 11

4 Prospects for gravitational wave experiments **13**
- 4.1 Cold dark matter halos 14
- 4.2 Fuzzy dark matter halos 15
- 4.3 Dark matter stars 17

5 Conclusions **18**

1 Introduction

Gravitational wave (GW) observations have opened a new window into the Universe and may uncover deep questions in modern physics. With the LIGO-Virgo observations we have, for example, learned about the black hole (BH) population [1], tested general relativity in the strong regime [2], constrained the graviton mass [2], the speed of GWs [2] and the nuclear matter equation of state [3]. Cosmology with GWs is also promising. Already LIGO-Virgo has been able to measure the Hubble constant [4] and put constraints on GW backgrounds [5–8]. The new more sensitive observatories, including ET [9], CE [10], AEDGE [11, 12], LISA [13], and BBO [14], will in the near future expand the frequency band providing more precise measurements and potentially answer further questions.

One of the most puzzling aspects of cosmology is the elusive nature of dark matter (DM). In this regard, one of the ways that GW observatories may be capable of detecting DM is through the gravitational interaction of GWs with DM substructures. For example, the cold dark matter (CDM) paradigm predicts an abundance of light halos which may act as lenses for a GW passing nearby [15]. Halos lighter than about $10^8 M_\odot$ do not retain enough baryonic matter to hold stars [16] so their detection with conventional techniques is very challenging. Deviations from CDM, which appears to work well down to this mass scale, may appear at these lower masses.

At these low mass scales, structure can be wiped out due to a large free streaming length in the early Universe caused by the DM having a non-zero kinetic energy (i.e. being warm) during structure formation [17]. Similarly, in ultralight DM scenarios, the Compton wavelength of DM can become comparable to astrophysical distances and the formation of halos with characteristic size below this length scale is suppressed in a way closely analogous to the suppression of structure on small scales in warm DM scenarios [18, 19].

JCAP07(2023)007

There are two other ways that dense dark objects created in fuzzy (ultra-light) DM scenarios are different from those created in traditional CDM scenarios which will be particularly relevant for this work — The first is the fact that above that suppression length/mass scale, the necessarily high occupation number of ultra-light DM in dense regions leads to the formation of solitonic cores at the center of dark halos [18, 20, 21]. While such halos host almost constant density cores with lower central densities than might occur in CDM scenarios, the mass as a function of radius can be slightly higher close to the edge of the central solitonic region which can affect gravitational lensing.

The second is related to the particle physics of the situation in the early Universe — In models where the ultra light DM is a Nambu-Goldstone boson of a broken U(1) symmetry, as is the case for axion-like particles if the symmetry is broken after inflation, the Kibble Mechanism results in large differences in field expectation values from Hubble volume to Hubble volume. This scaling with the Hubble volume continues until the resulting energy differences are converted into dense halos at the moment where the mass of the boson becomes comparable to the Hubble expansion rate, and these halos posses larger density contrasts which start to grow earlier than conventional CDM structures emerging from small adiabatic fluctuations on larger scales [22–24].

The evolution of these halos and their subsequent mergers is still a subject of debate, current thought being that they form a spectrum around the characteristic mass imprinted at the stage where the scaling halts [24, 25]. Across a broad mass range, the existence of such cores can be probed through optical microlensing [25, 26].

What is clear is that studying these smaller objects may shed light on the nature of DM and the community is advocating for the importance of finding new ways to probe DM structures at these lower masses by observing the disruption of stellar streams [27, 28] and the detection of smaller halos through lensing of photons [29–31]. In this work, we aim to investigate the possibility of probing such halos using their lensing of GWs.

The theory of gravitational lensing of GWs was originally addressed in [32, 33], and further developed in [34–40]. The interest in this topic has grown in the recent years [41–71] although the searches performed by the LIGO-Virgo collaboration have not shown evidence for lensing signatures [72]. In [62] the non-observation of microlensing effects by LIGO-Virgo was used to derive a new constraint on the abundance of compact DM objects. The major motivation for the lensing studies, however, arises from the proposed new GW observatories. These observatories will detect orders of magnitude more events than LIGO-Virgo, in a wider range of masses and up to much longer distances, and, consequently, several lensed events are expected.

In this paper, we focus on the prospects of detecting microlensing imprints caused by light, $M_{\rm v} < 10^8 M_\odot$, DM halos in the GW signals from black hole binaries. Similar studies done recently in [65, 67] have shown optimistic results with several detectable lensed events per year even in the traditional CDM scenarios. Our work introduces several improvements over the previous studies. First, we clarify in a systematic way what are the conditions under which microlensing effects are relevant. Second, we considered realistic profiles for the lenses, including for the first time the effect of a constant density profile, which is relevant for the fuzzy DM cores. Moreover, instead of the often used singular isothermal sphere (SIS) profile, we consider the more realistic NFW and Einasto profiles for the CDM halos. Third, we perform a systematical matched filtering analysis of the detectability of the microlensing effect and calculate the probability at which a given source is lensed in a way that is detectable accounting for realistic halo mass functions. With these improvements, we find that the

JCAP07(2023)007

probability of seeing events that are microlensed by halos or their solitonic cores is in general very small for all future GW observatories. This is a new and important result for the planning of new gravitational wave observatories.

2 Lensing by dark matter structures

Consider a binary emits GWs at the angular diameter distance D_s, that on their way to the detector pass by a DM structure at the angular diameter distance D_l. The DM structure curves the spacetime causing gravitational lensing of the signal. In this work, we consider microlensing effects that modify the waveform due to interference of the different paths that the GW takes around the DM structure.[1] In the frequency domain the lensed GW signal is $\tilde{\phi}_L(f) = F(f)\tilde{\phi}(f)$, where $\tilde{\phi}(f)$ denotes the signal that would be seen without the lens and $F(f)$ is the amplification factor. The latter can be written in the thin-lens approximation as [76]

$$F(w, \boldsymbol{y}) = \frac{w}{2i\pi} \int \mathrm{d}^2\boldsymbol{x}\, e^{iwT(\boldsymbol{x},\boldsymbol{y})}\,, \tag{2.1}$$

where the integral is over the lens plane. Using a characteristic length scale ξ_0 of the system, the dimensionless vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are defined as $\boldsymbol{x} \equiv \boldsymbol{\xi}/\xi_0$ and $\boldsymbol{y} \equiv D_l\boldsymbol{\eta}/(D_s\xi_0)$ and the dimensionless frequency w and the dimensionless time delay function T as

$$w \equiv \frac{(1+z_l)D_s}{D_l D_{ls}}\xi_0^2 2\pi f\,, \qquad T(\boldsymbol{x},\boldsymbol{y}) \equiv \frac{1}{2}|\boldsymbol{x}-\boldsymbol{y}|^2 - \psi(\boldsymbol{x}) - \phi(\boldsymbol{y})\,, \tag{2.2}$$

where f is the signal frequency, and $D_{ls} = D_s - D_l(1+z_l)/(1+z_s)$ the angular diameter distance between the lens and the source.[2] The vector $\boldsymbol{\eta}$ determines the displacement of the source from the line of sight to the lens, $\boldsymbol{y}$ is its dimensionless projection to the lens plane, and the vector $\boldsymbol{\xi}$ is in the lens plane and determines the point through which the path crosses it. We call $y \equiv |\boldsymbol{y}|$ the impact parameter. The function $\phi(\boldsymbol{y})$ is defined such that the minimum of $T(\boldsymbol{x},\boldsymbol{y})$ with respect to $\boldsymbol{x}$ is zero, $\min_{\boldsymbol{x}} T(\boldsymbol{x},\boldsymbol{y}) = 0$.[3]

The dimensionless deflection potential $\psi(\boldsymbol{x})$ is obtained from the gravitational potential $U(\boldsymbol{r})$ of the lens by integrating over the direction perpendicular to the lens plane. For a spherically symmetric lens potential, $\psi(x)$ can be integrated from [78]

$$\partial_x\psi(x) = \frac{4D_l D_{ls}}{\xi_0^2 D_s}\frac{M_{\mathrm{cyl}}(x)}{x}\,, \tag{2.3}$$

where $x = |\boldsymbol{x}|$. The projected mass in the lens plane within radius x is given by

$$M_{\mathrm{cyl}}(x) = 2\pi\int_0^x dx' \int_{-\infty}^{\infty} \mathrm{d}z\, x'\rho(x', z)\,, \tag{2.4}$$

where z denotes the direction perpendicular to the lens plane and $\rho(x, z)$ the mass density of the lens. The constant of integration from (2.3) can be included in $\phi(y)$.

In the geometric optics limit, i.e. the limit which roughly holds when the wavelength of the signal is much smaller than the Schwarzschild radius of the lens, the amplification factor

[1] For simplicity, we consider isolated spherical lenses. We note, however, that non-sphericities of the lens and larger structures surrounding the lens may affect the observed signal [73–75].

[2] The expression for D_{ls} is justified since the Planck and BAO observations indicate a very low cosmic curvature [77].

[3] $\phi(\boldsymbol{y})$ gives only an overall phase factor for F but ensures that the phase of the signal is same for all potentials.

JCAP07(2023)007

can be approximated by accounting only for the paths that extremize the time delay function, $\nabla_x T(\boldsymbol{x}_j, \boldsymbol{y}) = 0$ or equivalently $\boldsymbol{y} = \boldsymbol{x}_j - \nabla_x \psi(\boldsymbol{x}_j)$. The amplification factor in this limit reduces to

$$F(w, \boldsymbol{y}) = \sum_i |\mu(\boldsymbol{x}_j)|^{1/2} \, e^{iwT(\boldsymbol{x}_j, \boldsymbol{y}) - i\pi a_j} \,. \tag{2.5}$$

The magnification of the extremal path that crosses the lens plane at $\boldsymbol{x} = \boldsymbol{x}_i$ is given by

$$\mu(\boldsymbol{x}_j) = \frac{1}{\det\left[\partial_a \partial_b T(\boldsymbol{x}_j, y)\right]} = \frac{1}{\det\left[\delta_{ab} - \partial_a \partial_b \psi(\boldsymbol{x}_j)\right]} \,, \tag{2.6}$$

and a_j is 0, 1/2 or 1 depending on whether $\boldsymbol{x}_j$ is a minimum, saddle point or a maximum of T, respectively.

If the time difference between different paths is small, they will interfere at the detector. This happens if

$$\frac{t(\boldsymbol{x}, \boldsymbol{y})}{f} \left| \frac{\mathrm{d}f}{\mathrm{d}t} \right| \ll 1 \,, \tag{2.7}$$

where $t(\boldsymbol{x}, \boldsymbol{y}) = \omega T(\boldsymbol{x}, \boldsymbol{y})/(2\pi f)$. As the largest contribution to the integral comes from the paths near the paths $\boldsymbol{x}_j$ that extremize the time delay function, we can estimate if the interference condition (2.7) is satisfied by considering the differences between the paths $\boldsymbol{x}_j$. If the interference condition (2.7) is not satisfied, we enter the strong lensing regime where two or more signals from the same source can be distinguished.

In this work we will focus on the microlensing case where eq. (2.7) is fulfilled. To study the detectability of the lens we look for deformations of the unlensed signal (see section 3). Our analysis is therefore considerably different than in the strong lensing case where we should study whether we can confidently say that the two or more signals are lensed counterparts of the same event [79]. The prospects of seeing strongly lensed GW signals have been studied e.g. in refs. [41–43, 61]. In the following, we will show that the microlensing case is relevant for a broad range of lens and source masses.

2.1 Point mass

In order to get a rough estimate of the lens and source masses for which the interference effect is detectable, we start with the simplest case where the lens can be approximated as a point mass. In this case $M_{\rm cyl}(x) = M_l$ is a constant, and, using eq. (2.3), we get $\psi(x) = \ln x$ by choosing ξ_0 to be the Einstein radius,

$$\xi_0 = \sqrt{\frac{4M_l D_l D_{ls}}{D_s}} \equiv R_E(M_l) \,. \tag{2.8}$$

The dimensionless frequency is $w = 8\pi M_{lz} f$, where $M_{lz} \equiv (1 + z_l) M_l$. For a point mass lens the amplification factor $F(f)$ depends only on two parameters: y and M_{lz}.

The point mass lens potential supports one path that minimizes and one that is a saddle point of the time delay function. These are at $\boldsymbol{x}_\pm(\boldsymbol{y}) = \pm x_\pm(y)\boldsymbol{y}/y$, where $x_\pm(y) = (\pm y + \sqrt{y^2+4})/2$. By taking $\phi(y) = (x_+(y) - y)^2/2 - \ln x_+(y)$, the time delay function for the former path is zero. The integral over $\boldsymbol{x}$ in F can be done in terms of the gamma function and the confluent hypergeometric function of the first kind as [37]

$$F(w, y) = \exp\left[\frac{\pi w}{4} + i\frac{w}{2}\left(\ln\frac{w}{2} - 2\phi(y)\right)\right] \Gamma\left(1 - \frac{iw}{2}\right) {}_1F_1\left(i\frac{w}{2}; 1; \frac{i}{2} w y^2\right) \,. \tag{2.9}$$

JCAP07(2023)007

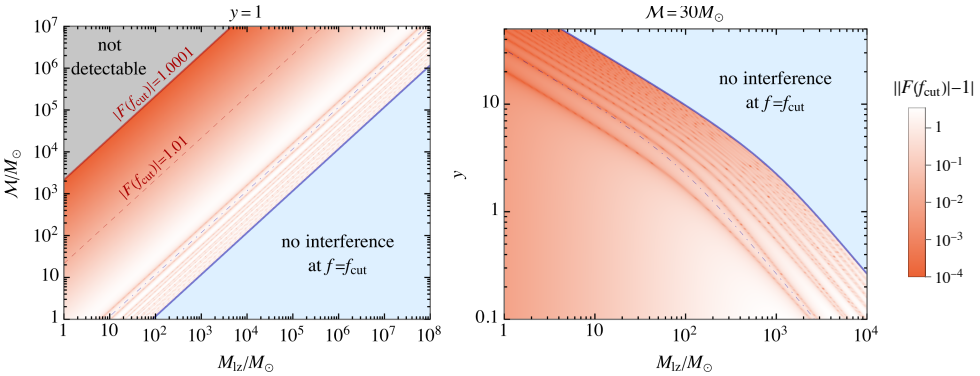


Figure 1. In the light blue region the left-hand side of the interference condition (2.13) is larger than one, and the dotdashed contour indicates where it is 0.1. In the light gray region $|F(f_{\rm cut})| < 1.0001$. The symmetric mass ratio of the source binary is fixed to $\eta = 0.23$.

The magnifications for the geometric optics paths crossing the lens plane at $\boldsymbol{x} = \boldsymbol{x}_\pm$ are $\mu_\pm(y) = 1/2 \pm (2+y^2)/(2y\sqrt{y^2+4})$, and their dimensionless arrival time difference is

$$\Delta T(y) = \frac{y\sqrt{y^2+4}}{2} + \ln\left(\frac{\sqrt{y^2+4}+y}{\sqrt{y^2+4}-y}\right). \tag{2.10}$$

The geometric optics approximation works for $w \gg 1/\Delta T(y)$. The time delay difference between the geometric optics paths is $\Delta t \equiv 4M_{lz}\Delta T$.

Let us check when the interference condition (2.7) is fulfilled. During the inspiral phase of a compact binary coalescence, the frequency of the GW signal evolves as

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{3f}{8(t_c - t)}\,, \tag{2.11}$$

where t_c denotes the time of coalescence of the binary. The frequency changes most rapidly at the end of the inspiral phase which we approximate by the innermost stable circular orbit (ISCO),

$$f_{\rm cut} = \frac{\eta^{3/5}}{6^{3/2}\pi\mathcal{M}}\,, \tag{2.12}$$

where η denotes the symmetric mass ratio of the binary. Thus, at the ISCO frequency, the interference condition (2.7) for a point mass lens is given by

$$\frac{8}{135}\Delta T\,\eta^{8/5}\,\frac{M_{lz}}{\mathcal{M}} \ll 1\,. \tag{2.13}$$

In the light blue region in figure 1 the left-hand side of eq. (2.13) is larger than one, and therefore the interference assumption is violated at the end of the inspiral. As shown in the right panel of figure 1, for $y = \mathcal{O}(1)$ and the sources similar to those that LIGO-Virgo has observed, the interference condition (2.13) is fulfilled as long as $M_{lz} \lesssim 10^3 M_\odot$. For heavier sources, the interference condition holds up to higher lens masses, as shown in the left panel.

For the lensing to be potentially detectable the mass of the lens cannot be several orders of magnitude lighter than the binary as in that case the amplification factor would

JCAP07(2023)007

not significantly deviate from unity. In figure 1 we show by the color coding $||F(f_{\rm cut})| - 1|$ in logarithmic scale. Along the red dashed contour in the left panel this quantity equals 10^{-2} and in the gray region it is smaller than 10^{-4}. The further we go toward the upper left corner of that plot, the more challenging the detection of the microlensing effect becomes. The boundaries of both "no interference" and "not detectable" regions scale as $\mathcal{M} \propto M_{lz}$, and there is a broad range of lens masses that can potentially be probed through microlensing of GW signals from binaries. Notice also, that at times much before the binary reaches the ISCO, the interference condition is satisfied even in the light blue region but if the merger is outside of the sensitivity range of the detector, then also the region that is not detectable moves to higher lens masses.

2.2 Uniform density sphere

A uniform sphere is characterized by its mass M_l and radius R, and the mass density of the lens is

$$\rho(r) = \frac{3M_l}{4\pi R^3}\,\theta(R-r)\,, \tag{2.14}$$

where θ denotes the Heaviside step function. The projected mass within radius r in the lens plane is

$$M_{\rm cyl}(r) = M_l\left[1 - \left(1 - \frac{r^2}{R^2}\right)^{3/2}\theta(R-r)\right]\,. \tag{2.15}$$

Using eq. (2.3), choosing ξ_0 in the same way as for a point mass lens and defining $b \equiv R/\xi_0$, we find that the deflection potential for a uniform density sphere is

$$\psi(x) = \begin{cases} \frac{1}{3}\sqrt{1-\frac{x^2}{b^2}}\left(\frac{x^2}{b^2}-4\right) + \ln\left(b+\sqrt{b^2-x^2}\right)\,, & x \le b\,, \\ \ln x\,, & x > b\,. \end{cases} \tag{2.16}$$

As expected, the deflection potential coincides with that of the point mass lens for $x > b$. The dimensionless frequency is the same as in the point mass case, $w = 8\pi M_{lz} f$. The amplification function can be computed utilizing the analytical result of the point mass approximation (2.9), so that the integral over the lens plane needs to be performed numerically only up to $x = b$.[4]

There can be at most two extremal paths that cross the lens plane outside of the sphere ($x > b$). These are the same paths as for the point mass lens. The amplification factor for the sphere is heavily influenced by whether these paths exist. Also extremal paths that pass through the interior ($x \le b$) of the sphere can exist. These can be easily found numerically. We find that the following three cases can be realized:

- Type I: if
$$0 < b \le 1 \quad \text{and} \quad y \le \frac{1-b^2}{b} \tag{2.17}$$
there are two extremal paths at $x > b$ and one at $x \le b$.

- Type II: if
$$b > 0 \quad \text{and} \quad y > \frac{|1-b^2|}{b} \tag{2.18}$$
there is one extremal path at $x > b$. In a narrow range just above the lower bound of y, there is also two extremal paths at $x < b$.

[4]Notice that $\phi(y)$ still needs to be calculated with the full lens potential (2.16). Given y and b, we resolve $\phi(y)$ numerically.

JCAP07(2023)007

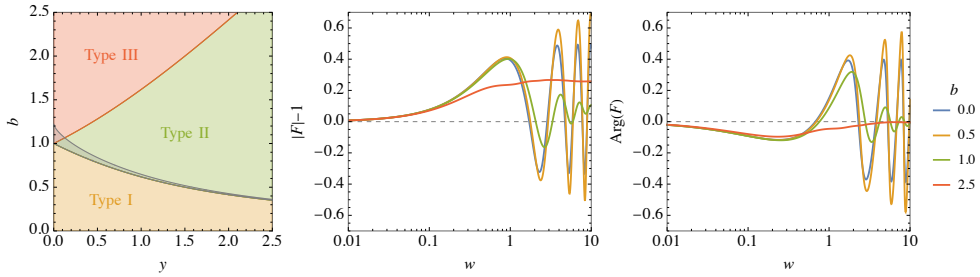


Figure 2. The left panel shows the regions corresponding to the three types of lensing by a uniform sphere discussed in the main text. The middle and right panels show the modulus and argument of the amplification function F as a function of the dimensionless frequency w for different lens sizes b and $y = 1$. The $b = 0$ case corresponds to the point mass lens, $b = 0.5$ is what we refer to as type I, $b = 1$ as type II, and $b = 2.5$ as type III lensing, as shown in the left panel.

- Type III: if

$$b > 1 \quad \text{and} \quad y \leq \frac{b^2 - 1}{b}\,, \tag{2.19}$$

  and there is one extremal path at $x \leq b$. In a small region at $b < \sqrt{3/2}$ two additional extremal paths at exist at $x < b$.

In figure 2 we illustrate these cases: For the most compact objects type I is realized and the amplification function is very close to that of a point mass. The path crossing through the interior of the sphere may, however, boost the amplification, as can be seen by comparing the blue and yellow curves. Instead, for type II, shown by the green curves, the interference effect is smaller than for a point mass lens and the amplification converges to a constant deep in the geometric optics limit since the lens potential supports only one extremal time path. Finally, in the type III case, shown by the red curves, the amplification function does not have any significant oscillations, and the lens causes only a small overall amplification of the signal. However, even in this case the small increase in the amplification as a function of the frequency may give a chance to detect the lens. In the narrow band shown in gray shading in the left panel two additional extremal paths exist that cross through the sphere.

2.3 Navarro-Frenk-White halo

The Navarro-Frenk-White (NFW) density profile is characterized by the scale radius r_s which separates the central $\rho \propto 1/r$ scaling from the outer $\rho \propto 1/r^3$:

$$\rho(r) = \frac{r_s^3 \rho_s}{r\,(r_s + r)^2}\,. \tag{2.20}$$

The projected mass within radius r in the lens plane for the NFW profile is

$$M_{\rm cyl} = 4\pi\rho_s r_s^3 \left[\ln\left(\frac{r}{r_s}\right) + \frac{r_s}{\sqrt{r_s^2 - r^2}}\,\text{arcoth}\left(\frac{r_s}{\sqrt{r_s^2 - r^2}}\right)\right]\,, \tag{2.21}$$

and, using eq. (2.3), we find that the NFW deflection potential is given by

$$\psi(x) = \kappa \left[\ln^2\left(\frac{x}{2b}\right) - \text{artanh}^2\sqrt{1 - \frac{x^2}{b^2}}\right]\,, \tag{2.22}$$

JCAP07(2023)007

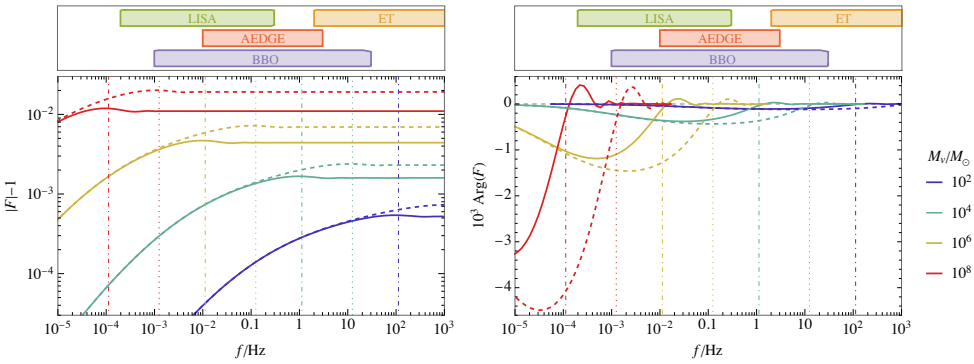


Figure 3. The modulus and the argument of the amplification function F for NFW halo lens as a function of frequency for different virial masses of the halo. The solid and dashed curves correspond to $y = 1$ and $y = 0.3$, respectively, and the source and lens luminosity distances are fixed to $D_{\rm sL} = 5\,{\rm Gpc}$ and $D_{\rm lL} = 2.5\,{\rm Gpc}$. The vertical lines indicate the position of the maximum of $|F|$. The frequency ranges relevant for different experiments are also indicated.

where we have defined dimensionless parameters $\kappa \equiv 2\pi\rho_s r_s^3/M_{\rm v}$ and $b \equiv r_s/\xi_0$. The distances are scaled by $\xi_0 = R_E(M_v)$ and with this choice of ξ_0 the dimensionless frequency is $w = 8\pi(1+z_l)M_{\rm v}f$. For example, for a halo of mass $M_{\rm v} = 10^5 M_\odot$ at luminosity distance $D_{\rm sL} = 1\,{\rm Gpc}$ and a lens at luminosity distance $D_{\rm lL} = 0.5\,{\rm Gpc}$, we get $\xi_0 \approx 2\,{\rm pc}$.

The NFW parameters r_s and ρ_s are related, and can be expressed as functions of the halo viral mass $M_{\rm v}$, $r_s = r_s(M_{\rm v})$ and $\rho_s = \rho_s(M_{\rm v})$. Consequently, the prefactor κ is fixed by $M_{\rm v}$. The parameter b instead depends on $M_{\rm v}$ and the distances to the source and to the lens.

We calculate the amplification function for the NFW lens using the numerical strategy described in [80]. The modulus and the argument of the amplification function are shown for different halo viral masses in figure 3. For $M_{\rm v} < 10^{10} M_\odot$ we find that $\kappa \lesssim 0.3$ and $b > 2$. For these values of κ and b there is only one path that extremizes the time delay T. This path crosses the lens plane at $\boldsymbol{x} \approx \boldsymbol{y}$. As shown in figure 3, the amplification function is similar to the Type III case for a uniform-density sphere lens, showing no clear interference pattern at the detector.

In the geometric optics limit, the effect of the halo is a constant magnification of the signal that biases the apparent distance of the source. The magnification of the signal can be approximated by finding numerically the path that minimizes the time delay function using eq. (2.6), and the frequency above which the wave optics effects are no longer important is $w_{\rm geo} \approx 2/y^2$ which implies $f_{\rm geo} \approx 1/(4\pi y^2(1+z_{\rm l})M_{\rm v})$. This frequency gives approximately the position of the maximum in $|F|$, as indicated by the vertical lines in figure 3.

For $M_{\rm v} > 10^4 M_\odot$ the wave optics regime is at $f \lesssim 1.6\,{\rm Hz}/(y^2(1+z_l))$. Therefore, LIGO and ET, which are sensitive at $f > 1\,{\rm Hz}$, can not distinguish the lensed signal from the unlensed one if the virial mass of the lens halo is $M_{\rm v} < 10^4 M_\odot$, unless $y \ll 1$ which makes the probability of such events very low. Instead, for $M_{\rm v} < 10^4 M_\odot$ the magnification that the NFW halos can generate is for realistic distances and impact parameters, at most $\mathcal{O}(10^{-3})$. This effect is so small that neither LIGO nor ET can distinguish the lensed signal from an unlensed one. Therefore, in the case of NFW halos, we will focus on LISA, AEDGE and BBO, which are sensitive at frequencies $f < 1\,{\rm Hz}$, and can probe the wave optics effects caused by halos with $M_{\rm v} > 10^4 M_\odot$.

JCAP07(2023)007

High-definition galaxy simulations tend to prefer the Einasto profile, especially for low-mass DM halos [81]. The Einasto profile is

$$\rho(r) = \rho_{-2} \exp\left[-\frac{2}{\alpha}\left(\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right)\right], \tag{2.23}$$

where r_{-2} and ρ_{-2} denote radius and density where the logarithm scale slope of the profile is -2. In the same way as for the NFW profile, these are functions of $M_{\rm v}$. For the shape parameter we consider the value $\alpha = 0.158$. The relevant quantities for lensing can be analytically expressed in terms of Fox H function [82]. We have, however, simply computed them numerically. We have found that the Einasto and the NFW profiles are sufficiently similar to not produce any significant differences in the scales relevant to GW lensing. Therefore, in the following we consider only the NFW profile for the CDM halos.

3 Differentiating the lensed signal from unlensed one

Consider data that consists of a GW signal ϕ and noise that we characterize by the noise power spectral density $S_{\rm n}(f)$.[5] The standard matched filtering analysis searches for the parameters $\boldsymbol{\theta}$ of a given template ϕ_T that maximize the log-likelihood [84]

$$\ln \Lambda_T(\phi) = (\phi|\phi_T(\boldsymbol{\theta})) - \frac{1}{2}(\phi_T(\boldsymbol{\theta})|\phi_T(\boldsymbol{\theta}))\,, \tag{3.1}$$

where $(a|b)$ denotes the noise-weighted inner product,

$$(a|b) \equiv 2\int \mathrm{d}f\, \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_{\rm n}(f)}\,. \tag{3.2}$$

For an optimal template $\phi_T = \phi_{\rm opt}$ we can find parameters for which $\phi = \phi_{\rm opt}$. If we know the optimal template, we can calculate how well another template ϕ_T fits the signal by maximizing the log-likelihood (3.1) over the parameters of the template and compare it to the optimal template by calculating the log-likelihood difference

$$\begin{aligned}\Delta\chi^2 &\equiv 2\left[\ln\Lambda_{\rm opt}(\phi) - \max_{\boldsymbol{\theta}} \ln\Lambda_T(\phi)\right] \\ &= \min_{\boldsymbol{\theta}} (\phi_{\rm opt} - \phi_T(\boldsymbol{\theta})|\phi_{\rm opt} - \phi_T(\boldsymbol{\theta})) \\ &= 4\min_{\boldsymbol{\theta}} \int \mathrm{d}f\, \frac{|\tilde{\phi}_{\rm opt}(f) - \tilde{\phi}_T(f;\boldsymbol{\theta})|^2}{S_{\rm n}(f)}\,.\end{aligned} \tag{3.3}$$

Assuming Gaussian statistics, if $\Delta\chi^2 > 4.00, 6.18, 8.02, 9.72$ for one, two, three or four parameter model, respectively, we can differentiate the templates ϕ_T and $\phi_{\rm opt}$ at 2σ confidence level.

We use eq. (3.3) to quantify the detectability of the lens effect in the GW signal from a black hole binary.[6] In this case we take $\phi_{\rm opt}$ as the lensed signal with given lens and source parameters, and try to fit that with the unlensed template. The unlensed template is given by

[5] In addition to the detector noise, we account for the noise coming from the unresolved white dwarf and black hole binary backgrounds [83].

[6] The same method can be used also to study whether different types of lenses can be differentiated from each other. We leave that analysis for future studies.

JCAP07(2023)007

$\tilde{\phi}_T(f;\boldsymbol{\theta}) = A(f;\boldsymbol{\theta})e^{-i\Psi(f;\boldsymbol{\theta})}$. We consider the lowest order approximations for the amplitude $A(f;\boldsymbol{\theta})$ and the phase $\Psi(f;\boldsymbol{\theta})$:

$$A(f;\boldsymbol{\theta}) = \sqrt{\frac{5}{24}}\frac{\mathcal{M}^{5/6}}{\pi^{2/3}D_L}f^{-\frac{7}{6}}\,, \qquad \Psi(f;\boldsymbol{\theta}) = 2\pi f t_c - \phi_c + \frac{3}{128}\left(\pi\mathcal{M}f\right)^{-\frac{5}{3}}\,. \tag{3.4}$$

This template is defined by 4 parameters: $\boldsymbol{\theta} = \{\mathcal{M}, D_L, t_c, \phi_c\}$. The luminosity distance is related to D_s via $D_L = (1+z_s)D_s$, and the binary chirp mass is related to the binary total mass M and the binary mass ratio q by $\mathcal{M} = (1+z_s)M(q/(1+q))^{3/5}$.

The lensed GW signal is given by $\tilde{\phi}_{\rm opt}(f) = F(f;\boldsymbol{\theta}_{l,\rm opt})A(f;\boldsymbol{\theta}_{\rm opt})e^{-i\Psi(f;\boldsymbol{\theta}_{\rm opt})}$, where $F(f;\boldsymbol{\theta}_{l,\rm opt})$ is the lensing amplification factor defined by the lens parameters $\boldsymbol{\theta}_l = \boldsymbol{\theta}_{l,\rm opt}$. To quantify how accurately this signal can be fitted with the unlensed GW template, we evaluate eq. (3.3) finding the minimum over the 4 parameters of the unlensed GW template. We find in all cases that the minimum is reached for $\mathcal{M} \approx \mathcal{M}_{\rm opt}$, and for $t_c \approx t_{c,\rm opt}$. In practice we therefore scan only over $\lambda \equiv (D_L/D_{L,\rm opt})^{-1}$ and $\delta\phi_c \equiv \phi_c - \phi_{c,\rm opt}$. In this case, we can write $\Delta\chi^2$ as

$$\Delta\chi^2 = 4\min_{\lambda,\,\delta\phi_c}\int \mathrm{d}f\,\frac{A(f;\boldsymbol{\theta}_{\rm opt})^2|F(f;\boldsymbol{\theta}_{l,\rm opt}) - \lambda e^{-i\delta\phi_c}|^2}{S_{\rm n}(f)}\,. \tag{3.5}$$

Notice that if the merger does not happen within the sensitivity window of the observatory, then $\Delta\chi^2 \propto \mathcal{M}^{5/3}/D_L^2$.

We evaluate $\Delta\chi^2$ in the frequency range corresponding to at most a year of inspiralling observation finishing at the ISCO orbit. Once $\Delta\chi^2$ is computed as a function of the impact parameter y, it is possible to invert the relation to find the maximum impact parameter $y_{\rm max}$ that leads to a detectable lens effect. We define $y_{\rm max}$ so that for $y < y_{\rm max}$ the lensed template can be separated from the unlensed one by more than 2σ confidence level. As the unlensed template, eq. (3.4), includes 4 parameters, this means that $\Delta\chi^2 > 9.72$ for $y < y_{\rm max}$. The maximal detectable impact parameter is relevant as it determines the detectable lens cross-section and consequently the probability that the lens effect can be detected in a given GW signal. We return to this in section 4. In the following, we compute $y_{\rm max}$ for different lenses, sources, and detectors.

3.1 Cold dark matter halos

For CDM halos, we consider the NFW density profile described in section 2.3. As discussed in the Introduction, the most interesting mass range is $M_{\rm v} \lesssim 10^8 M_\odot$ as these halos don't include many stars and their detection is challenging by conventional techniques. As shown in section 2.3, this in fact is the mass range where the effect of lensing of GW signals by DM halos is within the frequency window of the future GW interferometers.

In section 2.3, we show why LIGO and ET cannot probe the NFW halos through gravitational lensing. Instead, as LISA, AEDGE and BBO are sensitive at lower frequencies, they can probe the lens effects from DM halos of virial mass $M_{\rm v} > 10^4 M_\odot$ for which the lens effect can be relatively strong, about a 1% effect over the unlensed signal. We have computed $\Delta\chi^2(y)$ for differentiating the lens effect from the unlensed template for these experiments, and the corresponding maximal impact parameter $y_{\rm max}$. As shown in figure 4, each of these observatories can detect the lens effect for sources up to $\mathcal{O}$(Gpc) distances with a reasonably large $y_{\rm max}$. The mass ranges in both the chirp mass and the virial mass reflect the frequency ranges of these experiments. The lens effect becomes easier to detect the closer the source is, as for such sources the signal is stronger. The Einstein radius of the system vanishes at the lens distances $D_{\rm lL} = 0$ and at $D_{\rm lL} = D_{\rm sL}$, and consequently such cases are not detectable.

JCAP07(2023)007

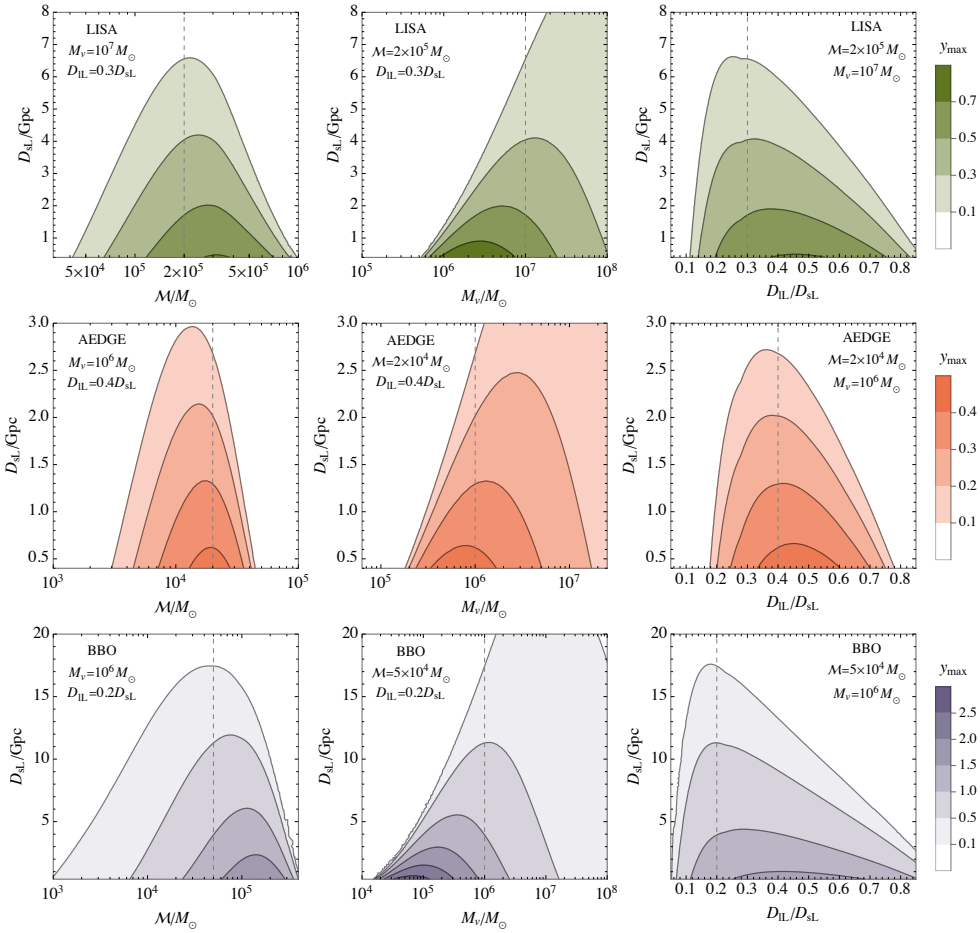


Figure 4. The maximum value of the impact parameter $y_{\rm max}$ that allows for detection of the lens effect by an NFW halo as a function of the binary chirp mass $\mathcal{M}$, the lens halo virial mass $M_{\rm v}$, and the source and lens luminosity distances $D_{\rm sL}$ and $D_{\rm lL}$. The symmetric mass ratio of the binary is fixed to $\eta = 0.23$.

3.2 Fuzzy dark matter halos

DM halos consisting of fuzzy DM are expected to have relatively compact cores. We approximate them with the uniform density sphere deflection potential. In figure 5 we show in the mass-radius plane of the uniform density spheres the maximum impact parameter $y_{\rm max}$ that is detectable with LIGO and ET. If the radius R_c of the object is much smaller than the Einstein radius of a point mass lens of the same mass M_c, $R_c \ll R_E(M)$, then the object behaves as a point mass and $y_{\rm max}$ does not depend on R_c. For $R_c \gtrsim R_E(M)$ the lens effect becomes suppressed, but as shown in figure 5, ET can detect objects whose radius is almost an order of magnitude larger than $R_E(M)$.

JCAP07(2023)007

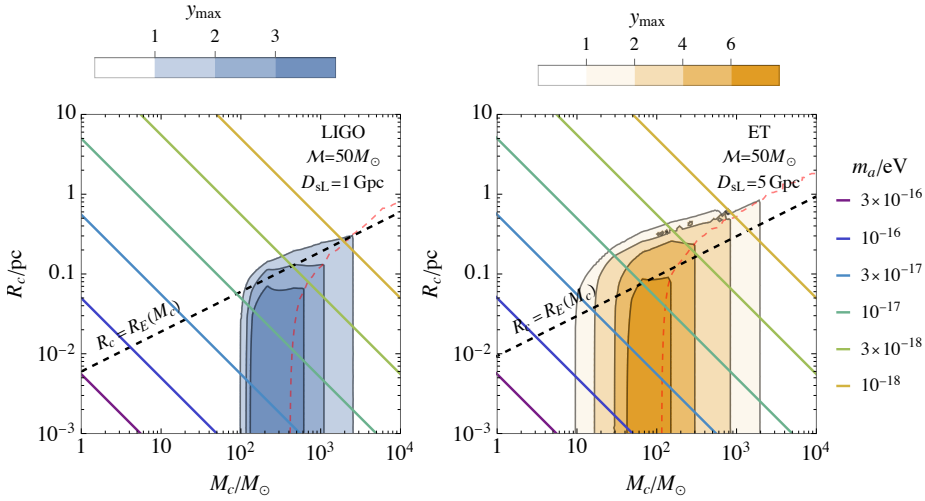


Figure 5. The color shading shows the maximum value of the impact parameter $y_{\max}$ that allows for detection of the lens effect by a uniform density sphere of mass M_c and radius R_c. The binary chirp mass and luminosity distance are fixed as shown in the plot, the symmetric mass ratio of the binary is fixed to $\eta = 0.23$ and the lens distance to $D_{\rm lL} = 0.5 D_{\rm sL}$. The lines show the $M_c - R_c$ relation for DM core for different axion masses m_a, following eq. (3.6). Right from the red dashed curve $y_{\max}$ is set by the interference condition, and the black dashed curve shows the point mass Einstein radius, $R_c = R_E(M)$.

The properties of the cores have been studied by numerical simulations of structure formation of fuzzy DM [20, 85, 86]. These simulations give us predictions for the mass-radius relationships of cores in these halos as well as the core mass/halo mass ratio. There is some uncertainty with regards to the expected ratio between the core and halo mass but the predictions for the mass-radius relations of the cores themselves are more settled and neglecting the scatter found in these mass-radius relationships, we use the fit provided in [20] which gives

$$R_c = 5.0\,\mathrm{pc}\left(\frac{m_a}{10^{-17}\,\mathrm{eV}}\right)^{-2}\left(\frac{M_c}{M_\odot}\right)^{-1}, \tag{3.6}$$

where m_a denotes the fuzzy DM particle mass. We overlap in figure 5 this mass-radius relationship for different fuzzy DM masses. We can see that LIGO is sensitive, assuming some realistic impact parameters, to DM masses from 10^{-16} to 3×10^{-18} eV. ET can instead probe a wider range of masses going down to 10^{-18} eV and lighter less compact cores.

In figure 6 we show the maximum impact parameter as a function of the source parameters and the core mass for a fixed DM mass $m_a = 10^{-17}$ eV. The maximum impact parameter monotonically decreases with the distance and has a maximum in the chirp mass at around $100M_\odot$. This is expected because if the binary is heavier the GW frequency starts to be too low to be detected with ET. The maximum impact parameter grows up to $M_c \gtrsim 100M_\odot$ where the paths take too much time two arrive at the detector and the interaction is no longer satisfying the interference condition.

JCAP07(2023)007

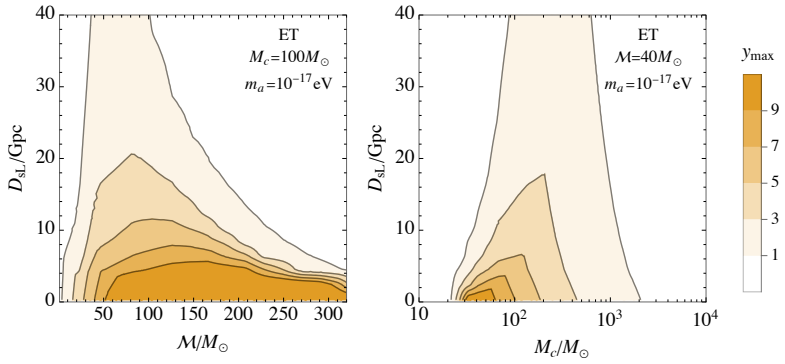


Figure 6. Maximum value of the impact parameter $y_{\max}$ for which ET can detect microlensing by a fuzzy DM core for the DM mass $m_a = 10^{-17}$eV as a function of the luminosity distance and the binary chirp mass (left panel) or the core mass (right panel).

4 Prospects for gravitational wave experiments

The probability for an event to be lensed assuming a Poisson distribution is

$$P_l = 1 - e^{-\tau}\,, \tag{4.1}$$

where τ denotes the optical depth associated with the lensing cross-section. The optical depth is given by

$$\tau(z_s) = \int_0^{D_s} \mathrm{d}V\, n(D_l) \frac{\sigma(D_l)}{4\pi D_l^2} = \int_0^{z_s} \mathrm{d}z_l \frac{n(z_l)\sigma(z_l)}{(1+z_l)H(z_l)}\,, \tag{4.2}$$

where $\mathrm{d}V = 4\pi D_l^2 \mathrm{d}D_l$ denotes the differential volume element, D_l is the angular diameter distance of the lens and z_l the corresponding redshift, n is the number density of the lenses and H the Hubble expansion rate. The cross-section for detecting the lens effect in a given GW signal is

$$\sigma = \pi \xi_{\max}^2 = \pi \xi_0^2 y_{\max}^2\,, \tag{4.3}$$

where $\xi_{\max} = \xi_0 y_{\max}$ is the maximal impact parameter of the source-lens-detector system for which the lens effect is detectable.

Accounting for the present-day halo mass function, $\mathrm{d}n_0/\mathrm{dlog}_{10}M_{\rm v}$, and assuming that the cross-section depends directly on the virial mass of the halo, we get

$$\tau(z_s) = \int_0^{z_s} \mathrm{d}z_l \int \mathrm{dlog}_{10}M_{\rm v} \frac{\mathrm{d}n}{\mathrm{dlog}_{10}M_{\rm v}} \frac{\sigma(z_l, M_{\rm v})}{(1+z_l)H(z_l)}\,. \tag{4.4}$$

In the case of lensing by the cores of fuzzy DM halos, we also need to account for the spread in the halo mass — core mass relationship [86], and the optical depth is given by

$$\tau(z_s) = \int_0^{z_s} \mathrm{d}z_l \int \mathrm{dlog}_{10}M_{\rm v} \frac{\mathrm{d}n}{\mathrm{dlog}_{10}M_{\rm v}} \int \mathrm{dlog}_{10}M_{\rm c}\, f(M_{\rm c}, M_{\rm v}) \frac{\sigma(z_l, M_{\rm c})}{(1+z_l)H(z_l)}\,, \tag{4.5}$$

where $f(M_{\rm c}, M_{\rm v})$ denotes the probability distribution function of the core mass $M_{\rm c}$ for a halo of virial mass $M_{\rm v}$, which is normalized such that $\int \mathrm{dlog}_{10}M_{\rm c}\, f(M_{\rm c}, M_{\rm v}) = 1$.

JCAP07(2023)007

From the lensing probability and the rate of GW events, we can calculate the expected number of events that a given detector will see:

$$N_l = \mathcal{T} \int \mathrm{d}\lambda \, p_{\rm det} \, P_l \,, \tag{4.6}$$

where $\mathcal{T}$ denotes the observation period and $p_{\rm det}$ is the detection probability that accounts for averaging over the binary sky location, binary inclination, and polarization of the signal [87]. The differential merger rate per unit time is given by

$$\mathrm{d}\lambda = \frac{1}{1+z_s} \frac{\mathrm{d}R}{\mathrm{d}m_1 \mathrm{d}m_2} \frac{\mathrm{d}V_c}{\mathrm{d}z_s} \mathrm{d}m_1 \mathrm{d}m_2 \mathrm{d}z_s \,, \tag{4.7}$$

where V_c is the comoving volume. Without the P_l factor eq. (4.6) gives the expected total number of detectable GW events.

4.1 Cold dark matter halos

In the case of CDM halos, which we describe with the NFW profiles we adopt the EPS formalism [88, 89] to calculate the galactic halo mass function,

$$\frac{\mathrm{d}n(M,t)}{\mathrm{d}\ln M} = \frac{\rho_0}{M} \sqrt{\frac{2}{\pi}} \frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M} \frac{\delta_c(z)}{\sigma(M)} e^{-\frac{\delta_c(z)^2}{2\sigma^2(M)}} \,, \tag{4.8}$$

where ρ_0 is the background matter density today, $\sigma^2(M)$ is the variance of the matter fluctuations, $\delta_c(z)$ is given by the linear growth function $D(z)$ as $\delta_c(z) \approx 1.686/D(z)$ [90]. Using the results for $y_{\rm max}$ shown in figure 4 we calculate the probability P_l for a given GW source and detector by eqs. (4.1) and (4.4).

The probability of an event to be lensed and detected by LISA, AEDGE and BBO is shown in figure 7. Although $y_{\rm max}$ increases with decreasing source distance as the signal becomes stronger, the maximal lensing probability is reached at a finite source distance because the number of lenses grows with volume. The maximal probability of detecting lensing by halos of mass $M_{\rm v} < 10^8 M_\odot$ is $P_l \approx 0.8 \times 10^{-3}$ for a source of mass $\mathcal{M} \approx 10^5 M_\odot$ at $D_{\rm sL} \approx 4\,$Gpc for BBO, $P_l \approx 5 \times 10^{-5}$ for a source of mass $\mathcal{M} \approx 3 \times 10^5 M_\odot$ and $D_{\rm sL} \approx 3\,$Gpc for LISA, and $P_l \approx 10^{-5}$ for a source of mass $\mathcal{M} \approx 2 \times 10^4 M_\odot$ at $D_{\rm sL} \approx 3\,$Gpc for AEDGE. For BBO we also show the probability of lensing by halos smaller than $10^6 M_\odot$, which is at most 0.8×10^{-3}.

The merger rate of intermediate-mass BHs (IMBHs) that we have considered as sources is uncertain. The observed dynamics of gas accretion and tidal disruption events provide possible candidates for nearby IMBHs [91–96], but the merger rate and population properties of IMBHs will be known only after enough binaries are detected with GW observatories. The component masses of the heaviest binary seen so far by LIGO-Virgo, GW190521, are $m_1 \approx 85 M_\odot$ and $m_2 \approx 66 M_\odot$ [97]. Searches have been performed for binaries of total mass up to $800 M_\odot$ and the non-observation of such binaries indicates an upper bound on their merger rate. For binaries consisting of two $100 M_\odot$ BHs, the merger rate is smaller than $0.2\,\mathrm{Gpc}^{-3}\mathrm{yr}^{-1}$ [98]. Merger rates have been proposed at the local universe due to repeated merges in dense stellar environments like globular and nuclear clusters [99–103], but these are mostly mergers of IMBHs with stellar mass BHs. In this work, we have focused on the merger of BHs of a similar mass. Halo mergers may generate such IMBH binaries and consequently, their merger rate follows from the halo merger rate [104–106]. Using the expected merger

JCAP07(2023)007

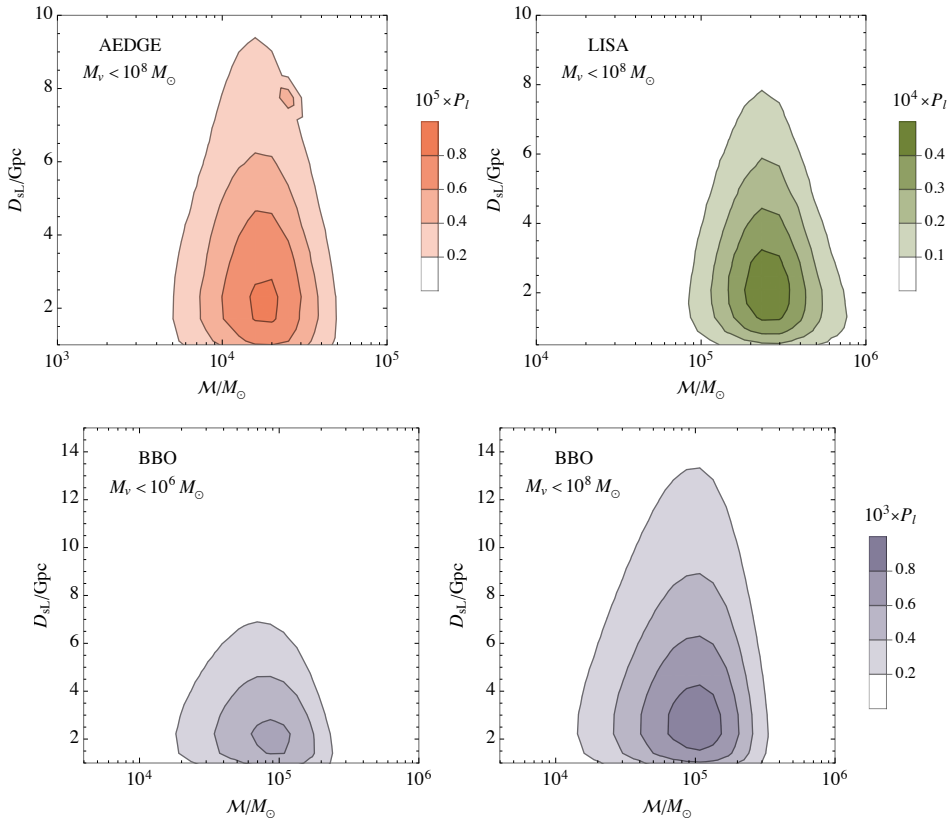


Figure 7. Probability P_l that a GW signal emitted by a binary of chirp mass $\mathcal{M}$ and symmetric mass ratio $\eta = 0.23$ at distance $D_{\rm sL}$ is lensed by a NFW halo. The probabilities are integrated over the virial mass range shown in each panel with the CDM halo mass function.

rate for IMBHs predicted by halo merger theory in ref. [104], we find that LISA is expected to see 3×10^{-7}, AEDGE 1×10^{-5} and BBO 1×10^{-2} events per year that are lensed by halos of mass $M_{\rm v} < 10^8 M_\odot$. When integrating the probability over those halos that contain hardly any baryonic matter $(M_{\rm v} < 10^6 M_\odot)$, the number of expected events drops to 1×10^{-3} events per year for BBO.

4.2 Fuzzy dark matter halos

In the fuzzy DM scenarios, the halo mass function may deviate significantly from the CDM case. First, the small scale fluctuations are suppressed which results in a low mass cut-off of the halo mass function at $M_v \sim (m_a/10^{-15}{\rm eV})^{-3/2} M_\odot$ (see e.g. [107]). Second, in addition to the structures formed after matter-radiation equality, which we assume to follow the mass function (4.8), miniclusters that can have formed before matter-radiation equality contributes to the total halo mass function. This is relevant for the axionic fuzzy DM scenarios where the Peccei-Quinn symmetry breaks after the cosmic inflation [22]. Using the results presented in ref. [25], we find that the present day the DM halos generated from these miniclusters give

JCAP07(2023)007

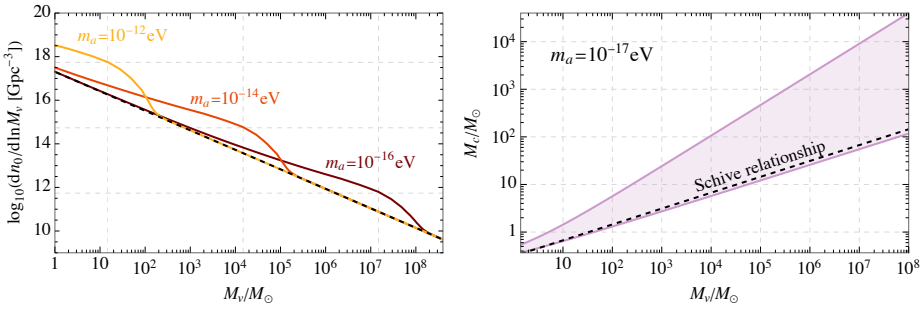


Figure 8. *Left panel:* The present-day fuzzy DM halo mass function from [25] with a bump caused by miniclusters formed before matter-radiation equality. The dashed line shows the CDM halo mass function (4.8). *Right panel:* The spread in the fuzzy DM core mass — halo mass relationship from [86].

rise to a bump in the halo mass function at $M_v \simeq 5\times10^5(m_a/10^{-15}\mathrm{eV})^{-3/2}M_\odot$ (more than 5 orders of magnitude above the cut-off mass) that exceeds the halo mass function (4.8) by factor $\simeq 10(m_a/10^{-15}\mathrm{eV})^{0.15}$. In the left panel of figure 8, we show this bump in the halo mass function for different values of the axion mass m_a.

Moreover, the fuzzy DM halos have solitonic cores that we approximate with a sphere of constant density and radius given by eq. (3.6). The core mass M_c can have a range of values for a given halo mass M_v. We estimate the scatter of the core masses using the results of ref. [86] assuming that the distribution of the core masses is uniform in logarithmic scale in the range $M_{\mathrm{c},1} < M_\mathrm{c} < M_{\mathrm{c},2}$ given by the fitting formula

$$\frac{M_{\mathrm{c},j}}{M_\odot} = \beta_j\left(\frac{m_a}{10^{-17}\mathrm{eV}}\right)^{-3/2} + \gamma_j\left(\frac{m_a}{10^{-17}\mathrm{eV}}\right)^{3(\alpha_j-1)/2}\left(\frac{M_\mathrm{v}}{10^6 M_\odot}\right)^{\alpha_j} \tag{4.9}$$

with $\alpha_1 = 0.33$, $\beta_1 = 0.062$ and $\gamma_1 = 26$, and $\alpha_2 = 0.64$, $\beta_2 = 0.19$ and $\gamma_2 = 2.1\times10^3$. We show this mass range for a benchmark value of m_a in the right panel of figure 8. For comparison, the black dashed line shows the relationship obtained in ref. [20], $M_\mathrm{c} = 31(m_a/10^{-17}\mathrm{eV})^{-1}(M_\mathrm{v}/10^6 M_\odot)^{1/3}$.

As discussed in section 2.1, the microlensing effect is negligible if the binary is much heavier than the lens. So, even if there are solitonic cores inside fuzzy DM halos, we will take into account only the lensing by the DM halo when considering heavy binaries that can be seen with LISA, AEDGE and BBO. For DM masses between 10^{-13} eV up to 10^{-16} eV the halo mass function is bumped at scales that can be probed with these observatories (see figure 3 and the left panel of figure 8). In figure 9 we show the lensing probability for BBO for the halo mass function containing both the CDM halos and the bump from miniclusters for three different DM masses m_a. By comparing to figure 7, we see that the minicluster bump in the halo mass function increases the lensing probability by $\mathcal{O}(1)$ factor.

We focus on LIGO and ET to detect the cores inside the fuzzy DM halos since they are sensitive to the masses where the cores may be more abundant. The effect of the halo surrounding the core is negligible since for the frequencies of LIGO and ET the effect is in the geometric limit and the constant magnification of the signal would bias the luminosity distance by less than 10^{-3} (see figure 3). The lighter m_a is, the heavier the cores are per halo so there is more abundance of heavy cores. In figure 5 we can see that for $m_a < 10^{-18}$ eV the cores are not compact enough to be detectable. We show in figure 10 the probability

JCAP07(2023)007

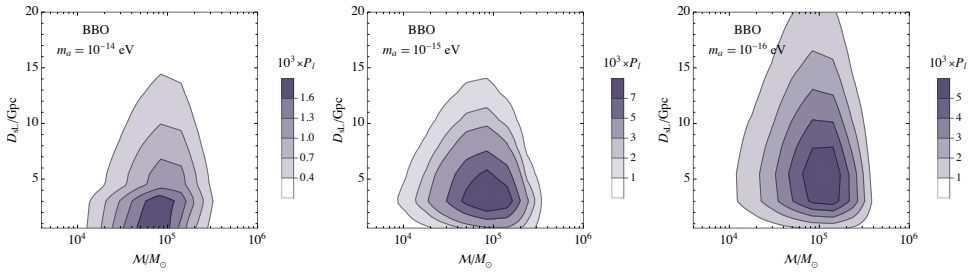


Figure 9. Probability of lensing by a NFW DM halo with $M_{\rm v} < 10^8 M_\odot$ considering the contribution of miniclusters for three fuzzy DM masses.

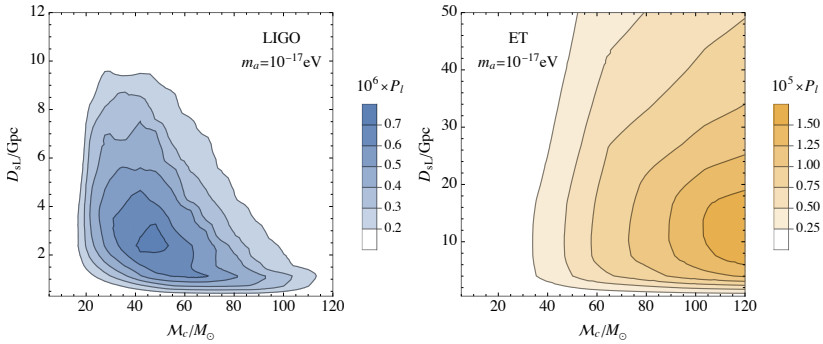


Figure 10. Probability to detect the lensing by fuzzy DM cores for $m_a = 10^{-17}$ eV as a function of the source parameters.

to detect the cores for $m_a = 10^{-17}$ eV, both with LIGO and ET. Roughly, this DM mass leads to the biggest microlensing probability because if DM mass is lighter, then the cores are not compact enough and if it is heavier then the lensing quickly enters the regime where the interference condition is violated.

We have a good understanding of the merger rates of the BBH that LIGO and ET may detect. To estimate the merger rate, we use the results from the most recent LIGO-Virgo catalog [1] and assume a redshift dependence that follows the star formation rate [108],

$$\mathrm{SFR}(z) \propto \frac{(1+z)^{2.7}}{1+[(1+z)/2.9]^{5.6}}\,, \tag{4.10}$$

which is in excellent agreement with the data. The expected number of detectable lensed events for $m_a = 10^{-17}$eV is $0.001^{+0.003}_{-0.0005}$ for LIGO and $0.03^{+0.04}_{-0.02}$ for ET during a year of observation. The uncertainties in the expected number of detectable lensed events reflect the uncertainties in the LIGO-Virgo merger rate.

4.3 Dark matter stars

In order to demonstrate The reason why the expected number of lensed events in the above cases is small is that there isn't sufficiently many lens halos. To demonstrate that an increase in the abundance of lens objects indeed leads to large number of detectable lensed events, we

JCAP07(2023)007

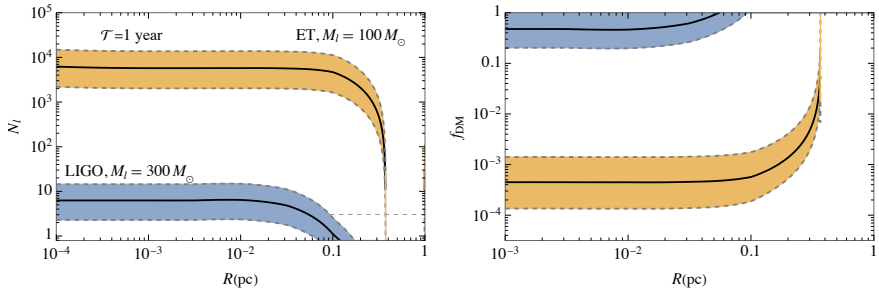


Figure 11. *Left panel:* Expected number of lensed events assuming that all DM is in the form of compact objects of radius R and mass $M_{\rm l} = 100 M_\odot$ during a year of observation with ET (yellow) or with LIGO (blue). The error bands reflect the uncertainties in the BH merger rate. *Right panel:* Projected constraints from the non-observation of microlensing signatures.

consider a scenario where fraction $f_{\rm DM}$ of all DM is in objects that can be approximated by a constant density sphere profile, that is, a fraction of DM constitutes DM stars.

From figure 5, we see that the best case to detect the microlensing effect induced by such objects is at $M_l \simeq 100 M_\odot$ for ET and at $M_l \simeq 300 M_\odot$ for LIGO. Taking these values as benchmark cases, we show in the left panel of figure 11 the expected number of detectable microlensed events assuming they constitute all DM, $f_{\rm DM} = 1$, as a function of their radius R. Here we have used again the merger rate indicated by the LIGO-Virgo observations. We see that if these objects are sufficiently compact LIGO will in this case see $\mathcal{O}(1)$ lensed event per year and ET $\mathcal{O}(1000)$ lensed events per year.[7] The finite size of the lens becomes relevant at $R > 0.01\,{\rm pc}$ for LIGO and at $R > 0.1\,{\rm pc}$ for ET, and the expected number of lensed events quickly falls to zero. In the right panel of figure 11 we show the projected sensitivity of these experiments on the fraction of DM in these DM stars as a function of their radius. If ET will not see any events, it will be able to put a strong constraint, $f_{\rm DM} \leq 10^{-3}$, on the abundance of such objects if their radius is $R < 0.37\,{\rm pc}$. The projected constraint from LIGO in its design sensitivity may reach $f_{\rm DM} \simeq 0.5$, depending on the BH merger rate.

5 Conclusions

In this work, we have considered microlensing of GWs from BH coalescence events by small dark halos along the path of the wave as a probe of small scale DM structures. We have shown that for the consistency of the microlensing formalism, the classical paths that the GW takes around the DM structure should get to the detector at approximately the same frequency. This imposes constraints on the values of the impact parameter y and the lens mass M_{lz}, as shown in figure 1 when we impose the microlensing interference condition up to the ISCO orbit. For the microlensing effect of the lens to be relevant, we have also shown in figure 1 that the mass lens cannot be much lighter than the chirp mass of the binary $\mathcal{M}$, as otherwise, the classical paths do not produce any interference effect at the detector. The combination of both effects imposes a range of lens masses for a given y and $\mathcal{M}$ where the microlensing

[7] Notice that, our analysis includes only the inspiral phase, whereas, in particular for LIGO, the merger and ringdown phases are important and including them would increase the expected number of detectable events. We leave a study going beyond the inspiral phase for future work.

JCAP07(2023)007

formalism describes an interference effect that could be detectable. The interference can happen in the geometric limit or in the wave optics regime. The former correctly describes the effect roughly when the wavelength of the GW is smaller than the Schwarzschild radius of the lens.

First, we have considered microlensing by CDM halos. This search is relevant for halos lighter than $10^8 M_\odot$, where the number of stars is small enough to challenge conventional techniques. We have approximated the CDM halos by NFW profiles, which we have checked to be close enough to the Einasto one that is slightly preferred from numerical simulations. Such light spherically symmetric DM halos are not compact enough to make more than one classical path around them and in this case only the wave effects around the shortest path may produce a detectable effect. We have checked that only AEDGE, LISA and BBO are going to be sensitive to the wave effect for halos of virial mass in the range from $10^4 M_\odot$ to $10^8 M_\odot$, see figure 3. For these detectors, we have computed, assuming the CDM halo mass function, what is the probability for an event to be lensed. We have shown in figure 7 that the maximal probabilities for LISA, AEDGE and BBO are, respectively 5×10^{-5}, 10^{-5}, and 1.3×10^{-3}. There are big uncertainties in the merger rate of IMBHs that these observatories may detect. We show that for current estimates, even for BBO the number of lensed events a year is small, $\mathcal{O}(0.01/\text{year})$.

Second, we have considered fuzzy DM halos, which are expected to host solitonic cores that we have approximated by uniform density sphere profiles. Although it is not possible to find analytical expressions for the microlensing effect of the cores, we have found three main lensing types depending on how many classical paths exist outside the sphere. The different regimes are shown in figure 2. The finite size effects are relevant for the interaction with the cores if their radius is larger or comparable to the Einstein radius. In particular, we have shown in figure 5 that ET can probe the microlensing effect for lenses with sizes up to an order magnitude bigger than their Einstein radius. The relevant fuzzy DM masses for detecting the microlensing effect by LIGO or ET range roughly from $10^{-18}\,\text{eV}$ to $10^{-16}\,\text{eV}$. We have shown that the maximal probability to detect the microlensing is 10^{-6} for LIGO and 3×10^{-5} for ET, realized for DM mass $m_a \simeq 10^{-17}\text{eV}$. Using the merger rate of $\mathcal{O}(10 M_\odot)$ BHs indicated by the LIGO-Virgo observations and assuming that the redshift dependence follows the star formation rate, we have estimated the number of events lensed by the cores. For the best case DM mass, the expected number of lensed events is $\mathcal{O}(0.01/\text{year})$. If fuzzy DM is in the form of axion-like particles and the PQ symmetry was broken before matter-radiation equality, we expect to see deviations from the CDM halo mass function since there are contributions also of axion miniclusters. This, however, induces only an $\mathcal{O}(1)$ increase in the expected number of lensed events.

Third, we have considered microlensing by DM stars. We have shown that, assuming their mass is in the range detectable with LIGO or ET (see figure 5), the expected number of detectable microlensed events can be large, $\mathcal{O}(10^3)$ for ET and $\mathcal{O}(1)$ for LIGO. The finite radius of these stars becomes relevant for ET at $R > 0.1\,\text{pc}$ and for LIGO at $R > 0.01\,\text{pc}$, and the microlensing effects become undetectable if the radius is significantly larger than that. As shown in figure 11, the large expected number of lensed events turns into a strong constraint on the abundance of these objects, in case such microlensing events will not be observed.

This work shows that probing dark structures through microlensing of GWs is difficult. With developments in future sensitivity, it may be possible to discover compact dark structures, and DM scenarios that predict compact objects in the stellar mass range can lead to predictions using this methodology. In the case of relatively compact DM objects, such as fuzzy DM

JCAP07(2023)007

cores, the lensing probability is limited by imposing the microlensing interference condition. This implies that accounting for the strong lensing regime may help in detecting such objects. We leave the study of the detectability of the lens effects ranging from microlensing to strong lensing due to compact DM structures for future work. Moreover, whereas the probability of seeing any individual events lensed by NFW halos is very low, it may be possible to see the lensing effects statistically, in the way of weak lensing, by studying the whole population of gravitational wave events. We leave also such studies for future work.

Acknowledgments

The authors are extremely grateful for conversations with Doddy Marsh. This work was supported by European Regional Development Fund through the CoE program grant TK133 and by the Estonian Research Council grant PRG803. The work of VV has been partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736. MF gratefully received funding via the STFC particle theory grant STFC-ST/T000759/1.

References

[1] KAGRA et al. collaborations, *Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3*, *Phys. Rev. X* **13** (2023) 011048 [arXiv:2111.03634] [INSPIRE].

[2] LIGO Scientific et al. collaborations, *Tests of General Relativity with GWTC-3*, arXiv:2112.06861 [INSPIRE].

[3] LIGO Scientific and Virgo collaborations, *GW170817: Measurements of neutron star radii and equation of state*, *Phys. Rev. Lett.* **121** (2018) 161101 [arXiv:1805.11581] [INSPIRE].

[4] LIGO Scientific et al. collaborations, *Constraints on the Cosmic Expansion History from GWTC-3*, *Astrophys. J.* **949** (2023) 76 [arXiv:2111.03604] [INSPIRE].

[5] LIGO Scientific et al. collaborations, *Constraints on Cosmic Strings Using Data from the Third Advanced LIGO-Virgo Observing Run*, *Phys. Rev. Lett.* **126** (2021) 241102 [arXiv:2101.12248] [INSPIRE].

[6] KAGRA et al. collaborations, *Upper limits on the isotropic gravitational-wave background from Advanced LIGO and Advanced Virgo's third observing run*, *Phys. Rev. D* **104** (2021) 022004 [arXiv:2101.12130] [INSPIRE].

[7] A. Romero-Rodriguez et al., *Search for a Scalar Induced Stochastic Gravitational Wave Background in the Third LIGO-Virgo Observing Run*, *Phys. Rev. Lett.* **128** (2022) 051301 [arXiv:2107.11660] [INSPIRE].

[8] A. Romero et al., *Implications for First-Order Cosmological Phase Transitions from the Third LIGO-Virgo Observing Run*, *Phys. Rev. Lett.* **126** (2021) 151301 [arXiv:2102.01714] [INSPIRE].

[9] M. Punturo et al., *The Einstein Telescope: A third-generation gravitational wave observatory*, *Class. Quant. Grav.* **27** (2010) 194002 [INSPIRE].

[10] D. Reitze et al., *Cosmic Explorer: The U.S. Contribution to Gravitational-Wave Astronomy beyond LIGO*, *Bull. Am. Astron. Soc.* **51** (2019) 035 [arXiv:1907.04833] [INSPIRE].

[11] AEDGE collaboration, *AEDGE: Atomic Experiment for Dark Matter and Gravity Exploration in Space*, *EPJ Quant. Technol.* **7** (2020) 6 [arXiv:1908.00802] [INSPIRE].

JCAP07(2023)007

[12] L. Badurina et al., *Prospective sensitivities of atom interferometers to gravitational waves and ultralight dark matter*, *Phil. Trans. A. Math. Phys. Eng. Sci.* **380** (2021) 20210060 [arXiv:2108.02468] [INSPIRE].

[13] LISA collaboration, *Laser Interferometer Space Antenna*, arXiv:1702.00786 [INSPIRE].

[14] J. Crowder and N.J. Cornish, *Beyond LISA: Exploring future gravitational wave missions*, *Phys. Rev. D* **72** (2005) 083005 [gr-qc/0506015] [INSPIRE].

[15] J. Diemand et al., *Clumps and streams in the local dark matter distribution*, *Nature* **454** (2008) 735 [arXiv:0805.1244] [INSPIRE].

[16] T. Sawala et al., *The chosen few: the low-mass haloes that host faint galaxies*, *Mon. Not. Roy. Astron. Soc.* **456** (2016) 85 [arXiv:1406.6362] [INSPIRE].

[17] P. Bode, J.P. Ostriker and N. Turok, *Halo formation in warm dark matter models*, *Astrophys. J.* **556** (2001) 93 [astro-ph/0010389] [INSPIRE].

[18] L. Hui, J.P. Ostriker, S. Tremaine and E. Witten, *Ultralight scalars as cosmological dark matter*, *Phys. Rev. D* **95** (2017) 043541 [arXiv:1610.08297] [INSPIRE].

[19] K.K. Rogers and H.V. Peiris, *Strong Bound on Canonical Ultralight Axion Dark Matter from the Lyman-Alpha Forest*, *Phys. Rev. Lett.* **126** (2021) 071302 [arXiv:2007.12705] [INSPIRE].

[20] H.-Y. Schive, T. Chiueh and T. Broadhurst, *Cosmic Structure as the Quantum Interference of a Coherent Dark Wave*, *Nature Phys.* **10** (2014) 496 [arXiv:1406.6586] [INSPIRE].

[21] D.J.E. Marsh and A.-R. Pop, *Axion dark matter, solitons and the cusp-core problem*, *Mon. Not. Roy. Astron. Soc.* **451** (2015) 2479 [arXiv:1502.03456] [INSPIRE].

[22] E.W. Kolb and I.I. Tkachev, *Axion miniclusters and Bose stars*, *Phys. Rev. Lett.* **71** (1993) 3051 [hep-ph/9303313] [INSPIRE].

[23] E.W. Kolb and I.I. Tkachev, *Large amplitude isothermal fluctuations and high density dark matter clumps*, *Phys. Rev. D* **50** (1994) 769 [astro-ph/9403011] [INSPIRE].

[24] D. Ellis et al., *Structure of axion miniclusters*, *Phys. Rev. D* **106** (2022) 103514 [arXiv:2204.13187] [INSPIRE].

[25] M. Fairbairn, D.J.E. Marsh, J. Quevillon and S. Rozier, *Structure formation and microlensing with axion miniclusters*, *Phys. Rev. D* **97** (2018) 083502 [arXiv:1707.03310] [INSPIRE].

[26] M. Fairbairn, D.J.E. Marsh and J. Quevillon, *Searching for the QCD Axion with Gravitational Microlensing*, *Phys. Rev. Lett.* **119** (2017) 021101 [arXiv:1701.04787] [INSPIRE].

[27] N. Banik and J. Bovy, *Effects of baryonic and dark matter substructure on the Pal 5 stream*, *Mon. Not. Roy. Astron. Soc.* **484** (2019) 2009 [arXiv:1809.09640] [INSPIRE].

[28] N. Banik et al., *Evidence of a population of dark subhaloes from Gaia and Pan-STARRS observations of the GD-1 stream*, *Mon. Not. Roy. Astron. Soc.* **502** (2021) 2364 [arXiv:1911.02662] [INSPIRE].

[29] D. Gilman et al., *Probing dark matter structure down to 10^7 solar masses: flux ratio statistics in gravitational lenses with line-of-sight haloes*, *Mon. Not. Roy. Astron. Soc.* **487** (2019) 5721 [arXiv:1901.11031] [INSPIRE].

[30] M. Fairbairn, *Galactic Anomalies and Particle Dark Matter*, *Symmetry* **14** (2022) 812 [INSPIRE].

[31] J.W. Nightingale et al., *Scanning For Dark Matter Subhalos in Hubble Space Telescope Imaging of 54 Strong Lenses*, arXiv:2209.10566 [INSPIRE].

[32] J.K. Lawrence, *Focusing of gravitational radiation by the galactic core*, *Phys. Rev. D* **3** (1971) 3239 [INSPIRE].

JCAP07(2023)007

[33] H.C. Ohanian, *On the focusing of gravitational radiation*, *Int. J. Theor. Phys.* **9** (1974) 425 [INSPIRE].

[34] Y. Wang, A. Stebbins and E.L. Turner, *Gravitational lensing of gravitational waves from merging neutron star binaries*, *Phys. Rev. Lett.* **77** (1996) 2875 [astro-ph/9605140] [INSPIRE].

[35] T.T. Nakamura, *Gravitational lensing of gravitational waves from inspiraling binaries by a point mass lens*, *Phys. Rev. Lett.* **80** (1998) 1138 [INSPIRE].

[36] A.F. Zakharov and Y.V. Baryshev, *Influence of gravitational lensing on sources of gravitational radiation*, *Class. Quant. Grav.* **19** (2002) 1361 [INSPIRE].

[37] R. Takahashi and T. Nakamura, *Wave effects in gravitational lensing of gravitational waves from chirping binaries*, *Astrophys. J.* **595** (2003) 1039 [astro-ph/0305055] [INSPIRE].

[38] N. Seto, *Strong gravitational lensing and localization of merging massive black hole binaries with LISA*, *Phys. Rev. D* **69** (2004) 022002 [astro-ph/0305605] [INSPIRE].

[39] N. Matsunaga and K. Yamamoto, *The finite source size effect and the wave optics in gravitational lensing*, *JCAP* **01** (2006) 023 [astro-ph/0601701] [INSPIRE].

[40] M. Sereno et al., *Strong lensing of gravitational waves as seen by LISA*, *Phys. Rev. Lett.* **105** (2010) 251101 [arXiv:1011.5238] [INSPIRE].

[41] A. Piórkowska, M. Biesiada and Z.-H. Zhu, *Strong gravitational lensing of gravitational waves in Einstein Telescope*, *JCAP* **10** (2013) 022 [arXiv:1309.5731] [INSPIRE].

[42] M. Biesiada, X. Ding, A. Piórkowska and Z.-H. Zhu, *Strong gravitational lensing of gravitational waves from double compact binaries — perspectives for the Einstein Telescope*, *JCAP* **10** (2014) 080 [arXiv:1409.8360] [INSPIRE].

[43] X. Ding, M. Biesiada and Z.-H. Zhu, *Strongly lensed gravitational waves from intrinsically faint double compact binaries — prediction for the Einstein Telescope*, *JCAP* **12** (2015) 006 [arXiv:1508.05000] [INSPIRE].

[44] L. Dai, T. Venumadhav and K. Sigurdson, *Effect of lensing magnification on the apparent distribution of black hole mergers*, *Phys. Rev. D* **95** (2017) 044011 [arXiv:1605.09398] [INSPIRE].

[45] G.P. Smith et al., *What if LIGO's gravitational wave detections are strongly lensed by massive galaxy clusters?*, *Mon. Not. Roy. Astron. Soc.* **475** (2018) 3823 [arXiv:1707.03412] [INSPIRE].

[46] K.K.Y. Ng, K.W.K. Wong, T. Broadhurst and T.G.F. Li, *Precise LIGO Lensing Rate Predictions for Binary Black Holes*, *Phys. Rev. D* **97** (2018) 023012 [arXiv:1703.06319] [INSPIRE].

[47] S. Jung and C.S. Shin, *Gravitational-Wave Fringes at LIGO: Detecting Compact Dark Matter by Gravitational Lensing*, *Phys. Rev. Lett.* **122** (2019) 041103 [arXiv:1712.01396] [INSPIRE].

[48] K. Liao et al., *Precision cosmology from future lensed gravitational wave and electromagnetic signals*, *Nature Commun.* **8** (2017) 1148 [*Erratum ibid.* **8** (2017) 2136] [arXiv:1703.04151] [INSPIRE].

[49] S.-S. Li, S. Mao, Y. Zhao and Y. Lu, *Gravitational lensing of gravitational waves: A statistical perspective*, *Mon. Not. Roy. Astron. Soc.* **476** (2018) 2220 [arXiv:1802.05089] [INSPIRE].

[50] M. Oguri, *Effect of gravitational lensing on the distribution of gravitational waves from distant binary black hole mergers*, *Mon. Not. Roy. Astron. Soc.* **480** (2018) 3842 [arXiv:1807.02584] [INSPIRE].

[51] G. Congedo and A. Taylor, *Joint cosmological inference of standard sirens and gravitational wave weak lensing*, *Phys. Rev. D* **99** (2019) 083526 [arXiv:1812.02730] [INSPIRE].

[52] L. Dai et al., *Detecting Lensing-Induced Diffraction in Astrophysical Gravitational Waves*, *Phys. Rev. D* **98** (2018) 104029 [arXiv:1810.00003] [INSPIRE].

JCAP07(2023)007

[53] T. Broadhurst, J.M. Diego and G. Smoot, *Reinterpreting Low Frequency LIGO/Virgo Events as Magnified Stellar-Mass Black Holes at Cosmological Distances*, arXiv:1802.05273 [INSPIRE].

[54] T. Broadhurst, J.M. Diego and G.F. Smoot, *Twin LIGO/Virgo Detections of a Viable Gravitationally-Lensed Black Hole Merger*, arXiv:1901.03190 [INSPIRE].

[55] S. Cao et al., *Direct test of the FLRW metric from strongly lensed gravitational wave observations*, *Sci. Rep.* **9** (2019) 11608 [arXiv:1910.10365] [INSPIRE].

[56] S. Cao et al., *Measuring the viscosity of dark matter with strongly lensed gravitational waves*, *Mon. Not. Roy. Astron. Soc.* **502** (2021) L16 [arXiv:2012.12462] [INSPIRE].

[57] Y. Li, X. Fan and L. Gou, *Constraining Cosmological Parameters in the FLRW Metric with Lensed GW+EM Signals*, *Astrophys. J.* **873** (2019) 37 [arXiv:1901.10638] [INSPIRE].

[58] M. Oguri and R. Takahashi, *Probing Dark Low-mass Halos and Primordial Black Holes with Frequency-dependent Gravitational Lensing Dispersions of Gravitational Waves*, *Astrophys. J.* **901** (2020) 58 [arXiv:2007.01936] [INSPIRE].

[59] O.A. Hannuksela, T.E. Collett, M. Çalışkan and T.G.F. Li, *Localizing merging black holes with sub-arcsecond precision using gravitational-wave lensing*, *Mon. Not. Roy. Astron. Soc.* **498** (2020) 3395 [arXiv:2004.13811] [INSPIRE].

[60] K. Liao, S. Tian and X. Ding, *Probing compact dark matter with gravitational wave fringes detected by the Einstein Telescope*, *Mon. Not. Roy. Astron. Soc.* **495** (2020) 2002 [arXiv:2001.07891] [INSPIRE].

[61] A. Piórkowska-Kurpas et al., *Inspiraling Double Compact Object Detection and Lensing Rate: Forecast for DECIGO and B-DECIGO*, *Astrophys. J.* **908** (2021) 196 [arXiv:2005.08727] [INSPIRE].

[62] J. Urrutia and V. Vaskonen, *Lensing of gravitational waves as a probe of compact dark matter*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 1358 [arXiv:2109.03213] [INSPIRE].

[63] J.M. Diego, T. Broadhurst and G. Smoot, *Evidence for lensing of gravitational waves from LIGO-Virgo data*, *Phys. Rev. D* **104** (2021) 103529 [arXiv:2106.06545] [INSPIRE].

[64] C. Dalang, G. Cusin and M. Lagos, *Polarization distortions of lensed gravitational waves*, *Phys. Rev. D* **105** (2022) 024005 [arXiv:2104.10119] [INSPIRE].

[65] H.G. Choi, C. Park and S. Jung, *Small-scale shear: Peeling off diffuse subhalos with gravitational waves*, *Phys. Rev. D* **104** (2021) 063001 [arXiv:2103.08618] [INSPIRE].

[66] Z. Gao et al., *A higher probability of detecting lensed supermassive black hole binaries by LISA*, *Mon. Not. Roy. Astron. Soc.* **512** (2022) 1 [arXiv:2102.10295] [INSPIRE].

[67] X. Guo and Y. Lu, *Probing the nature of dark matter via gravitational waves lensed by small dark matter halos*, *Phys. Rev. D* **106** (2022) 023018 [arXiv:2207.00325] [INSPIRE].

[68] P. Cremonese, D.F. Mota and V. Salzano, *Characteristic Features of Gravitational Wave Lensing as Probe of Lens Mass Model*, *Annalen Phys.* **535** (2023) 2300040 [arXiv:2111.01163] [INSPIRE].

[69] C.T. Mpetha, G. Congedo and A. Taylor, *Future prospects on testing extensions to ΛCDM through the weak lensing of gravitational waves*, *Phys. Rev. D* **107** (2023) 103518 [arXiv:2208.05959] [INSPIRE].

[70] M. Çalışkan et al., *Observability of lensing of gravitational waves from massive black hole binaries with LISA*, *Phys. Rev. D* **107** (2023) 043029 [arXiv:2206.02803] [INSPIRE].

[71] H. Zhou, Z. Li, K. Liao and Z. Huang, *Constraints on compact dark matter from lensing of gravitational waves for the third-generation gravitational wave detector*, *Mon. Not. Roy. Astron. Soc.* **518** (2022) 149 [arXiv:2206.13128] [INSPIRE].

JCAP07(2023)007

[72] LIGO Scientific and VIRGO collaborations, *Search for Lensing Signatures in the Gravitational-Wave Observations from the First Half of LIGO-Virgo's Third Observing Run*, *Astrophys. J.* **923** (2021) 14 [arXiv:2105.06384] [INSPIRE].

[73] R. Kormann, P. Schneider and M. Bartelmann, *Isothermal elliptical gravitational lens models*, *Astron. Astrophys.* **284** (1994) 285.

[74] J.M. Diego et al., *Observational signatures of microlensing in gravitational waves at LIGO/Virgo frequencies*, *Astron. Astrophys.* **627** (2019) A130 [arXiv:1903.04513] [INSPIRE].

[75] J.M. Diego, *Constraining the abundance of primordial black holes with gravitational lensing of gravitational waves at LIGO frequencies*, *Phys. Rev. D* **101** (2020) 123512 [arXiv:1911.05736] [INSPIRE].

[76] P. Schneider, J. Ehlers and E.E. Falco, *Gravitational Lenses*, Springer New York, U.S.A. (2012).

[77] Planck collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [*Erratum ibid.* **652** (2021) C4] [arXiv:1807.06209] [INSPIRE].

[78] C.R. Keeton, *A catalog of mass models for gravitational lensing*, astro-ph/0102341 [INSPIRE].

[79] K. Haris et al., *Identifying strongly lensed gravitational wave signals from binary black hole mergers*, arXiv:1807.07062 [INSPIRE].

[80] R. Takahashi, *Wave Effects in the Gravitational Lensing of Gravitational Waves from Chirping Binaries*, Ph.D. thesis, Department of Physics, Kyoto University, Japan (2004).

[81] J. Wang et al., *Universal structure of dark matter haloes over a mass range of 20 orders of magnitude*, *Nature* **585** (2020) 39 [arXiv:1911.09720] [INSPIRE].

[82] E. Retana-Montenegro et al., *Analytical properties of Einasto dark matter haloes*, *Astron. Astrophys.* **540** (2012) A70 [arXiv:1202.5242] [INSPIRE].

[83] M. Lewicki and V. Vaskonen, *Impact of LIGO-Virgo black hole binaries on gravitational wave background searches*, *Eur. Phys. J. C* **83** (2023) 168 [arXiv:2111.05847] [INSPIRE].

[84] P. Jaranowski and A. Krolak, *Gravitational-Wave Data Analysis. Formalism and Sample Applications: The Gaussian Case*, *Living Rev. Rel.* **8** (2005) 3 [arXiv:0711.1115] [INSPIRE].

[85] S. May and V. Springel, *Structure formation in large-volume cosmological simulations of fuzzy dark matter: impact of the non-linear dynamics*, *Mon. Not. Roy. Astron. Soc.* **506** (2021) 2603 [arXiv:2101.01828] [INSPIRE].

[86] H.Y.J. Chan et al., *The diversity of core-halo structure in the fuzzy dark matter model*, *Mon. Not. Roy. Astron. Soc.* **511** (2022) 943 [arXiv:2110.11882] [INSPIRE].

[87] D. Gerosa et al., *Multiband gravitational-wave event rates and stellar physics*, *Phys. Rev. D* **99** (2019) 103004 [arXiv:1902.00021] [INSPIRE].

[88] W.H. Press and P. Schechter, *Formation of galaxies and clusters of galaxies by selfsimilar gravitational condensation*, *Astrophys. J.* **187** (1974) 425 [INSPIRE].

[89] J.R. Bond, S. Cole, G. Efstathiou and N. Kaiser, *Excursion set mass functions for hierarchical Gaussian fluctuations*, *Astrophys. J.* **379** (1991) 440 [INSPIRE].

[90] S. Dodelson, *Modern Cosmology*, Academic Press, Amsterdam, The Netherlands (2003) [INSPIRE].

[91] T.J. Whalen et al., *Variable Hard X-ray Emission from the Candidate Accreting Black Hole in Dwarf Galaxy Henize 2–10*, *Astrophys. J.* **806** (2015) 37 [arXiv:1504.03331] [INSPIRE].

[92] D. Lin et al., *A luminous X-ray outburst from an intermediate-mass black hole in an off-centre star cluster*, *Nature Astron.* **2** (2018) 656 [arXiv:1806.05692] [INSPIRE].

JCAP07(2023)007

[93] Z.-K. Peng et al., *CDF-S XT1 and XT2: White Dwarf Tidal Disruption Events by Intermediate-Mass Black Holes?*, *Astrophys. J. Lett.* **884** (2019) L34 [arXiv:1909.12173] [INSPIRE].

[94] R.-F. Shen, *Fast, Ultraluminous X-Ray Bursts from Tidal Stripping of White Dwarfs by Intermediate-mass Black Holes*, *Astrophys. J. Lett.* **871** (2019) L17 [arXiv:1809.09359] [INSPIRE].

[95] M. Mićić, J.A. Irwin and D. Lin, *An Ultraluminous Supersoft Source in a Dwarf Galaxy of A85: An Intermediate-mass Black Hole Candidate*, *Astrophys. J.* **928** (2022) 117 [arXiv:2203.10136] [INSPIRE].

[96] V. Toptun et al., *Confirmation of intermediate-mass black holes candidates with X-ray observations*, in the proceedings of the *All-Russian Astronomical Conference 2021*, Moscow, Russian Federation, 23–28 August 2021, [DOI:10.51194/VAK2021.2022.1.1.117] [arXiv:2201.01075] [INSPIRE].

[97] LIGO Scientific and Virgo collaborations, *GW190521: A Binary Black Hole Merger with a Total Mass of* $150M_{\odot}$, *Phys. Rev. Lett.* **125** (2020) 101102 [arXiv:2009.01075] [INSPIRE].

[98] LIGO Scientific and Virgo collaborations, *Search for intermediate mass black hole binaries in the first and second observing runs of the Advanced LIGO and Virgo network*, *Phys. Rev. D* **100** (2019) 064064 [arXiv:1906.08000] [INSPIRE].

[99] A. Rasskazov, G. Fragione and B. Kocsis, *Binary intermediate-mass black hole mergers in globular clusters*, *Astrophys. J.* **899** (2020) 149 [arXiv:1912.07681] [INSPIRE].

[100] G. Fragione and B. Kocsis, *Black hole mergers from an evolving population of globular clusters*, *Phys. Rev. Lett.* **121** (2018) 161103 [arXiv:1806.02351] [INSPIRE].

[101] G. Fragione, I. Ginsburg and B. Kocsis, *Gravitational Waves and Intermediate-mass Black Hole Retention in Globular Clusters*, *Astrophys. J.* **856** (2018) 92 [arXiv:1711.00483] [INSPIRE].

[102] A. Sesana, J. Gair, E. Berti and M. Volonteri, *Reconstructing the massive black hole cosmic history through gravitational waves*, *Phys. Rev. D* **83** (2011) 044036 [arXiv:1011.5893] [INSPIRE].

[103] G. Fragione, A. Loeb, B. Kocsis and F.A. Rasio, *Merger Rates of Intermediate-mass Black Hole Binaries in Nuclear Star Clusters*, *Astrophys. J.* **933** (2022) 170 [arXiv:2204.03745] [INSPIRE].

[104] A.L. Erickcek, M. Kamionkowski and A.J. Benson, *Supermassive Black Hole Merger Rates: Uncertainties from Halo Merger Theory*, *Mon. Not. Roy. Astron. Soc.* **371** (2006) 1992 [astro-ph/0604281] [INSPIRE].

[105] E. Barausse, *The evolution of massive black holes and their spins in their galactic hosts*, *Mon. Not. Roy. Astron. Soc.* **423** (2012) 2533 [arXiv:1201.5888] [INSPIRE].

[106] R. Valiante et al., *Unveiling early black hole growth with multifrequency gravitational wave observations*, *Mon. Not. Roy. Astron. Soc.* **500** (2020) 4095 [arXiv:2010.15096] [INSPIRE].

[107] M. Kulkarni and J.P. Ostriker, *What is the halo mass function in a fuzzy dark matter cosmology?*, *Mon. Not. Roy. Astron. Soc.* **510** (2021) 1425 [arXiv:2011.02116] [INSPIRE].

[108] K. Belczynski, D.E. Holz, T. Bulik and R. O'Shaughnessy, *The first gravitational-wave source from the isolated evolution of two 40–100 Msun stars*, *Nature* **534** (2016) 512 [arXiv:1602.04531] [INSPIRE].

JCAP07(2023)007

Appendix 2

II
J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational wave microlensing by dressed primordial black holes. *Phys. Rev. D*, 108(2):023507, 2023

Gravitational wave microlensing by dressed primordial black holes

Juan Urrutia,[1,2,*] Ville Vaskonen,[1,3,4,†] and Hardi Veermäe[1,‡]

[1]*Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia*
[2]*Departament of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia*
[3]*Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy*
[4]*Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 11 April 2023; accepted 23 June 2023; published 11 July 2023)

We study gravitational wave microlensing by primordial black holes (PBHs), accounting for the effect of a particle dark matter minihalo surrounding them. Such minihaloes are expected when PBHs make up only a fraction of all dark matter. We find that current LIGO-Virgo detections disfavor dark matter in the form of PBHs heavier than $100M_\odot$ at 1.4σ. The next generation observatories can potentially probe PBHs as light as $0.01M_\odot$ and down to 2×10^{-4} fraction of all dark matter. We also show that these detectors can distinguish between dressed and naked PBHs, providing a novel way to study the distribution of particle dark matter around black holes and potentially shed light on the origins of black holes.

DOI: 10.1103/PhysRevD.108.023507

I. INTRODUCTION

Primordial black holes (PBHs) as a potential dark matter (DM) candidate, have gained renewed interest due to their testability through gravitational wave (GW) observations [1,2]. Given the existing constraints on PBH [3], they may comprise all of DM only in the asteroid mass window $10^{-16}M_\odot \lesssim m_{\rm PBH} \lesssim 10^{-11}M_\odot$. Yet, heavier PBHs may be related to the seeding of cosmic structures [4–6] including the high redshift surprisingly luminous galaxies observed by the James Webb telescope [7–9].

After the first detections of black hole (BH) binaries by LIGO [10], speculations of their possible primordial origin were presented [11–13]. The subsequent analyses of the observed binary population [14–16] indicate that many of these BHs are likely to have an astrophysical origin [17–20], while the observed merger rate suggests that stellar mass PBHs cannot account for more than a percent of all DM [17,21–26]. The next-generation GW observatories can probe PBH binary populations across a broad parameter range [27,28].

Gravitational lensing has provided important probes of PBH DM, with $10^{-11}M_\odot \lesssim m_{\rm PBH} \lesssim 30M_\odot$ PBHs constrained by stellar microlensing [29–36] and heavier PBHs by the lensing of type Ia supernovae [37,38] or GWs [39]. At high masses, $m_{\rm PBH} \gtrsim 100M_\odot$, the most stringent constraints arise from the accretion of baryons into PBHs [40–46].

The search for PBHs and other compact astrophysical objects in the stellar mass range can be conducted through GW lensing, with next-generation GW observatories, such as the Einstein Telescope (ET), having the potential to confirm or exclude the primordial origin for the observed BH mergers [39,47–51]. Conventional optical microlensing searches rely on the lens transiting through its Einstein radius and become challenging when the transit time exceeds the duration of the experiment. GW microlensing, on the other hand, relies on the interference of the multiple paths the GW takes around the lens [52], allowing for the detection of much heavier lenses. Although various different DM substructures have been considered as lenses [53–57], the expected rate of such events has been shown to be low [58].

In this paper, we constrain the abundance of PBHs using GW microlensing and the LIGO-Virgo observations and examine the capabilities of current and future GW detectors, including LIGO O5 (A+) [59] and ET [60]. We include the effect of a particle dark matter (PDM) minihalo surrounding the PBHs, which is expected to form when PBHs make up only a fraction of DM. These PDM minihaloes have been extensively studied as they are relevant for precise predictions of lensing surveys, WIMP annihilation signals, and PBH accretion [35,36,40,46,61–69].

Our work incorporates several key improvements when compared to existing literature. On top of considering the well-motivated effect of the DM dress around the PBHs in the context of GW lensing, we do so in full detail without relying on approximations of the lens profile. We use GW waveform templates that include the inspiral, merger, and ringdown phases, and employ a log-likelihood analysis to accurately assess the detectability of lensing. These considerations

*juan.urrutia@kbfi.ee
†ville.vaskonen@pd.infn.it
‡hardi.veermae@cern.ch

 © 2023 American Physical Society

allow us to place reliable constraints on the PBH abundance from existing LIGO-Virgo data and to derive prospects for the upcoming LIGO observational runs and ET. Additionally, we discuss the ability of the next-generation GW detectors to differentiate between BH environments. We show that ET will be able to identify the presence of DM minihalos around BHs. This would provide a new way to gain insight into the origins of binary black hole populations that is independent of the merger rate and could also shed light on the nature of DM.

II. DM DRESS OF PBHs

For GW microlensing, the details of the inner profile slope are not crucial, and we can adapt the density profile [64,68],[1]

$$\rho_h(r) = \rho_{\rm PDM,eq}\left(\frac{r}{r_{\rm eq}}\right)^{-\frac{9}{4}}\theta(R_h - r) \approx \frac{0.4M_\odot}{\rm pc^3}(1-f_{\rm PBH})\left[\frac{m_{\rm PBH}}{M_\odot}\right]^{\frac{3}{4}}\left[\frac{r}{\rm pc}\right]^{-\frac{9}{4}}\theta(R_h - r), \quad (1)$$

where $\rho_{\rm PDM,eq} = (1-f_{\rm PBH})\rho_{\rm eq}\Omega_{\rm DM}/\Omega_{\rm M}$ is the average proper cold particle DM energy density at matter-radiation equality, $f_{\rm PBH}$ is the fraction of DM in the form of PBHs, $r_{\rm eq} = 0.4(m_{\rm PBH}\rho_{\rm eq})^{1/3}$ is the turnaround radius at matter-radiation equality, and R_h denotes the proper size of the minihalo. As can be seen in Fig. 1, this profile is in good agreement with the results of N-body simulations of isolated PBHs [46,65], but it predicts a less dense and thus, a less massive PDM minihalo than was estimated by approximate analytic arguments [40,62,63].

To determine R_h, we first note that the PDM minihalo mass M_h is limited by the available amount of PDM in the Universe. This implies that $f_{\rm PBH}M_h \le (1-f_{\rm PBH})m_{\rm PBH}$. This bound is saturated in the idealized case when all of the PDM is contained in PBH minihaloes. A stronger condition for R_h, however, arises from the formation of DM structures—as the minihaloes grow via secondary accretion in the matter-dominated phase [61], one expects that the minihalo growth is halted once the PBHs are absorbed into nonlinear DM structures.[2] If $f_{\rm PBH}$ is large, the PBHs themselves can begin to form small clusters of a few PBHs at the onset of matter domination [22,25,45,65]. We estimate that this effect stops the minihalo growth when half of the PBHs have formed structures containing two or more PBHs. This happens at $1+z_{\rm NL} \approx 5500 f_{\rm PBH}$ [65]. Additionally, we expect that the usual CDM cosmology

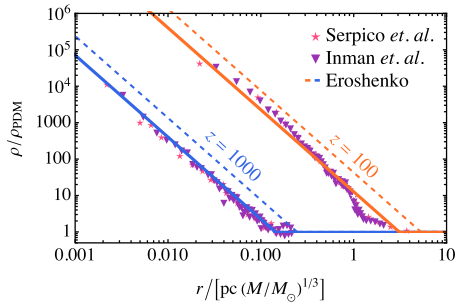


FIG. 1. The PDM minihalo profile (1) (solid lines) at redshifts $z = 1000$ and $z = 100$ compared to the estimate in Ref. [63] (dashed lines) and profiles of isolated PBHs obtained via N-body simulations [46,65] (points).

holds at larger scales so that nonlinear structure formation becomes relevant not later than at redshift 30, which sets the lower bound on $z_{\rm NL}$. Defining the size of the PDM minihalo by the radius at which the density of the minihalo matches the ambient PDM density, we find that $\rho_h(R_h) \approx \rho_{\rm PDM}(z_{\rm NL})$, which gives

$$R_h \approx 14\ {\rm pc}\left[\frac{m_{\rm PBH}}{M_\odot}\right]^{\frac{1}{3}}\min(1, 10^{-3}f_{\rm PBH}^{-\frac{4}{3}}). \quad (2)$$

Consequently, the minihalo mass is

$$M_h \approx (1-f_{\rm PBH})m_{\rm PBH}\min(50, 0.28 f_{\rm PBH}^{-1}). \quad (3)$$

Therefore, the PDM minihalo can be at most 50 times heavier than the central BH, and at most 28% of PDM can be bound to PBHs.

The minihalo profile is not well understood for large $f_{\rm PBH}$ and at small z. N-body simulations [65] indicate that the minihalo around isolated PBHs becomes steeper at larger $f_{\rm PBH}$, potentially due to the tidal effects that the neighboring PBHs have on the outer shells of the PDM minihalo. On the other hand, clusters of several PBHs possess much shallower PDM minihaloes. In this paper, we consider the region of the density profile that was formed when most PBHs were isolated. By the above arguments, most PBHs were isolated at $z \gtrsim 100$ when $f_{\rm PBH} \gtrsim 0.02$, that is, at the epoch when the profile (1) has been numerically tested. Moreover, if $f_{\rm PBH} \gtrsim 0.2$, the PBHs are heavier than the minihalos surrounding them, so the GW microlensing is only mildly affected by the latter. In summary, we expect that a more detailed description of the PBH dressing would not strongly affect our results.

III. MICROLENSING FORMALISM

We consider a compact object binary emitting GWs at the angular diameter distance D_s, which, on their way to the

[1]Geometric units $c = G = 1$ are used throughout this paper.
[2]Virial velocities can be quite large even in small early PBHs clusters and stifle the binding of PDM to PBHs. In general, minihalo formation is poorly understood in this regime. Thus, to be conservative, we make the assumption that the minihalo growth stops at the onset of nonlinear structure formation.

detector, pass by a PBH surrounded by a DM minihalo at the angular diameter distance D_l. The microlensing effect on the Fourier transform of the GW signal $\tilde{\phi}(f)$ can be characterized by the amplification factor $F(f)$ so that the Fourier transform of the lensed GW signal is $\tilde{\phi}_L(f) = F(f)\tilde{\phi}(f)$. In the thin-lens and point source[3] approximation [71],

$$F(w,\boldsymbol{y}) = \frac{w}{2i\pi}\int \mathrm{d}^2x e^{iwT(\boldsymbol{x},\boldsymbol{y})}, \tag{4}$$

where the integral is over the lens plane and $\boldsymbol{y}$ is a vector in the lens plane that determines the displacement of the source from the line of sight to the lens. The dimensionless frequency w and the dimensionless time delay function T are

$$\begin{aligned} w &= 2\pi f \frac{(1+z_l)D_s}{D_l D_{ls}}\xi_0^2, \\ T(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{2}|\boldsymbol{x}-\boldsymbol{y}|^2 - \psi(\boldsymbol{x}) - \phi_m(\boldsymbol{y}), \end{aligned} \tag{5}$$

where $D_{ls} = D_s - D_l(1+z_l)/(1+z_s)$, ξ_0 is a characteristic lens scale, $\psi(\boldsymbol{x})$ denotes the deflection potential and the function $\phi_m(\boldsymbol{y})$ is defined so that the minimum of $T(\boldsymbol{x},\boldsymbol{y})$ for a fixed $\boldsymbol{y}$ is zero. The lens potential is determined by the density profile of the lens (see, e.g., [72]).

We choose the characteristic lens scale ξ_0 to be the Einstein radius of the naked PBH,

$$\xi_0 = \sqrt{\frac{4m_{\mathrm{PBH}}D_l D_{ls}}{D_s}} \approx 0.014\ \mathrm{pc}\sqrt{\frac{m_{\mathrm{PBH}}}{M_\odot}\frac{D_s}{\mathrm{Gpc}}\frac{D_l D_{ls}}{D_s^2}}, \tag{6}$$

which generally is smaller than the PDM minihalo radius (2). The deflection potential induced by the PBH dressed with the DM minihalo with a density profile of Eq. (1) is $\psi(x) = \ln(x) + \psi_h(x)$, where the first term corresponds to the contribution from the PBH, which we describe as a point mass, and the second term is the contribution from the PDM minihalo around the PBH,

$$\begin{aligned} \psi_h(x) \approx{}& 12(1-f_{\mathrm{PBH}})\left[\frac{m_{\mathrm{PBH}}}{M_\odot}\right]^{\frac{1}{4}}\left[\frac{R_h}{\mathrm{pc}}\right]^{\frac{3}{4}} \\ &\times \begin{cases} \left[\frac{x\xi_0}{R_h}\right]^{\frac{3}{4}}, & x\xi_0 \le R_h \\ 1+\frac{3}{4}\ln\left[\frac{x\xi_0}{R_h}\right], & x\xi_0 > R_h \end{cases}. \end{aligned} \tag{7}$$

For the numerical computation of F, we use the method described in [73], and for $wy > 5$, we switch to the geometric limit (see, e.g., [52]).

[3]As shown in [70], the point source approximation is very good in the parameter range considered in this work.

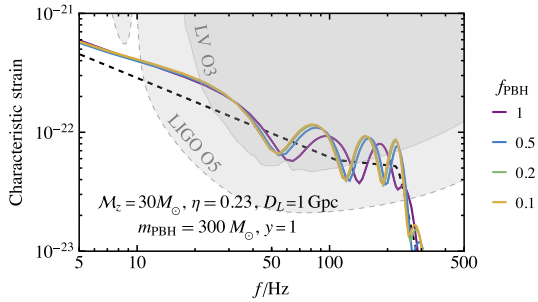


FIG. 2. The dimensionless characteristic GW strain, $\sqrt{f}|\tilde{\phi}(f)|$, from a BH binary merger. The black dashed curve shows the unlensed signal and the solid curves show the effect of a $300M_\odot$ PBH lens dressed in a PDM minihalo for different fractions of PBHs as DM. The shaded regions show the LIGO-Virgo O3 [16] and LIGO O5 (A+) [59] sensitivities.

In Fig. 2, we show examples of the lensed signal for various values of f_{PBH}. The variations in the interference effect that modulates the amplitude of the signal are clearly visible when compared to the unlensed signal shown by the dashed curve. Moreover, we see that this effect depends on the fraction of DM in PBHs, as it, together with the PBH mass, determines the size of the PDM minihalo around the PBH.

IV. ANALYSIS

A. Detectability of the lens

The Fourier transform of the GW signal can be expressed as $\tilde{\phi}(f) = A(f;\boldsymbol{\theta})e^{-i\Psi(f;\boldsymbol{\theta})}$. We compute the amplitude $A(f;\boldsymbol{\theta})$ of the unlensed signal using the inspiral-merger-ringdown template [74],

$$\begin{aligned} A(f;\boldsymbol{\theta}) ={}& \sqrt{\frac{5}{24}}\frac{\mathcal{M}_z^{\frac{5}{6}}}{\pi^{\frac{2}{3}}D_L} \\ &\times \begin{cases} f^{-\frac{7}{6}} & f < f_{\mathrm{merg}} \\ f_{\mathrm{merg}}^{-\frac{1}{2}} f^{-\frac{2}{3}} & f_{\mathrm{merg}} \le f < f_{\mathrm{ring}} \\ f_{\mathrm{merg}}^{-\frac{1}{2}} f_{\mathrm{ring}}^{-\frac{2}{3}} \frac{\sigma^2}{4(f-f_{\mathrm{ring}})^2+\sigma^2} & f_{\mathrm{ring}} \le f < f_{\mathrm{cut}} \end{cases} \end{aligned} \tag{8}$$

and the phase as

$$\Psi(f;\boldsymbol{\theta}) = 2\pi f t_c - \phi_c + \frac{3}{128}(\pi\mathcal{M}_z f)^{-\frac{5}{3}}. \tag{9}$$

Here, $\mathcal{M}_z$ denotes the redshifted chirp mass of the binary, $\mathcal{M}_z \equiv (1+z)(m_1 m_2)^{\frac{3}{5}}/(m_1+m_2)^{\frac{1}{5}}$, and D_L its luminosity distance. The frequencies f_{merg}, f_{ring}, f_{cut}, and σ are parameterized as $f_j = \eta^{\frac{3}{5}}(a_j\eta^2 + b_j\eta + c_j)/(\pi\mathcal{M}_z)$, where η denotes the symmetric mass ratio of the binary,

$\eta \equiv m_1 m_2/(m_1+m_2)^2$, and a_j, b_j, and c_j are coefficients whose fitted values are given in Table I of [74]. This unlensed GW template includes five parameters: $\boldsymbol{\theta} = \{\mathcal{M}_z, \eta, D_L, t_c, \phi_c\}$.

We determine the detectability of the lens by computing the log-likelihood difference (see, e.g., [58]),

$$\Delta\chi^2(\boldsymbol{\theta}_s) = 4\bar{\omega} \min_{\boldsymbol{\theta}} \int \mathrm{d}f \frac{|\tilde{\phi}_L(f;\boldsymbol{\theta}_s) - \tilde{\phi}_T(f;\boldsymbol{\theta})|^2}{S_\mathrm{n}(f)}, \tag{10}$$

where $\tilde{\phi}_L(f;\boldsymbol{\theta}_s)$ is the lensed GW signal, $\tilde{\phi}_T(f;\boldsymbol{\theta})$ the unlensed GW template, the minimum is taken over the parameters of the unlensed GW template, and $S_n(f)$ is the noise power spectral density, including the intrinsic detector noise and the noise from undetectable compact object binaries [75]. The prefactor $\bar{\omega}$ accounts for the sky location and the inclination of the binary, and the polarization of the signal. Averaging leads to $\bar{\omega} = 4/25$ and $\bar{\omega} = 3/25$ for L- and triangular-shaped detectors.

We compute $\Delta\chi^2(\boldsymbol{\theta}_s)$ numerically and find the maximal impact parameter $y_{\max}(\boldsymbol{\theta}_s)$ so that $\Delta\chi^2(\boldsymbol{\theta}_s) > 11.31$ for $y < y_{\max}(\boldsymbol{\theta}_s)$. This implies that for $y < y_{\max}(\boldsymbol{\theta}_s)$ the lensed template fits the signal better than the unlensed template at the 2σ confidence level.

We can further use the log-likelihood test (10) to analyze whether different lens profiles can be experimentally resolved. To do this, we consider the lensed waveform characterized by the parameters $\boldsymbol{\theta} = \{\mathcal{M}_z, \eta, D_L, t_c, \phi_c, y, m_\mathrm{PBH}, f_\mathrm{PBH}\}$ as the template $\tilde{\phi}_T(f;\boldsymbol{\theta})$. For this eight-parameter fit, the 2σ confidence level corresponds to $\Delta\chi^2(\boldsymbol{\theta}_s) = 15.79$.

B. Lensing probability

The computation of $y_{\max}(\boldsymbol{\theta}_s)$ is of crucial importance since it determines the lensing cross section $\sigma(\boldsymbol{\theta}_s) = \pi\xi_0^2 y_{\max}(\boldsymbol{\theta}_s)^2$. The probability of a GW signal to be lensed follows from the probability of a lens object lying within the cross section σ from the line-of-sight to the GW source. It is given by $P_l(\boldsymbol{\theta}_s) = 1 - e^{-\tau(\boldsymbol{\theta}_s)}$ where

$$\tau(\boldsymbol{\theta}_s) = \int_0^{z_s} \mathrm{d}z_l \frac{\sigma(\boldsymbol{\theta}_s) n(z_l)}{(1+z_l)H(z_l)} \tag{11}$$

is the optical depth. The number density of the lens objects is $n(z_l) = (1+z_l)^3 f_\mathrm{PBH}\rho_\mathrm{DM}/m_\mathrm{PBH}$.

We note that in (11) we have assumed that PBHs were uniformly distributed in space. However, they would have inevitably clustered during the structure formation [22,25,45,65]. Studies of the effect of PBH clustering on optical microlensing [76,77] suggest the effect is negligible on the resulting constraints, except in cases where PBHs were initially formed in very compact clusters. The latter case would, however, clash with Lyman-α observations [78]. In short, our results remain robust also in the presence of observational allowed relatively mild PBH clustering.

The expected number of lensed events from existing observations is given by

$$\bar{N}_l(m_\mathrm{PBH}, f_\mathrm{PBH}) = \sum_{j=1}^{N} P_l(\boldsymbol{\theta}_j), \tag{12}$$

where N is the number of observed events. The prospects for future observations, on the other hand, can be obtained by integrating the expected number of lensed events from a model of the merger rate,

$$\bar{N}_l(m_\mathrm{PBH}, f_\mathrm{PBH}) = \mathcal{T} \int \mathrm{d}\lambda P_l(\boldsymbol{\theta}_s) p_\mathrm{det}\left(\frac{\mathrm{SNR}_c}{\mathrm{SNR}(\boldsymbol{\theta}_s)}\right), \tag{13}$$

where $\mathcal{T}$ is the observation time, p_det is the detection probability [79], SNR is the signal-to-noise ratio [80] and

$$\mathrm{d}\lambda = \frac{1}{1+z_s}\frac{\mathrm{d}R}{\mathrm{d}m_1\mathrm{d}m_2}\frac{\mathrm{d}V_c}{\mathrm{d}z_s}\mathrm{d}m_1\mathrm{d}m_2\mathrm{d}z_s \tag{14}$$

is the differential rate of compact binary mergers. We use $\mathrm{SNR}_c = 8$ as the threshold SNR. The SNR is computed for an optimally oriented source-detector system and the averaging over the binary sky location and inclination and the polarization of the signal is accounted for by the detection probability p_det.

We parametrize the merger rate of stellar mass BH binaries as

$$\frac{\mathrm{d}R}{\mathrm{d}m_1\mathrm{d}m_2} = \frac{R_0}{Z_\psi}\eta^\beta \psi(m_1)\psi(m_2)\mathrm{SFR}(z), \tag{15}$$

assuming redshift dependence proportional to the star formation rate [81]. At $z < 2$, this behaves roughly as $\propto (1+z)^{2.7}$, which is in good agreement with the LIGO-Virgo results [82]. The factor Z_ψ is chosen such that $R(z=0) = R_0$. We consider a cut double power-law shape for the BH mass function,

$$\psi(m) \propto \begin{cases} (m/m_2)^{\gamma_1}, & m_1 \le m < m_2 \\ (m/m_2)^{\gamma_2}, & m \ge m_2, \end{cases} \tag{16}$$

and fix $m_1 = 3M_\odot$ and $m_2 = 55M_\odot$, corresponding to the lower and higher mass gaps. We fit the rest of the parameters, R_0, β, γ_1, and γ_2, to the LIGO-Virgo BH binary merger data [14–16] following Ref. [17]. In the following, we use the best fit values $R_0 = 15\ \mathrm{yr}^{-1}\ \mathrm{Gpc}^{-3}$, $\beta = 4.8$, $\gamma_1 = -1.8$, and $\gamma_2 = -6.0$.

When assessing prospective sensitivities, we assume that the astrophysically motivated model (15) can describe the GW sources. However, if the abundance of stellar mass

PBHs is sufficiently large, $f_{\rm PBH} \gtrsim 10^{-4}$, detectable PBH merger events are expected [17,19]. As we fix the magnitude of the merger rate from existing observations, a partially primordial origin of the GW sources would mainly imply that the merger rate grows monotonously with redshift, unlike the star formation rate. We will, however, not consider this possibility to ensure conservative constraints that are independent of the PBH merger rate. We stress also that the uncertainties related to the merger rate model do not apply to our analysis of the observed LIGO-Virgo events.

V. RESULTS AND DISCUSSION

During the first three observing runs, the LIGO-Virgo network detected a total of 90 compact binary coalescence events. We take source parameters and the corresponding sensitivity curves of the detectors from the LIGO-Virgo catalogs [14–16]. For each event, we scale the sensitivity $S_n(f)$ so that the computed SNR matches the mean network SNR value reported in the catalogs. Then, using the scaled $S_n(f)$ and taking $\bar{\omega} = 1$ in Eq. (10), we compute the lensing probability for each of these events as a function of $m_{\rm PBH}$ and $f_{\rm PBH}$, including the DM minihalos around the PBHs. This procedure accounts for the sky location, polarization and inclination of the binary, and uncertainties in our estimate of $S_n(f)$. When summing over the LIGO-Virgo events, we find that the expected total number of lensed events is $\bar{N}_l \approx 1.75$ if $m_{\rm PBH} > 100 M_\odot$ and $f_{\rm PBH} = 1$, and $\bar{N}_l \approx 0.5$ if $f_{\rm PBH} = 0.2$ or if $f_{\rm PBH} = 1$ and $m_{\rm PBH} \approx 20 M_\odot$. The LIGO-Virgo Collaboration has searched for lensing signatures in its catalog with negative results [83]. So, the expected number of lensed events implies a mild 1σ tension with $f_{\rm PBH} \geq 0.7$ for $m_{\rm PBH} > 100 M_\odot$ for dressed PBHs and $f_{\rm PBH} \geq 0.8$ for naked PBHs and $f_{\rm PBH} < 1$ for $m_{\rm PBH} > 50 M_\odot$ in both scenarios, as shown by the orange curves in Fig. 3. Our constraint is weaker than estimated in Ref. [39], but can reach lower masses. This is because the analysis in [39] was based on the accuracy at which the lensed waveform parameters can be measured, but it did not consider whether a lensed template gives a significantly better fit than an unlensed one.

We estimate the sensitivity of the LIGO O5 and ET to the PBH abundance by using the BH binary merger rate model (15).[4] The prospects are shown by the green and blue curves in Fig. 3.[5] We find that LIGO O5 can probe $f_{\rm PBH} \geq 0.1$ for masses $m_{\rm PBH} > 10 M_\odot$, and ET can reach abundances $f_{\rm PBH} = 10^{-3}$ and subsolar mass PBHs with masses $m_{\rm PBH} \gtrsim 0.01 M_\odot$. Importantly, abundances of the order

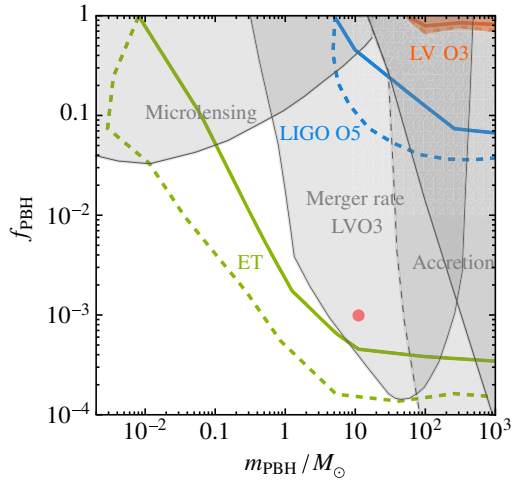


FIG. 3. The blue and green curves show the projected sensitivities of LIGO O5 and ET, and the orange curve shows the region disfavored at 1σ by the current (O3) LIGO-Virgo observations. The gray curves show PBH constraints from microlensing [30], merger rate measurements [17], and accretion [46]. The solid and dashed curves show the results with and without the effect of the PDM minihalo. The red dot corresponds to the scenario in which all observed BH mergers would be primordial [17–19].

$f_{\rm PBH} = 10^{-3}$ correspond to the scenario in which a significant fraction of the observed BH-BH merger events were primordial [17–19]. So, GW lensing with ET can provide a robust independent probe for PBH scenarios of the observed BH binary mergers.

As shown in Fig. 3, the PDM dress enhances the sensitivity to $f_{\rm PBH}$ by a factor of $\mathcal{O}(1$–$10)$ and the effect is slightly stronger at higher $f_{\rm PBH}$.[6] Notably, the enhancement in the sensitivity at $f_{\rm PBH} \leq 10^{-2}$ is smaller than one would naively expect from the PDM minihalo mass (3) because PDM minihalo is much larger than the typical Einstein radius for low $f_{\rm PBH}$. We remark that our constraints can also be extended to sufficiently compact extended objects, like DM stars, as was discussed in [58].

Finally, we investigate whether a heavier point mass can mimic a dressed PBH. In previous studies, the dressed PBH was modeled as a point mass lens whose mass was equal to the integrated mass up to the Einstein radius [35]. If this approximation fails, analyzing a lensed event could potentially reveal information about the environment surrounding the BH. We find that LIGO O5 is unlikely to see events

[4]As a cross-check, we derived the LIGO-Virgo O3 constraint using the merger rate (15) instead of the observed individual events and found a very good agreement between the results.

[5]Figure 3 is derived assuming a monochromatic PBH mass function. The results for extended PBH mass functions can be computed using the method described in Ref. [84].

[6]We find that the PDM minihalo has a much weaker effect than was suggested in the recent strong lensing study [35]. Apart from differences in the simplifications, this discrepancy is mostly because a much denser PDM minihalo was used in [35].

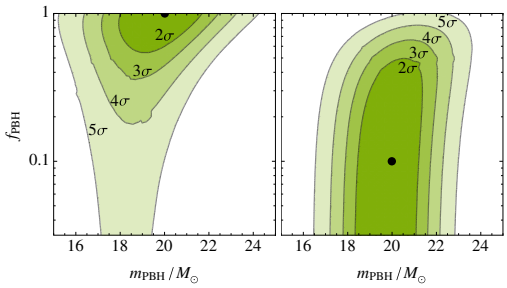


FIG. 4. Sensitivity of ET to the lens profile for two benchmark cases with $\mathcal{M}_z = 40M_\odot$, $\eta = 0.23$, $D_L = 5$ Gpc, $m_{\rm PBH} = 20M_\odot$, and $y = 5$. The GW signal is assumed to correspond to a naked PBH lens ($f_{\rm PBH} = 1$) in the left panel and to a dressed lens (with $f_{\rm PBH} = 0.1$) in the right panel. The shading shows parameter regions in which ET can distinguish the assumed waveform (black dot) from a lensed signal characterized by $(m_{\rm PBH}, f_{\rm PBH})$ at the $2\sigma - 5\sigma$ confidence levels.

for which dressed and naked lenses can be distinguished. On the other hand, ET is expected to see several sufficiently high SNR events allowing it to resolve the lens profile. Two realistic benchmark cases where ET would be able to measure the PDM dress of a PBH lens are shown in Fig. 4, illustrating that the presence or absence of the dress may be excluded at the 4σ confidence level. Such events can shed light on the origin of BHs in the Universe and, if the lens is interpreted as a PBH, would give a probe of $f_{\rm PBH}$ through the measurement of the PDM minihalo properties. In this way, GW lensing can be a powerful tool in constraining potential PBH scenarios, especially when combined with other observables such as the PBH binary merger rate.

VI. CONCLUSIONS

We have examined gravitational wave microlensing by primordial black holes, accounting for the effect of the particle dark matter minihalos when primordial black holes make up only a fraction of dark matter. We have further improved the existing analyses by including the full lens profile, employing a more complete waveform that contains the inspiral, merger, and ringdown phases, and performing a matched filtering analysis of the detectability of the lens.

We find that the existing LIGO-Virgo observations disfavor $f_{\rm PBH} > 0.7$ at 1σ in the mass range $m_{\rm PBH} \gtrsim 50M_\odot$. Moreover, we have shown that future observing runs of LIGO will be sensitive to $f_{\rm PBH} \gtrsim 0.1$ and $m_{\rm PBH} > 5M_\odot$ and next-generation gravitational wave observatories like ET will be sensitive $f_{\rm PBH} \gtrsim 2 \times 10^{-4}$ and $m_{\rm PBH} \gtrsim 0.01M_\odot$. In particular, they have the potential to rule out or confirm the primordial origin of the observed black hole merger events independently of the black hole binary merger rate measurements. We have also shown that ET can differentiate between dressed and naked black holes through gravitational wave lensing, providing an independent probe of the origins of black holes in the Universe and the particle dark matter properties, in case a lensed gravitational wave event would be detected.

Our results are expected to remain valid even if the standard minihalo profile and primordial black hole cosmology are modified in realistic ways (e.g., mild initial clustering). In summary, our study offers new insights into GW lensing by primordial black holes and opens up new avenues for exploring the phenomenology of dark matter and primordial black holes.

ACKNOWLEDGMENTS

We thank Gert Hütsi for useful discussions. This work was supported by European Regional Development Fund through the CoE Program Grant No. TK133 and by the Estonian Research Council Grants No. PRG803 and No. PSG86. The work of V. V. has been partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101065736.

[1] T. Nakamura, M. Sasaki, T. Tanaka, and K. S. Thorne, Astrophys. J. Lett. **487**, L139 (1997).
[2] Y. N. Eroshenko, J. Phys. Conf. Ser. **1051**, 012010 (2018).
[3] B. Carr, K. Kohri, Y. Sendouda, and J. Yokoyama, Rep. Prog. Phys. **84**, 116902 (2021).
[4] K. Freese, R. Price, and D. N. Schramm, Astrophys. J. **275**, 405 (1983).
[5] B. J. Carr and J. Silk, Astrophys. J. **268**, 1 (1983).
[6] B. Carr and J. Silk, Mon. Not. R. Astron. Soc. **478**, 3756 (2018).
[7] B. Liu and V. Bromm, Astrophys. J. Lett. **937**, L30 (2022).
[8] G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe, Phys. Rev. D **107**, 043502 (2023).
[9] G.-W. Yuan, L. Lei, Y.-Z. Wang, B. Wang, Y.-Y. Wang, C. Chen, Z.-Q. Shen, Y.-F. Cai, and Y.-Z. Fan, arXiv:2303.09391.
[10] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[11] M. Sasaki, T. Suyama, T. Tanaka, and S. Yokoyama, Phys. Rev. Lett. **117**, 061101 (2016); **121**, 059901(E) (2018).
[12] S. Bird, I. Cholis, J. B. Muñoz, Y. Ali-Haïmoud, M. Kamionkowski, E. D. Kovetz, A. Raccanelli, and A. G. Riess, Phys. Rev. Lett. **116**, 201301 (2016).

[13] S. Clesse and J. García-Bellido, Phys. Dark Universe **15**, 142 (2017).
[14] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **9**, 031040 (2019).
[15] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **11**, 021053 (2021).
[16] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2111.03606 [Phys. Rev X (to be published)].
[17] G. Hütsi, M. Raidal, V. Vaskonen, and H. Veermäe, J. Cosmol. Astropart. Phys. 03 (**2021**) 068.
[18] A. Hall, A. D. Gow, and C. T. Byrnes, Phys. Rev. D **102**, 123524 (2020).
[19] G. Franciolini, V. Baibhav, V. De Luca, K. K. Y. Ng, K. W. K. Wong, E. Berti, P. Pani, A. Riotto, and S. Vitale, Phys. Rev. D **105**, 083526 (2022).
[20] J. He, H. Deng, Y.-S. Piao, and J. Zhang, arXiv:2303.16810.
[21] M. Raidal, V. Vaskonen, and H. Veermäe, J. Cosmol. Astropart. Phys. 09 (2017) 037.
[22] M. Raidal, C. Spethmann, V. Vaskonen, and H. Veermäe, J. Cosmol. Astropart. Phys. 02 (**2019**) 018.
[23] V. Vaskonen and H. Veermäe, Phys. Rev. D **101**, 043015 (2020).
[24] K. W. K. Wong, G. Franciolini, V. De Luca, V. Baibhav, E. Berti, P. Pani, and A. Riotto, Phys. Rev. D **103**, 023026 (2021).
[25] V. De Luca, V. Desjacques, G. Franciolini, and A. Riotto, J. Cosmol. Astropart. Phys. 11 (**2020**) 028.
[26] G. Franciolini, I. Musco, P. Pani, and A. Urbano, Phys. Rev. D **106**, 123526 (2022).
[27] O. Pujolas, V. Vaskonen, and H. Veermäe, Phys. Rev. D **104**, 083521 (2021).
[28] S. Barsanti, V. De Luca, A. Maselli, and P. Pani, Phys. Rev. Lett. **128**, 111104 (2022).
[29] R. A. Allsman *et al.* (Macho Collaboration), Astrophys. J. Lett. **550**, L169 (2001).
[30] P. Tisserand *et al.* (EROS-2 Collaboration), Astron. Astrophys. **469**, 387 (2007).
[31] K. Griest, A. M. Cieplak, and M. J. Lehner, Astrophys. J. **786**, 158 (2014).
[32] H. Niikura *et al.*, Nat. Astron. **3**, 524 (2019).
[33] N. Smyth, S. Profumo, S. English, T. Jeltema, K. McKinnon, and P. Guhathakurta, Phys. Rev. D **101**, 063005 (2020).
[34] H. Niikura, M. Takada, S. Yokoyama, T. Sumi, and S. Masaki, Phys. Rev. D **99**, 083503 (2019).
[35] M. Oguri, V. Takhistov, and K. Kohri, arXiv:2208.05957.
[36] R.-G. Cai, T. Chen, S.-J. Wang, and X.-Y. Yang, J. Cosmol. Astropart. Phys. 03 (2023) 043.
[37] M. Zumalacarregui and U. Seljak, Phys. Rev. Lett. **121**, 141101 (2018).
[38] S. Dhawan and E. Mörtsell, arXiv:2301.10204.
[39] J. Urrutia and V. Vaskonen, Mon. Not. R. Astron. Soc. **509**, 1358 (2021).
[40] M. Ricotti, J. P. Ostriker, and K. J. Mack, Astrophys. J. **680**, 829 (2008).
[41] B. Horowitz, arXiv:1612.07264.
[42] Y. Ali-Haïmoud and M. Kamionkowski, Phys. Rev. D **95**, 043534 (2017).
[43] V. Poulin, P. D. Serpico, F. Calore, S. Clesse, and K. Kohri, Phys. Rev. D **96**, 083524 (2017).
[44] A. Hektor, G. Hütsi, L. Marzola, M. Raidal, V. Vaskonen, and H. Veermäe, Phys. Rev. D **98**, 023503 (2018).
[45] G. Hütsi, M. Raidal, and H. Veermäe, Phys. Rev. D **100**, 083016 (2019).
[46] P. D. Serpico, V. Poulin, D. Inman, and K. Kohri, Phys. Rev. Res. **2**, 023204 (2020).
[47] S. Jung and C. S. Shin, Phys. Rev. Lett. **122**, 041103 (2019).
[48] J. M. Diego, Phys. Rev. D **101**, 123512 (2020).
[49] K. Liao, S. Tian, and X. Ding, Mon. Not. R. Astron. Soc. **495**, 2002 (2020).
[50] S. Basak, A. Ganguly, K. Haris, S. Kapadia, A. K. Mehta, and P. Ajith, Astrophys. J. **926**, L28 (2022).
[51] H. Zhou, Z. Li, K. Liao, and Z. Huang, Mon. Not. R. Astron. Soc. **518**, 149 (2022).
[52] R. Takahashi and T. Nakamura, Astrophys. J. **595**, 1039 (2003).
[53] J. M. Diego, O. A. Hannuksela, P. L. Kelly, T. Broadhurst, K. Kim, T. G. F. Li, G. F. Smoot, and G. Pagano, Astron. Astrophys. **627**, A130 (2019).
[54] M. Oguri and R. Takahashi, Astrophys. J. **901**, 58 (2020).
[55] H. G. Choi, C. Park, and S. Jung, Phys. Rev. D **104**, 063001 (2021).
[56] X. Guo and Y. Lu, Phys. Rev. D **106**, 023018 (2022).
[57] G. Tambalo, M. Zumalacárregui, L. Dai, and M. H.-Y. Cheung, arXiv:2212.11960.
[58] M. Fairbairn, J. Urrutia, and V. Vaskonen, arXiv:2210.13436.
[59] L. Barsotti, L. McCuller, M. Evans, and P. Fritschel, Report No. LIGO-T1800042-v4, 2018, https://dcc.ligo.org/LIGO-T1800042-v4/public.
[60] S. Hild *et al.*, Classical Quantum Gravity **28**, 094013 (2011).
[61] E. Bertschinger, Astrophys. J. Suppl. Ser. **58**, 39 (1985).
[62] K. J. Mack, J. P. Ostriker, and M. Ricotti, Astrophys. J. **665**, 1277 (2007).
[63] Y. N. Eroshenko, Astron. Lett. **42**, 347 (2016).
[64] J. Adamek, C. T. Byrnes, M. Gosenca, and S. Hotchkiss, Phys. Rev. D **100**, 023506 (2019).
[65] D. Inman and Y. Ali-Haïmoud, Phys. Rev. D **100**, 083528 (2019).
[66] B. Carr, F. Kuhnel, and L. Visinelli, Mon. Not. R. Astron. Soc. **506**, 3648 (2021).
[67] V. De Luca, G. Franciolini, P. Pani, and A. Riotto, J. Cosmol. Astropart. Phys. 04 (**2020**) 052.
[68] M. Boudaud, T. Lacroix, M. Stref, J. Lavalle, and P. Salati, J. Cosmol. Astropart. Phys. 08 (**2021**) 053.
[69] E. U. Ginés, O. Mena, and S. J. Witte, Phys. Rev. D **106**, 063538 (2022).
[70] N. Matsunaga and K. Yamamoto, J. Cosmol. Astropart. Phys. 01 (2006) 023.
[71] P. Schneider, J. Ehlers, and E. Falco, *Gravitational Lenses*, Astronomy and Astrophysics Library (Springer, New York, 2012).
[72] C. R. Keeton, arXiv:astro-ph/0102341.
[73] R. Takahashi, Wave effects in the gravitational lensing of gravitational waves from chirping binaries, Ph.D. thesis, Department of Physics, Kyoto University, 2004.
[74] P. Ajith *et al.*, Phys. Rev. D **77**, 104017 (2008); **79**, 129901(E) (2009).
[75] M. Lewicki and V. Vaskonen, Eur. Phys. J. C **83**, 168 (2023).
[76] M. Petač, J. Lavalle, and K. Jedamzik, Phys. Rev. D **105**, 083520 (2022).

[77] M. Gorton and A. M. Green, J. Cosmol. Astropart. Phys. 08 (2022) 035.
[78] V. De Luca, G. Franciolini, A. Riotto, and H. Veermäe, Phys. Rev. Lett. **129**, 191302 (2022).
[79] D. Gerosa, S. Ma, K. W. K. Wong, E. Berti, R. O'Shaughnessy, Y. Chen, and K. Belczynski, Phys. Rev. D **99**, 103004 (2019).
[80] L. S. Finn, Phys. Rev. D **46**, 5236 (1992).
[81] K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy, Nature (London) **534**, 512 (2016).
[82] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Phys. Rev. X **13**, 011048 (2023).
[83] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. **923**, 14 (2021).
[84] B. Carr, M. Raidal, T. Tenkanen, V. Vaskonen, and H. Veermäe, Phys. Rev. D **96**, 023514 (2017).

Appendix 3

III
J. Urrutia and V. Vaskonen. Dark timbre of gravitational waves. *Phys. Rev. D*, 111(12):123047, 2025

Dark timbre of gravitational waves

Juan Urrutia[1,2,*] and Ville Vaskonen[1,3,4,†]

[1]*Keemilise ja Bioloogilise Füüsika Instituut, Rävala puiestee 10, 10143 Tallinn, Estonia*
[2]*Departament of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia*
[3]*Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy*
[4]*Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 10 May 2024; revised 6 September 2024; accepted 7 May 2025; published 27 June 2025)

Gravitational wave timbre, the relative amplitude and phase of the different frequency harmonics, can change due to interactions with low-mass halos. We focus on binaries in the LISA range and find that the integrated lens effect of cold dark matter structures can be used to probe the existence of $M_{\rm v} \lesssim 10M_\odot$ halos if a single binary with eccentricity $e = 0.3$–0.6 is detected with a signal-to-noise ratio $100 - 10^3$ and it is at $z_s = 0.5$.

DOI: 10.1103/PhysRevD.111.123047

I. INTRODUCTION

Gravitational waves (GWs) are often associated with sound because they are characterized by waveforms, and in general there is a poorer angular resolution than with electromagnetic signals. Like with sound, for GWs there is also the notion of timbre since eccentric binaries emit GWs in different harmonics simultaneously. The relative powers of the harmonics are determined by the binary eccentricity and the harmonic numbers [1].

GWs are sensitive to wave optics interactions with small halos, which can leave detectable imprints on the waveform [2–11] or a phase difference with respect to an electromagnetic counterpart [12–15]. Both methods rely on frequency-dependent effects for detectability. Moreover, in the case of strong lensing, the lensed image can deviate from unlensed waveforms in the presence of higher harmonic modes, precession or eccentricity [16,17].

In this work, we propose a novel probe of the wave optics effects: measurements of the timbre of the GW signal from an eccentric binary. We study if the measurements of the timbre can be used to probe the low-mass end of the dark matter (DM) halo mass function (HMF) where deviations from the cold dark matter (CDM) predictions may appear. For example, the small-scale structures are suppressed in warm or ultralight DM models [18–20].

We focus on signals whose frequency does not change significantly. In the LISA sensitivity range, such signals can originate, e.g., from white dwarf (WD) binaries, intermediate-mass black hole (IMBH) binaries, or extreme mass ratio binaries [21–26]. The signals from WD binaries are weak and only sources in the Milky Way can be seen individually. Such signals can get lensed by Milky Way substructures. As we are interested in lensing by DM halos, the most relevant sources are IMBH binaries that produce the strongest extragalactic signals.

We compute the integrated effect of the DM halo population and show that very light halos, which are integrated out in weak lensing studies as a constant density field [27], induce changes in the amplitude and phases of the different harmonics of a mHz GW signal, effectively changing the timbre. The effect is detectable with LISA in the first harmonics, for example, if a binary is at $z_s = 0.5$ with a signal-to-noise between 100–10^3 and has eccentricity $e = 0.3$–0.6. We find that the effect mainly comes from halos of $\mathcal{O}(10M_\odot)$.

The reason for the dark timbre effect is that lensing by DM halos in wave optics causes a small frequency-dependent enhancement of the GW amplitude that accumulates as the GW encounters several halos over cosmological distances. While the probability of detecting lensing by a single heavier halo is too low [3,11], the integrated effect of small halos is always present if such halos exist. The CDM model predicts a large abundance of light halos that have so far eluded observations [28]. Therefore, the detection of dark timbre would provide a probe of the low-mass tail of the HFM and severely constrain deviations from CDM at small scales.

II. LENSING BY A SINGLE HALO

Consider a binary whose orbital frequency remains almost constant during the observation. We denote the angular diameter distance of the binary by $D_s = d_{\rm c}(z_s)/(1+z_s)$, where z_s is the corresponding redshift and $d_{\rm c}$ is the comoving distance, and assume that the GW

*Contact author: juan.urrutia@kbfi.ee
†Contact author: ville.vaskonen@pd.infn.it

 © 2025 American Physical Society

signal emitted by the binary interacts with a halo at angular diameter distance D_l. In the frequency domain, the lensed waveform is $\tilde{\phi}_L(f) = F(f)\tilde{\phi}(f)$, where $\tilde{\phi}(f)$ denotes the unperturbed GW signal. The amplification factor $F(f)$ under the thin lens, Born, and eikonal approximations, is given by [29,30] (see Appendix A)

$$F(f) = \frac{w(f)}{2i\pi}\int \mathrm{d}^2\boldsymbol{x}\, e^{iw(f)T(\boldsymbol{x},\boldsymbol{y})}. \tag{1}$$

The integral accounts for all the paths that the GW can take through it. The prefactor ensures that $F = 1$ in the absence of lenses. The dimensionless vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are defined as $\boldsymbol{x} \equiv \boldsymbol{\xi}/\xi_0$ and $\boldsymbol{y} \equiv D_l\boldsymbol{\eta}/(D_s\xi_0)$, where ξ_0 denotes a characteristic length scale of the lens, and $\boldsymbol{\eta}$ and $\boldsymbol{\chi}$ are dimensionful vectors that give, respectively, the position of the source in the source plane and the position at which the GW crosses the lens plane. The dimensionless frequency w and the dimensionless time delay function T are given by

$$w(f) \equiv \frac{(1+z_l)D_s}{D_l D_{ls}}\xi_0^2 2\pi f, \tag{2}$$

$$T(\boldsymbol{x},\boldsymbol{y}) \equiv \frac{1}{2}|\boldsymbol{x}-\boldsymbol{y}|^2 - \psi(\boldsymbol{x}) - \phi(\boldsymbol{y}), \tag{3}$$

where $D_{ls} = D_s - (1+z_l)/(1+z_s)D_l$ and $\phi(\boldsymbol{y})$ is defined such that $\min_{\boldsymbol{x}} T(\boldsymbol{x},\boldsymbol{y}) = 0$.

We approximate the halos by the Navarro-Frenk-White (NFW) profile $\rho(r) = r_s^3\rho_s/[r(r_s+r)^2]$, which corresponds to the lens potential (see, e.g., [3])

$$\psi(x) = \kappa\left[\ln^2\left(\frac{x}{2b}\right) - \mathrm{artanh}^2\sqrt{1-\frac{x^2}{b^2}}\right], \tag{4}$$

where $x = |\boldsymbol{x}|$ and $\kappa \equiv 2\pi\rho_s r_s^3/M_{\mathrm{v}}$ and $b \equiv r_s/\xi_0$ are dimensionless parameters. The distances are scaled by $\xi_0 = R_E(M_v)$, where R_E denotes the point mass Einstein radius, so the dimensionless frequency simplifies to $w = 8\pi(1+z_l)M_{\mathrm{v}}f$. The NFW profile is parametrized by the scale radius r_s and the corresponding density ρ_s, which are both functions of the halo virial mass M_{v} through $\int_0^{R_{200}} \mathrm{d}r r^2\rho(r) = 200\rho_c R_{200}^3/3$ where $R_{200} = C(M_{\mathrm{v}})r_s$ and ρ_c denotes the critical density of the Universe. For the concentration parameter $C(M_{\mathrm{v}})$ we use the fit provided in [31], which agrees with cosmological simulations in the relevant range of masses for this study [32]. For halos as light as the ones we will consider, interactions with stars can affect the profile. However, the survival rate is expected to be high [33–39] so we can safely use the NFW profile.

In the geometric limit, the GW follows a single classical trajectory, the minimal time path, around the NFW halo. The interaction lacks interference patterns and induces a frequency-independent amplification [2],

$$|F| = (\det[\partial_a\partial_b T(\boldsymbol{x}_j, y)])^{-2}, \tag{5}$$

where $y = |\boldsymbol{y}|$ and $\boldsymbol{x}_j$ corresponds to the classical path. However, in the wave optics regime, the amplification becomes frequency dependent, and numerical computation of Eq. (1) is necessary.[1]

In the upper panel of Fig. 1, we show the amplification $|F|$ for different halo masses and distances in the lens plane. We have fixed the signal frequency to $f_{\mathrm{GW}} = 2$ mHz, the source redshift to $z_s = 0.5$, and the lens redshift to $z_l = 0.25$. For a fixed halo mass in the wave optics regime, the interaction is independent of the projected position

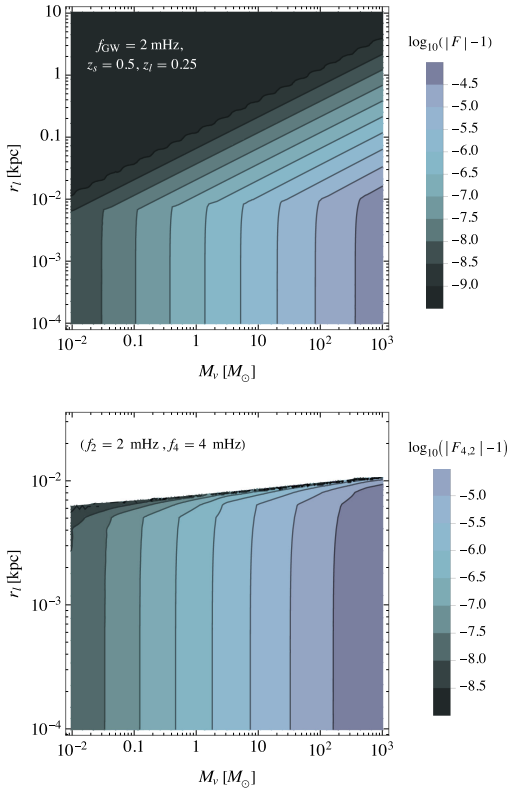


FIG. 1. Upper panel: the absolute value of the lensing amplification as a function of the halo mass and the projected distance of the source in the lens plane, for a source with $f_{\mathrm{GW}} = 2$ mHz at $z_s = 0.5$ and a lens at $z_l = 0.25$. Lower panel: the magnitude of the relative amplification factor (7) for the same source-lens configuration as in the upper panel.

[1]We do the numerical computation using the method introduced in [40].

$r_l \equiv \xi_0 y$ of the source in the lens plane. In the geometric optics regime, the amplification effect rapidly decreases with r_l, roughly as $|F| - 1 \propto r_l^{-4}$, but increases with the lens mass,[2] roughly as $|F| - 1 \propto M_{\rm v}^{7/6}$. The dividing distance between the wave optics and geometric optics regimes, on the other hand, increases very slowly with $M_{\rm v}$ remaining slightly below $r_l \approx 10$ pc in the shown mass range.

Whereas a circular binary emits GWs only at twice its orbital frequency $f_{\rm orb}$, an eccentric binary emits at all integer multiples of it [1],

$$f_n = n f_{\rm orb}. \tag{6}$$

These different harmonics get amplified by different amounts because the lensing effect is frequency dependent. This is the dark timbre effect. In the lower panel of Fig. 1, we show the magnitude of the ratio of amplification factors that characterize the shift in the timbre,

$$F_{i,j} \equiv \frac{|F(f_i)|}{|F(f_j)|} e^{i\Delta\phi_{i,j}}, \qquad \Delta\phi_{i,j} \equiv \phi_i - \phi_j, \tag{7}$$

between the second and the fourth harmonic of an eccentric binary with orbital frequency $f_{\rm orb} = 1$ mHz. As expected, there is no difference in the geometric optics regime's amplification factor for high-impact parameters. Only the halos in the wave optics regime, $r_l < 10$ pc, modify the timbre of the signal. The phase difference is similar in magnitude to the amplification factor ratio.

III. TOTAL LENS EFFECT

The effect of a single lens is characterized by the lens redshift z_l, the projected distance of the source in the lens plane r_l, and the halo mass $M_{\rm v}$. We estimate the modification of the timbre caused by all of the halos the signal encounters in the wave optics regime on its way from the source to the detector by generating configurations of the heaviest lenses and estimating the expected effect from the light halos.

We consider only halos inside the wave optics regime, $r_l < r_{\rm max}(z_l, M_{\rm v})$, since those are the only ones causing a frequency-dependent effect. The total number of halos within that radius is

$$N = \int {\rm d}M_{\rm v} \int_0^{z_s} {\rm d}z_l \int_0^{r_{\rm max}} {\rm d}r_l \frac{2\pi(1+z_l)^2 r_l}{H(z_l)} \frac{{\rm d}n(z_l)}{{\rm d}M_{\rm v}}, \tag{8}$$

where ${\rm d}n(z_l)/{\rm d}M_{\rm v}$ is the HMF which we compute using the Sheth-Tormen approximation [41–43] with the CDM power spectrum [44]. From Eq. (8), we can identify the probability densities of the lens parameters:

$$P(z_l, M_{\rm v}) = \frac{\pi(1+z_l)^2 r_{\rm max}^2}{N H(z_l)} \frac{{\rm d}n(z_l)}{{\rm d}M_{\rm v}},$$
$$P(r_l | z_l, M_{\rm v}) = \theta(r_l - r_{\rm max}) \frac{2 r_l}{r_{\rm max}^2}. \tag{9}$$

We generate realizations of the lenses with mass $M_{\rm v} > 0.01 M_\odot$ that the signal encounters by sampling the probability distributions (9) and approximate their total effect by taking a product of the amplifications caused by each of the lenses (see Appendix B):

$$F_H(f) \approx \prod_{j=1}^{N_{\rm h}} F(f | M_{{\rm v},j}, z_{l,j}, r_{l,j})$$
$$\approx 1 + \sum_{j=1}^{N_h} \left[F(f | M_{{\rm v},j}, z_{l,j}, r_{l,j}) - 1 \right]. \tag{10}$$

Here $N_{\rm h}$ is sampled from a Poisson distribution with mean N in the mass range $M_{\rm v} > 0.01 M_\odot$ and the approximation holds because $F(f | M_{{\rm v},j}, z_{l,j}, r_{l,j}) \approx 1$. We generate 10^3 realizations to estimate the distribution of $F_H(f)$.

For light halos with $M_{\rm v} < 0.01 M_\odot$, the Poisson fluctuations are negligible because their number inside the wave optics regime is very large. Therefore, the effect they cause does not change significantly between different paths, and the distribution of the total lensing factor

$$F_{\rm tot}(f) = F_H(f) F_L(f) \approx F_H(f) + F_L(f) - 1 \tag{11}$$

can then be estimated by convoluting the distribution of F_H with $\delta(F - F_L + 1)$, where

$$F_L(f) = \sum_j \left[1 + \langle F(f) - 1 \rangle_j N_{{\rm h},j} (1 - e^{-N_{{\rm h},j}}) \right] \tag{12}$$

is the expected effect of the light halos.[3] The sum in (12) is over mass bins, $\langle F(f) - 1 \rangle_j$ is the expected deviation of the lensing factor from unity for halos in the mass bin j, and $N_{{\rm h},j}$ is the number of halos in the wave optics regime in that bin. This results in a translation of the distribution from heavy halos by adding the integrated effect of the light ones.

In the upper panels of Fig. 2 we show the effect of adding the lightest halos. In the left panel, we can see that halos heavier than $100 M_\odot$ do not contribute significantly. The reason is that they are too rare to fill the wave optics lensing tube. The main effect comes from halos in the range

[2] Linear scaling $|F| - 1 \propto M_{\rm v}$ at a fixed frequency is expected if κ and b didn't change with $M_{\rm v}$. We find a slightly stronger scaling mainly because the compactness of the halos mildly decreases with decreasing $M_{\rm v}$.

[3] The HMF scales at small masses roughly as ${\rm d}n/{\rm d}\ln M_{\rm v} \propto M_{\rm v}^{-0.9}$. So, since $|F| - 1 \propto M_{\rm v}^{7/6}$, the total amplification effect integrated over all halo masses is finite.

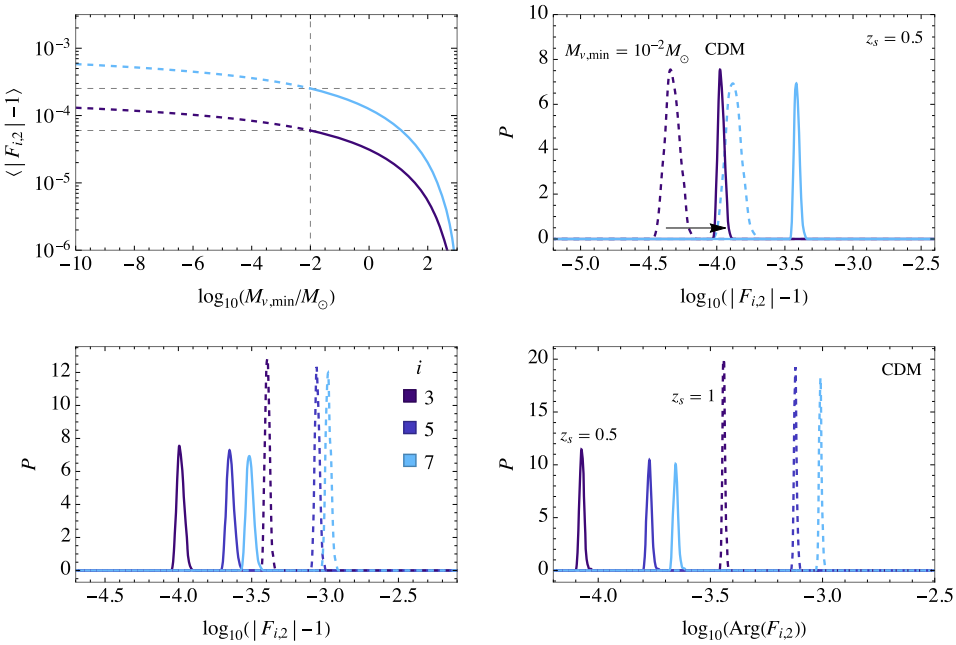


FIG. 2. Upper panels: the left plot shows the expected value for $|F_{i,2}| - 1$ for the third (dark) and seventh (light) harmonic when integrating the halos from $M_{\mathrm{v,min}}$. The right plot shows how the distributions from the heavy halos, $M_{\mathrm{v}} > 10^{-2} M_\odot$, (dashed curves) change when adding the contribution from the light ones. Lower panels: the probability distributions of $|F_{i,2}| - 1$ and $\mathrm{Arg}(F_{i,2})$ for $i = 3, 5, 7$ and a source with $f_2 = 2$ mHz at $z_s = 0.5$ (solid) and $z_s = 1$ (dashed).

$0.01 M_\odot - 10 M_\odot$. Below $0.01 M_\odot$ the effect increases very slowly and we have explicitly checked that it converges. Halos lighter than $0.01 M_\odot$ change the total lensing effect by an $\mathcal{O}(1)$ factor. In the right panel, we can see that the distributions are shifted by adding the contribution from the lightest halos.

In the lower panels of Fig. 2 we show the probability distribution of the modulus and the argument of the amplification ratio between the second and the third, the fifth and the seventh harmonic for the same benchmark case as in Fig. 1, for $z_s = 0.5$ (solid lines) and $z_s = 1$ (dashed lines). In the left panel, we see that the differences are of order 3×10^{-3} for $z_s = 0.5$ and up to 10^{-3} for $z_s = 1$. The distributions of the phase differences, shown in the lower right panel, scale similar to the distributions of the amplifications.

Previous GW microlensing studies have focused on searches of halos with $M_{\mathrm{v}} \gtrsim 10^6 M_\odot$, which cause larger effects. However, for realistic DM halo profiles, such as Einasto or NFW, the effect in the geometric limit is simply a constant amplification [3] that is degenerate with the source properties. The lensing tube where wave optics effects are relevant is fixed by the frequency of the GW and the mass of the halo. For LISA frequencies and halos with $M_{\mathrm{v}} \gtrsim 10^6 M_\odot$, the expected number of halos in the wave optics tube is very small and, consequently, it is not expected that LISA would see any such lensed events [3]. In this study, we instead consider much lighter halos $M_{\mathrm{v}} \lesssim 10 M_\odot$ which are expected, assuming the CDM model, to induce wave optics effects in all of the GW events that LISA will see. The next section discusses the precision necessary to test and detect this effect.

IV. DETECTABILITY AND PROSPECTS

To detect the dark timbre effect, we require that the signal with the shifted timbre cannot be fitted well with a template using the standard timbre. In general, the wellness of a template fit is given by [45]

$$\Delta\chi^2 = 4 \min_{\vec{\theta}} \int \mathrm{d}f \frac{|\tilde{h}(f) - \tilde{h}_{\mathrm{T}}(f, \vec{\theta})|^2}{S(f)}, \tag{13}$$

where $\tilde{h}(f)$ refers to the GW signal and $\tilde{h}_{\mathrm{T}}(f, \vec{\theta})$ the template parametrized by $\vec{\theta}$. For a binary that evolves very slowly, the waveform is approximately a sum of delta functions and Eq. (13) simplifies to

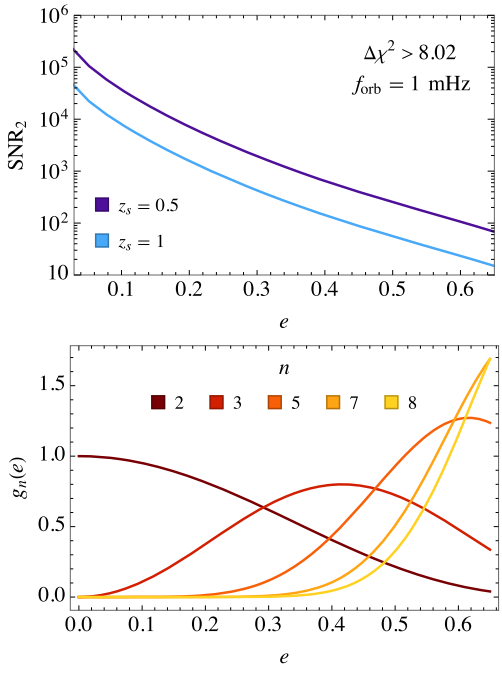


FIG. 3. Upper panel: the SNR of the second harmonic (SNR$_2$) necessary for detection of the shifted timbre as a function of the eccentricity of a binary orbiting at 1 mHz. The dark line is for a source at $z_s = 1$ and the light blue line for $z_s = 0.5$. Lower panel: The relative powers $g_n(e)$ of some of the lowest harmonics.

$$\Delta\chi^2 = 4\min_{\lambda,\phi,e'}\sum_{n=1}^{\infty}\frac{A^2\left|F(f_n)\sqrt{g_n(e)} - \lambda e^{i\phi}\sqrt{g_n(e')}\right|^2}{S(f_n)} = \frac{\mathrm{SNR}_2^2}{g_2(e)}\min_{\lambda,\phi,e'}\sum_{n=1}^{\infty}\frac{S(f_2)}{S(f_n)}\left|F_{n,2}\sqrt{g_n(e)} - \lambda e^{i\phi}\sqrt{g_n(e')}\right|^2, \tag{14}$$

where A characterizes the amplitude of the signal and SNR$_2$ is the signal-to-noise ratio of the second harmonic. The template is parametrized by an amplitude $\lambda \approx 1$, an eccentricity $e' \approx e$, and a phase $\phi \approx 0$. The function $g_n(e)$ gives the relative power radiated in the nth harmonic by a binary with eccentricity e [1]. We show $g_n(e)$ for some of the lowest harmonics in the lower panel of Fig. 3.

We use the expected LISA spectral sensitivity $S(f)$ [46,47] and include noise arising from the galactic and extragalactic foregrounds [48]. We considered sources at $z_s = 0.5$ and at $z_s = 1$ with $f_2 = 2$ mHz. To assess the detectability of the shifted timbre effect, we consider only the harmonics $n = 2$–7 that dominate the signal for $e \lesssim 0.6$, as seen in the lower panel of Fig. 3. We determine how much SNR$_2$ and e are required to detect the shifted timbre. We use the threshold $\Delta\chi^2 > 8.02$ corresponding to 2σ Confidence Level (CL) for a template with three parameters. The results are shown in the upper panel of Fig. 3. For example, for $z_s = 0.5$ and $e = 0.3$ the dark timbre is detected if SNR$_2 \gtrsim 3000$ while for $e = 0.5$ it is enough to have SNR$_2 \lesssim 500$. For $z_s = 1$, approximately factor four smaller SNR$_2$ is sufficient. An example of a source with approximately these properties could be an IMBH binary with $\mathcal{M} = \mathcal{O}(10^3/M_\odot)$ and $\eta \sim 0.1$. In scenarios with light supermassive BH seeds, the number of detectable binaries of this type would be $\mathcal{O}(100)$/yr and they would enter the LISA band at $\mathcal{O}$(years) before merger [49]. For some time, they would be approximately monochromatic with SNR of $\mathcal{O}(10^3)$. As shown in [50], also the referred eccentricities are realistic.

The precision at which the amplitudes of a harmonic i can be resolved is approximately

$$\frac{\sigma_{A_i}}{A_i} \sim \frac{1}{\mathrm{SNR}_i}, \tag{15}$$

which is easy to show from the Fisher analysis [51]. This implies that a higher SNR$_2$ is necessary to resolve the relative amplification factors $F_{i,j}$ rather than to detect the dark timbre. Optimistically, if a population of eccentric binaries is detected with SNR$_i \sim 10^3$ it would be possible to reconstruct the Probability Density Functions (PDFs) of $F_{i,j}$ and probe the HMF down to $M_{\rm v} \ll 10M_\odot$. However, already the detection of one sufficiently eccentric binary with SNR$_2 \gtrsim 10^2$ at $z_s = 0.5$ would tell about the existence of halos with $M_{\rm v} \lesssim 10M_\odot$.

We note that since the dark timbre effect is very small, corrections beyond the Newtonian order [1] in the different harmonics coefficients may become relevant, although we assume that the binaries are very far from merging, furthermore considering more realistic waveforms with higher modes would potentially help in the detectability [4]. Moreover, detecting the dark timbre requires identifying the frequency harmonics. LISA is expected to observe several nearly monochromatic GWs simultaneously at different frequencies. Some of these GWs may be harmonics from an eccentric IMBH binary, while others may originate, e.g., from galactic WD binaries [52]. Consequently, resolving which of these GWs are harmonics emitted by the same source can be challenging. The effects of higher modes and the identification problem of the frequency harmonics are left for future studies.

V. CONCLUSIONS

We have estimated for the first time the integrated lensing effect of light halos on the mid-frequency (mHz) GW signals. We have proposed to test these effects using the different harmonics of an eccentric binary. We have shown that the wave optics effects due to the low-mass DM halos, $M_{\rm v} \lesssim 10M_\odot$, induce frequency-dependent changes in the amplitude and phase of the harmonics—the timbre

of the signal. This shifted timbre is detectable in the seven dominant harmonics of the signal for signal-to-noise ratios between 100–10^3 if the binary eccentricity is $e = 0.3$–0.6 and it is $z_s = 0.5$. If such binaries exist, LISA could probe the dark timbre. This would open a new avenue to test low-mass DM halos and thus provide new insights into the nature of DM.

ACKNOWLEDGMENTS

We thank Luca Marzola and Hardi Veermäe for useful suggestions. This work was supported by the Estonian Research Council Grants No. PRG803, No. PSG869, No. RVTT3, and No. RVTT7 and the Center of Excellence program TK202. The work of V. V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101065736.

APPENDIX A: VALIDITY OF THE BORN AND EIKONAL APPROXIMATIONS

In the following, we make explicit the assumptions behind the formula of the amplification factor F used in the main text and showcase their validity.

To compute the scatter of a GW from the DM halo we will consider the metric around the halo to be given by

$$\mathrm{d}s^2 = -(1+2U(\boldsymbol{x}))\mathrm{d}t^2 + a^2(1-2U(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}^2, \tag{A1}$$

where U is the gravitational potential of the halo and a denotes the scale factor that we consider as a constant fixed by the halo redshift. Under the Born approximation, we consider only the leading order terms in $U \ll 1$. Moreover, we neglect the change in the polarization of the GW which is of the order $U \ll 1$ and consider only its amplitude $\phi(t,\boldsymbol{x})$. Then, the Fourier transform $\tilde{\phi}(\omega, r)$ of the scalar wave follows the wave equation:

$$(a^{-2}\nabla^2 + \omega^2)\tilde{\phi}(\omega,\boldsymbol{x}) = 4\omega^2 U\tilde{\phi}(\omega,\boldsymbol{x}). \tag{A2}$$

We want to express the solution as

$$\tilde{\phi}(\omega,\boldsymbol{x}) = F(\omega,\boldsymbol{x})\tilde{\phi}_{\mathrm{NL}}(\omega,\boldsymbol{x}), \tag{A3}$$

where the $\tilde{\phi}_{\mathrm{NL}}(\omega,\boldsymbol{x})$ is the solution of Eq. (A2) for $U = 0$. Plugging (A3) into the equation of motion (A2) we get that the amplification function $F(\omega, r)$ has to fulfill

$$\frac{1}{(ar)^2}\nabla_{\hat{\theta}}^2 F + \frac{1}{a^2}\frac{\partial^2 F}{\partial r^2} + \frac{2i\omega}{a}\frac{\partial F}{\partial r} = 4\omega^2 UF. \tag{A4}$$

Assuming that

$$\left|\frac{1}{a^2}\frac{\partial^2 F}{\partial r^2}\right| \ll \left|\frac{2i\omega}{a}\frac{\partial F}{\partial r}\right|, \tag{A5}$$

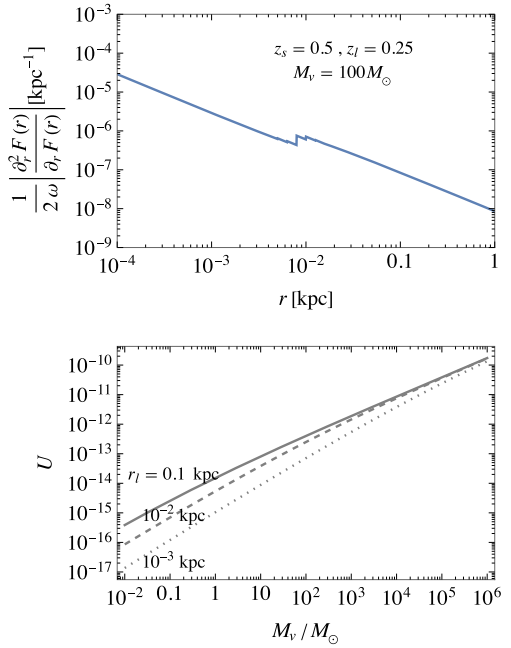


FIG. 4. Upper panel: comparison of the second and first derivatives of F. Lower panel: values of U for the typical impact parameters that contribute to the dark timbre.

Eq. (A4) simplifies to

$$\frac{i}{a}\frac{\partial F}{\partial r} = \frac{1}{2\omega(ar)^2}\nabla_{\hat{\theta}}^2 F + 2\omega UF, \tag{A6}$$

which is a Schrödinger equation that can be solved with the path integral approach that gives the formula of F we use throughout the paper.

In the lower panel of Fig. 4 we plot the NFW gravitational potential,

$$U(r) = -\frac{4\pi\rho_0 r_s^3}{r}\log\left(1+\frac{r}{r_s}\right), \tag{A7}$$

to show that the Born approximation is valid for the impact parameters we typically consider for the dark timbre. In the upper panel we demonstrate the validity of the eikonal approximation.

APPENDIX B: MULTILENS COMPUTATION

The solution of Eq. (A6) is given by the path integral

$$F = \int \mathcal{D}\boldsymbol{\theta} e^{i\int_0^{r_{N+1}} \mathrm{d}r\mathcal{L}}, \tag{B1}$$

where $\boldsymbol{\theta} = \theta(\cos(\phi), \sin(\phi))$ parametrizes the paths that the GW can take from the source at $r = 0$, $\boldsymbol{\theta} = 0$ to the observer at $r = r_{N+1}$, $\boldsymbol{\theta} = \boldsymbol{\theta}_{N+1} = 0$ and

$$\mathcal{L} = \frac{1}{2}\omega(ar)^2|\dot{\boldsymbol{\theta}}(r)|^2 - 2\omega U. \tag{B2}$$

We divide the path into N slices perpendicular to the radial vector from the source to the detector at distances $r = r_j$ of thickness ϵ, so that $r_{N+1} = (N+1)\epsilon$. Some of these slices include a lens. In the thin lens approximation, we project the gravitational potentials of each lens to a plane,

$$U(r, \boldsymbol{\theta}) = \frac{1}{2}\sum_{j=1}^{N} \psi_j(\boldsymbol{\theta})\delta(r - r_j) \tag{B3}$$

and consider paths that change direction only at the lens planes,

$$|\dot{\boldsymbol{\theta}}(r)| = \sum_{j=1}^{N} \Theta(r - r_j)\Theta(r_{j+1} - r)\frac{r_j r_{j+1}}{ar\epsilon}|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j|, \tag{B4}$$

where $\boldsymbol{\theta}_j$ denotes the position where the path crosses the jth lens plane and $r_{i,j} \equiv r_i r_j/(r_i - r_j)$. Then the integrated Lagrangian can be expressed as

$$\int_0^{r_{N+1}} \mathrm{d}r\mathcal{L} = \omega\sum_{j=1}^{N}\left[\epsilon\frac{r_j r_{j+1}}{2}\left|\frac{\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j+1}}{\epsilon}\right|^2 - \Psi_j(\boldsymbol{\theta}_j)\right], \tag{B5}$$

and THE amplification function as

$$F = \int\prod_{j=1}^{N}\frac{\omega r_j r_{j+1}}{2\pi i\epsilon}\mathrm{d}^2\boldsymbol{\theta}_j \times \exp\left[i\omega\left(\frac{r_j r_{j+1}}{2\epsilon}|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j|^2 - \psi_j(\boldsymbol{\theta}_j)\right)\right]. \tag{B6}$$

Here $\psi_j = 0$ for those slices that do not include a lens. Assuming that for all relevant paths $\omega\psi_j \ll 1$, we can expand to the lowest order in powers of $\omega\psi_j$:

$$\begin{aligned} F &\approx \int\prod_{j=1}^{N}\frac{\omega r_j r_{j+1}}{2\pi i\epsilon}\mathrm{d}^2\boldsymbol{\theta}_j \exp\left[i\omega\frac{r_j r_{j+1}}{2\epsilon}|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j|^2\right] \times [1 - i\omega\psi_j(\boldsymbol{\theta}_j)] \\ &\approx 1 - i\omega\sum_{l=1}^{N}\int\frac{\omega r_l r_{l+1}}{2\pi i\epsilon}\mathrm{d}^2\boldsymbol{\theta}_l\left[\prod_{j\neq l}\frac{\omega r_j r_{j+1}}{2\pi i\epsilon}\mathrm{d}^2\boldsymbol{\theta}_j\right] \times \psi_l(\boldsymbol{\theta}_l)\exp\left[i\omega\sum_{j=1}^{N}\frac{r_j r_{j+1}}{2\epsilon}|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j|^2\right]. \end{aligned} \tag{B7}$$

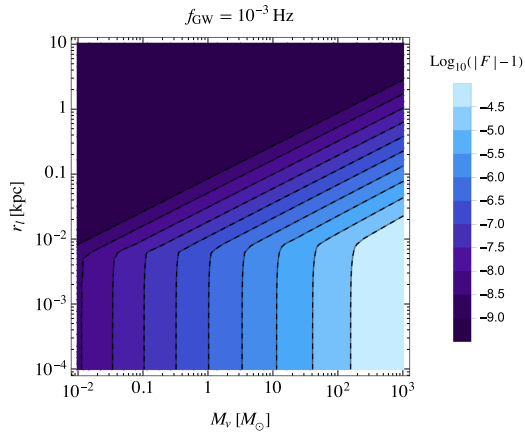


FIG. 5. Comparison between the amplification function for a single lens computed without (solid contours) and with (dashed contours) the expansion in $\omega\psi_j$.

The integrals for $j < l$ give unity and [53]

$$\sum_{j=l}^{N}\frac{r_j r_{j+1}}{2\epsilon}|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j|^2 = \epsilon\frac{r_l r_{N+1}}{r_{l,N+1}}|\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}_l|^2 + \sum_{j=l+1}^{N} r_j^2\frac{r_l r_{j+1}}{r_{l,j}}|\boldsymbol{\theta}_j - \boldsymbol{u}_{l,j}|^2, \tag{B8}$$

where $r_{l,j} = r_l - r_j$ and $\boldsymbol{u}_{l,j} = [r_l\boldsymbol{\theta}_j + (j-l)r_{j+1}\boldsymbol{\theta}_{j+1}]/(jr_{l,j+1})$. So, all integrals except that over $\boldsymbol{\theta}_l$ are trivial and give $\epsilon r_{N+1}/(r_{l+1}r_{l,N+1})$. This results in

$$\begin{aligned} F &\approx 1 - i\omega\sum_{l=1}^{N}\int\frac{\omega r_l r_{N+1}}{2\pi i r_{l,N+1}}\mathrm{d}^2\boldsymbol{\theta}_l\psi_l(\boldsymbol{\theta}_l) \times \exp\left[i\omega\frac{r_l r_{N+1}}{r_{l,N+1}}|\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}_l|^2\right] \\ &\approx 1 - \sum_{l=1}^{N}(F_j - 1). \end{aligned} \tag{B9}$$

In Fig. 5, we show that the expansion in $\omega\psi_j \ll 1$ works well for the NFW lenses. The solid and dashed contours are computed without and with the expansion and their excellent agreement is evident.

[1] P. C. Peters and J. Mathews, Phys. Rev. **131**, 435 (1963).
[2] R. Takahashi and T. Nakamura, Astrophys. J. **595**, 1039 (2003).
[3] M. Fairbairn, J. Urrutia, and V. Vaskonen, J. Cosmol. Astropart. Phys. 07 (2023) 007.
[4] M. Çalışkan, L. Ji, R. Cotesta, E. Berti, M. Kamionkowski, and S. Marsat, Phys. Rev. D **107**, 043029 (2023).
[5] M. Çalışkan, N. Anil Kumar, L. Ji, J. M. Ezquiaga, R. Cotesta, E. Berti, and M. Kamionkowski, Phys. Rev. D **108**, 123543 (2023).
[6] S. Savastano, G. Tambalo, H. Villarrubia-Rojo, and M. Zumalacarregui, Phys. Rev. D **108**, 103532 (2023).
[7] G. Tambalo, M. Zumalacárregui, L. Dai, and M. H.-Y. Cheung, Phys. Rev. D **108**, 103529 (2023).
[8] H. Ma, Y. Lu, Z. Chen, and Y. Chen, Mon. Not. R. Astron. Soc. **524**, 2954 (2023).
[9] P. Cremonese, J. M. Ezquiaga, and V. Salzano, Phys. Rev. D **104**, 023503 (2021).
[10] J. Urrutia, V. Vaskonen, and H. Veermäe, Phys. Rev. D **108**, 023507 (2023).
[11] G. Brando, S. Goyal, S. Savastano, H. Villarrubia-Rojo, and M. Zumalacárregui, Phys. Rev. D **111**, 024068 (2025).
[12] R. Takahashi, Astrophys. J. **835**, 103 (2017).
[13] T. Morita and J. Soda, arXiv:1911.07435.
[14] J. M. Ezquiaga, W. Hu, and M. Lagos, Phys. Rev. D **102**, 023531 (2020).
[15] T. Suyama, Astrophys. J. **896**, 46 (2020).
[16] L. Dai and T. Venumadhav, arXiv:1702.04724.
[17] J. M. Ezquiaga, D. E. Holz, W. Hu, M. Lagos, and R. M. Wald, Phys. Rev. D **103**, 064047 (2021).
[18] P. Bode, J. P. Ostriker, and N. Turok, Astrophys. J. **556**, 93 (2001).
[19] L. Hui, J. P. Ostriker, S. Tremaine, and E. Witten, Phys. Rev. D **95**, 043541 (2017).
[20] K. K. Rogers and H. V. Peiris, Phys. Rev. Lett. **126**, 071302 (2021).
[21] L. Barack and C. Cutler, Phys. Rev. D **75**, 042003 (2007).
[22] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. L. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, Phys. Rev. D **95**, 103012 (2017).
[23] C. P. L. Berry, S. A. Hughes, C. F. Sopuerta, A. J. K. Chua, A. Heffernan, K. Holley-Bockelmann, D. P. Mihaylov, M. C. Miller, and A. Sesana, Bull. Am. Astron. Soc. **51**, 42 (2019).
[24] J. R. Gair, S. Babak, A. Sesana, P. Amaro-Seoane, E. Barausse, C. P. L. Berry, E. Berti, and C. Sopuerta, J. Phys. Conf. Ser. **840**, 012021 (2017).
[25] O. A. Hannuksela, K. W. K. Wong, R. Brito, E. Berti, and T. G. F. Li, Nat. Astron. **3**, 447 (2019).
[26] M. Bonetti and A. Sesana, Phys. Rev. D **102**, 103023 (2020).
[27] M. Bartelmann and P. Schneider, Phys. Rep. **340**, 291 (2001).
[28] J. Zavala and C. S. Frenk, Galaxies **7**, 81 (2019).
[29] P. Schneider, J. Ehlers, and E. Falco, *Gravitational Lenses*, Astronomy and Astrophysics Library (Springer, New York, 2012).
[30] G. Braga, A. Garoffolo, A. Ricciardone, N. Bartolo, and S. Matarrese, J. Cosmol. Astropart. Phys. 11 (2024) 031.
[31] A. D. Ludlow, S. Bose, R. E. Angulo, L. Wang, W. A. Hellwing, J. F. Navarro, S. Cole, and C. S. Frenk, Mon. Not. R. Astron. Soc. **460**, 1214 (2016).
[32] J. Wang, S. Bose, C. S. Frenk, L. Gao, A. Jenkins, V. Springel, and S. D. M. White, Nature (London) **585**, 39 (2020).
[33] M. S. Delos, Phys. Rev. D **100**, 083529 (2019).
[34] E. Hayashi, J. F. Navarro, J. E. Taylor, J. Stadel, and T. R. Quinn, Astrophys. J. **584**, 541 (2003).
[35] F. C. van den Bosch, G. Ogiya, O. Hahn, and A. Burkert, Mon. Not. R. Astron. Soc. **474**, 3043 (2018).
[36] J. E. Taylor and A. Babul, Astrophys. J. **559**, 716 (2001).
[37] F. C. van den Bosch, G. Tormen, and C. Giocoli, Mon. Not. R. Astron. Soc. **359**, 1029 (2005).
[38] V. Berezinsky, V. Dokuchaev, and Y. Eroshenko, Phys. Rev. D **77**, 083519 (2008).
[39] J. Penarrubia and A. J. Benson, Mon. Not. R. Astron. Soc. **364**, 977 (2005).
[40] R. Takahashi, Wave effects in the gravitational lensing of gravitational waves from chirping binaries, Ph.D. thesis, Department of Physics, Kyoto University, 2004.
[41] W. H. Press and P. Schechter, Astrophys. J. **187**, 425 (1974).
[42] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophys. J. **379**, 440 (1991).
[43] R. K. Sheth and G. Tormen, Mon. Not. R. Astron. Soc. **308**, 119 (1999).
[44] D. J. Eisenstein and W. Hu, Astrophys. J. **496**, 605 (1998).
[45] P. Jaranowski and A. Krolak, Living Rev. Relativity **8**, 3 (2005).
[46] P. Amaro-Seoane *et al.*, arXiv:1702.00786.
[47] T. Robson, N. J. Cornish, and C. Liu, Classical Quantum Gravity **36**, 105011 (2019).
[48] M. Lewicki and V. Vaskonen, Eur. Phys. J. C **83**, 168 (2023).
[49] J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen, Astrophys. J. **964**, 11 (2024).
[50] E. K. Porter and A. Sesana, arXiv:1005.5296.
[51] E. Poisson and C. M. Will, Phys. Rev. D **52**, 848 (1995).
[52] V. Korol, E. M. Rossi, P. J. Groot, G. Nelemans, S. Toonen, and A. G. A. Brown, Mon. Not. R. Astron. Soc. **470**, 1894 (2017).
[53] T. T. Nakamura and S. Deguchi, Prog. Theor. Phys. Suppl. **133**, 137 (1999).

Appendix 4

IV
G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Did JWST observe imprints of axion miniclusters or primordial black holes? *Phys. Rev. D*, 107(4):043502, 2023

Did JWST observe imprints of axion miniclusters or primordial black holes?

Gert Hütsi,[1,*] Martti Raidal,[1,†] Juan Urrutia,[1,‡] Ville Vaskonen,[1,2,3,§] and Hardi Veermäe[1,‖]

[1]*Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia*
[2]*Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy*
[3]*Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 15 November 2022; accepted 23 January 2023; published 2 February 2023)

The James Webb Space Telescope has detected surprisingly luminous early galaxies that indicate a tension with the Λ cold dark matter. Motivated by scenarios including axion miniclusters or primordial black holes, we consider power-law modifications of the matter power spectrum. We show that the tension could be resolved if dark matter consists of 2×10^{-18} eV axions or if a fraction $f_{\rm PBH} > 0.005$ of dark matter is composed of compact heavy $4 \times 10^6 M_\odot (f_{\rm PBH}/0.005)^{-1}$ structures such as primordial black hole clusters. However, in both cases, the star formation efficiency needs to be significantly enhanced.

DOI: 10.1103/PhysRevD.107.043502

I. INTRODUCTION

The long-awaited next-generation space telescope—the James Webb Space Telescope (JWST) [1]—is finally delivering data. Arguably the most intriguing result so far has been the detection of high-redshift galaxies with surprisingly high stellar masses [2–7]. This is somewhat reminiscent of what happened almost three decades ago with the Hubble Space Telescope (HST) [8] discovering the unexpected richness of galaxies in the Hubble Deep Field. These early hints indicated severe challenges for the then standard Einstein–de Sitter cosmology and together with subsequent observational data (most importantly cosmic microwave background (CMB), type Ia supernovae and large-scale structure data from the large redshift surveys) led to the establishment of the current standard cosmological model—the ΛCDM model. If true, the new results from JWST might similarly call for a significant modification of the ΛCDM.

There have been several attempts in trying to make the concordance Λ cold dark matter (CDM) cosmology compatible with JWST measurements. For example, reduced dust attenuation in high-redshift galaxies makes them appear brighter [9,10]. However, it seems that altering the dust production alone is not enough to relieve these tensions—one also has to make the star formation significantly more effective [11]. Even then, by pushing the star formation efficiency to the extremes, it is still hard to accommodate these new observations within the ΛCDM cosmology [12,13]. Moreover, a high star formation efficiency leads to a high abundance of ionizing photons that may contradict the measurements of the cosmic reionization history. A potential solution was proposed in [14] within fuzzy DM models that suppress the abundance of small-scale structures.

Although there are significant uncertainties involved in pushing the Lyman-break technique to so high redshifts—difficulties in interpretation of spectral breaks or additional contamination due to emission lines [15,16]—in this paper we assume these observational inferences to be true and investigate beyond the ΛCDM physics that could resolve the tension. Different modifications to the small-scale physics boosting the abundance of $\mathcal{O}(10^{11})M_\odot$ dark matter (DM) halos at high redshifts $z > 9$ have been already proposed. These proposals include heavy primordial black holes (PBHs) [17], non-Gaussianities in primordial fluctuations [18], and modified dark energy equation of state [19].

In this paper, we focus on scenarios, including axion miniclusters or heavy PBHs, which boost the matter power spectrum at $k > 3h\ \mathrm{Mpc}^{-1}$. Our analysis uses extreme value statistics (EVS) [20,21] and considers two of the most extremal galaxies observed so far by JWST: CEERS-1749 of stellar mass $\log_{10}(M_*/M_\odot) = 9.6 \pm 0.2$ at redshift $z = 16 \pm 0.6$ [2] and Galaxy 14924 of stellar mass $\log_{10}(M_*/M_\odot) = 10.9 \pm 0.3$ at redshift $z = 9.9 \pm 0.5$ [3].

II. MATTER POWER SPECTRUM

The matter power spectrum is strictly constrained at scales $k \lesssim 3h\ \mathrm{Mpc}^{-1}$ by measurements of galaxy clustering, Lyman-alpha forest data and the UV luminosity function [22–24]. The latter is obtained using data from the HST and will improve with the JWST observations [25,26]. These observations agree well with the predictions of the

*gert.hutsi@kbfi.ee
†martti.raidal@cern.ch
‡juan.urrutia@kbfi.ee
§ville.vaskonen@pd.infn.it
‖hardi.veermae@cern.ch

 © 2023 American Physical Society

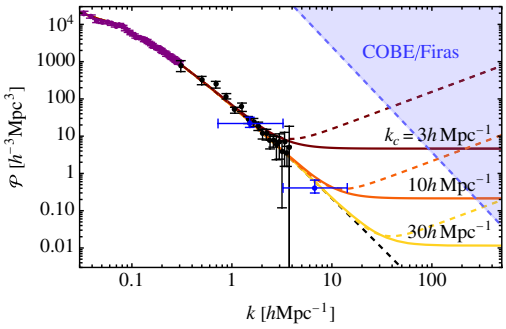


FIG. 1. The matter power spectrum for standard ΛCDM (black dashed) and for three values of k_c with $n = 0$ (solid) and $n = 1$ (dashed). The points with errors correspond to SDSS measurements of galaxy clustering from luminous red galaxies (purple) [22] and Lyman-α forest (black) [23] and HST measurements of UV luminosity function (blue) [24]. The blue dashed curve shows the maximal cutoff scale for k^4 growth of the adiabatic curvature power spectrum from COBE/Firas bound on μ and y distortions [28].

standard ΛCDM model. At smaller scales, significant deviations from the CDM matter power spectrum are possible and affect the abundance of $M \lesssim 10^{11} M_\odot$ halos. In particular, a matter power spectrum that exceeds the ΛCDM prediction at scales $\mathcal{O}(10)h$ Mpc^{-1} gives rise to an increased abundance of massive early galaxies, perhaps being compatible with the JWST observations.

We consider power-law modifications to the matter power spectrum at small scales:

$$\mathcal{P}(k) = \mathcal{P}_{\rm CDM}(k) + \mathcal{P}_{\rm CDM}(k_c)\left(\frac{k}{k_c}\right)^n, \tag{1}$$

where $\mathcal{P}_{\rm CDM}(k)$ denotes the standard ΛCDM matter power spectrum, $k_c > 3h$ Mpc^{-1} the scale above which the power-law behavior takes over and n the spectral index of the power-law. The power-law part is cut at some scale $k_{\rm cut} > k_c$. As discussed later, our results are insensitive to the exact value as long as $k_{\rm cut} \gtrsim 30h$ Mpc^{-1}. For the CDM matter power spectrum we consider a nearly scale-invariant curvature power spectrum with the spectral index $n_s = 0.96$ and the fitted transfer function from Ref. [27]. In Fig. 1 we show the CDM matter power spectrum and examples of the modified matter power spectra (1) together with the constraints from measurements of galaxy clustering [22], Lyman-alpha forest data [23] and UV luminosity function [24].

Deviations of the form of Eq. (1) from the standard ΛCDM matter power spectrum can arise, for example, in the following scenarios:

(1) In axionlike particle DM models, in which the global $U(1)$ symmetry is broken after the cosmic inflation, large amplitude small-scale fluctuations are generated via the Kibble mechanism leading to the formation of axion miniclusters [29]. This corresponds to an $n = 0$ contribution to the matter power spectrum with an amplitude [30–34]

$$A = \frac{4}{5}\frac{6\pi^2}{D(z_{\rm eq})^2} k_{\rm cut}^{-3}, \tag{2}$$

where $D(z_{\rm eq})$ denotes the growth factor (normalized so that $D(0) = 1$) at matter-radiation equality and $k_{\rm cut}$ the cutoff given by the scale that entered the horizon at the moment when the axion of mass m_a began to oscillate (i.e., $3H = m_a$ at $aH = k_{\rm cut}$),[1]

$$k_{\rm cut} \simeq 300 \text{ Mpc}^{-1} \sqrt{\frac{m_a}{10^{-18} \text{ eV}}}. \tag{3}$$

Using a power law approximation for $\mathcal{P}_{\rm CDM}$ around $k_c = \mathcal{O}(10)h$ Mpc^{-1}, we find that the scale k_c, at which $\mathcal{P}_{\rm CDM}(k_c) = A$, is given by

$$k_c \approx 3h \text{ Mpc}^{-1} \left(\frac{m_a}{10^{-18} \text{ eV}}\right)^{0.6}, \tag{4}$$

thus $k_{\rm cut} \gg k_c$ is satisfied in the relevant k_c range.

(2) In models with heavy PBHs the graininess associated with the discrete Poisson distribution of PBHs [35,36] or PBH clusters [37] (but possibly also other compact structures) generates an $n = 0$ contribution to the matter power spectrum with an amplitude [38,39]

$$A = \frac{1}{n_{\rm PBH}} \frac{f_{\rm PBH}^2}{D(z_{\rm eq})^2} = \frac{6\pi^2 f_{\rm PBH}^2}{D(z_{\rm eq})^2} k_{\rm cut}^{-3}, \tag{5}$$

where $f_{\rm PBH}$ denotes the fraction of DM in PBHs, $m_{\rm PBH}$ the PBH mass and $\rho_{\rm DM}$ the DM density. The cutoff scale of the spectrum is given by the average separation of PBHs,

$$\begin{aligned} k_{\rm cut} &\simeq (6\pi^2 n_{\rm PBH})^{1/3} \\ &= 900h \text{ Mpc}^{-1} \left(\frac{f_{\rm PBH} 10^4 M_\odot}{m_{\rm PBH}}\right)^{1/3}. \end{aligned} \tag{6}$$

Below this scale, one expects a single PBH in the volume of the corresponding comoving sphere and thus the seed effect [35,36,40–42] begins to dominate over the Poisson one.[2] The scale k_c is

[1] We assume, for simplicity, that the axion mass is temperature independent. For further discussion, see e.g. Refs. [30,33].

[2] Importantly, the usual Press-Schechter formalism breaks down above $k_{\rm cut}$ as the fluctuations in the PBH density are non-Gaussian. Note that the scale at which the PBH-induced density fluctuations become non-linear, $k_{\rm cut,NL} \approx (n_{\rm PBH}/f_{\rm PBH})^{1/3} > k_{\rm cut}$ is smaller.

$$k_c \approx 6h\ \mathrm{Mpc}^{-1}\left(\frac{f_{\mathrm{PBH}} m_{\mathrm{PBH}}}{10^4 M_\odot}\right)^{-0.4}. \quad (7)$$

As Eq. (4), the latter approximation works well at $k_c = \mathcal{O}(10)h\ \mathrm{Mpc}^{-1}$. We find that $k_c < k_{\mathrm{cut}}$ if $f_{\mathrm{PBH}} > 10^{-4}(m_{\mathrm{PBH}}/10^4 M_\odot)^{-0.09}$. Because CMB observations constrain $f_{\mathrm{PBH}} < 10^{-8}$ for $m_{\mathrm{PBH}} \gtrsim 10^4 M_\odot$ [43],[3] explanations for the JWST observations relying on the Poisson effect from heavy PBHs are not viable. However, the accretion constraint can be softened if lighter PBHs were formed in dense clusters (as described e.g. in [45–49]) with masses above $10^4 M_\odot$. Such clusters would then act as sources for the Poisson fluctuations instead of the individual PBHs [37].

We must stress that here we focused on the Poisson effect. As massive BHs may play an important role in the formation of the first galaxies [50], we cannot rule out the so-called seed effect from PBHs with $f_{\mathrm{PBH}} \leq 10^{-8}$ as a potential explanation. Assuming a monochromatic mass distribution, the expected number of PBHs of initial mass M_{PBH} in the JWST light cone [see Eq. (13)] at redshifts $15 < z < 17$ is

$$N_{\mathrm{PBH},15<z<17} = 2.8 \times \left(\frac{f_{\mathrm{PBH}}}{10^{-9}}\right)\left(\frac{M_{\mathrm{PBH}}}{10^6 M_\odot}\right)^{-1}. \quad (8)$$

However, the number of massive PBHs cannot be straightforwardly translated into the number of luminous galaxies (such as CEERS-1749). Due to its nonlinear nature and dependence on uncertain baryonic physics, the PBH seed scenario requires a dedicated study.

(3) Enhanced adiabatic power spectra can be generated, for example, in single field inflation if the inflaton potential has features which slow down the inflaton field [51–53]. Such scenarios are relevant for PBH formation [54,55] and are thus mostly studied in that context. However, they could be realized without any relation to PBHs. In typical models of single-field inflation, a curvature power spectrum with a constant spectral index n_s at $k < k_c$ can grow as $k^{5-|2-n_s|}$ at $k > k_c$ [56]. For $n_s = 0.96$ this translates in Eq. (1) to $n = 0.68$ for $k_c \simeq 1h\ \mathrm{Mpc}^{-1}$ and $n = 0.14$ for $k_c \gg 100h\ \mathrm{Mpc}^{-1}$. Although an k^4 growth of the curvature power spectrum is typical [57], steeper spectra can be realized in inflationary scenarios with increasingly fine-tuned features [58]. In the following, we consider at most $n = 1$.

In the first two cases, the power-law contribution to the matter power spectrum constitutes an isocurvature component to the matter fluctuations. The COBE/Firas constraint shown in Fig. 1 applies to the adiabatic component, while the constraint from μ and y distortions on the isocurvature component is much weaker [28]. In particular, the cutoff scales in the first two scenarios are far below that constraint.

In the third case, the fluctuations are adiabatic, and the COBE/Firas constraint excludes PBH formation if $k_c \lesssim 10^2 h\ \mathrm{Mpc}^{-1}$ unless the growth of the curvature power spectrum is relatively slow, less than $\propto k^3$, which translates to $n < 0$ in Eq. (1). However, steep growth of the curvature power spectrum at scales $k_c = \mathcal{O}(10)h\ \mathrm{Mpc}^{-1}$ that terminates with amplitude less than four orders of magnitude above the CMB amplitude is not excluded.

III. HALO MASS FUNCTION

The halo mass function can be expressed as

$$\frac{\mathrm{d}n}{\mathrm{d}\ln M} = \frac{\rho_m}{M}\nu f(\nu)\frac{\mathrm{d}\ln\nu}{\mathrm{d}\ln M}, \quad (9)$$

where ρ_m denotes the average present mass density and $\nu \equiv \delta_c(z)^2/\sigma_M^2$. The critical overdensity required for collapse is $\delta_c(z) = 1.686/D(z)$, where $D(z)$ is the growth factor [59],[4]

$$D(z) \propto H(z)\int_z^\infty \mathrm{d}z' \frac{(1+z')}{H(z')^3}, \quad (10)$$

normalized such that $D(0) = 1$. In the Press-Schechter formalism [60] with corrections from ellipsoidal dynamics [61]

$$\nu f(\nu) = A[1 + (q\nu)^{-p}]\sqrt{\frac{q\nu}{2\pi}}e^{-q\nu/2}, \quad (11)$$

where $p = 0.30$, $A = 0.322$ and $q = 0.75$ [62]. The variance σ_M^2 at the mass scale M is

$$\sigma_M^2 = \frac{1}{2\pi^2}\int \mathrm{d}k\, k^2 \mathcal{P}(k) W^2(kR)|_{R=R(M)}, \quad (12)$$

and the comoving smoothing scale R is related to the halo mass by $M = 4\pi\rho_m R^3/3$. We use a real space top-hat window function $W(x) = 3(\sin x - x\cos x)/x^3$.

In Fig. 2 we show the halo mass function corresponding to the modified power spectra shown in Fig. 1. We see that the deviation of the halo mass function from the standard ΛCDM prediction at the relevant scales depends mostly on the scale k_c and less on the slope n. This indicates that the exact shape of the modification in $\mathcal{P}$ has a relatively minor impact on our results and that the two-parameter model (1) is sufficient to capture the relevant qualitative features.

[3]For a review on constraints on PBHs, see e.g. [44].

[4]This growth factor is for adiabatic fluctuations. While it is slightly different for isocurvature fluctuations [38], we find that this does not make a significant difference in our results.

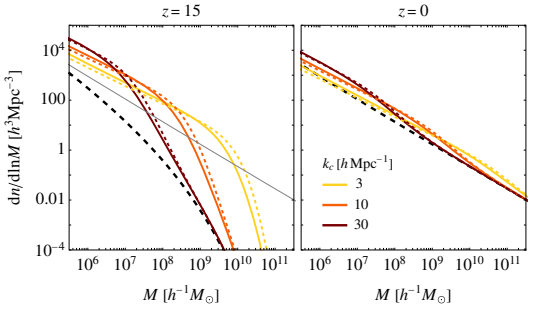


FIG. 2. The halo mass function for the matter power spectra shown in Fig. 1 at redshifts $z = 15$ and $z = 0$. The thin line shows the CDM halo mass function at $z = 0$.

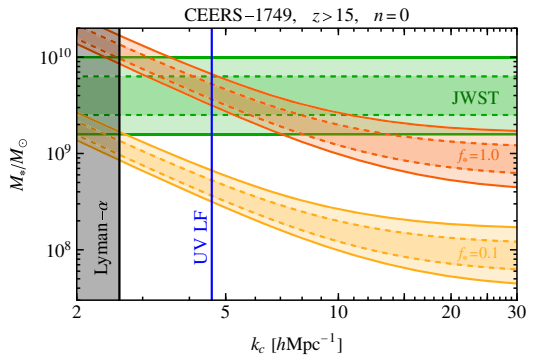


FIG. 3. The yellow and red bands show the 1σ (dashed) and 2σ (solid) stellar mass EVS confidence intervals at $z > 15$ for star formation efficiencies $f_* = 0.1$ (yellow) and $f_* = 1.0$ (red) and the spectral index $n = 0$. The gray region is excluded at 2σ by the constraints on the matter power spectrum from Lyman-α clustering [23] and the blue line shows the conservative 2σ constraint from the UV luminosity function [24]. The green band indicates the stellar mass estimate of CEERS-1749 at $z > 15$ from [2].

Moreover, we find that the halo mass function at the relevant mass scales does not significantly depend on the cutoff scale of the power-law part as long as $k_{\rm cut} \gtrsim 30h$ Mpc^{-1}.

IV. ANALYSIS AND RESULTS

The expected number of halos in the mass range $M_{\rm min} < M < M_{\rm max}$ observed in the redshift range $z_{\rm min} < z < z_{\rm max}$ is given by the light-cone integral

$$N = f_{\rm sky} \int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V_c}{{\rm d}z} \int_{M_{\rm min}}^{M_{\rm max}} {\rm d}M \frac{{\rm d}n}{{\rm d}M}, \tag{13}$$

where $f_{\rm sky}$ is the fraction of the sky being observed and V_c is the comoving Hubble volume at redshift z. For the JWST observations, the relevant quantity is the stellar mass of the halo, related to the halo mass M via [12]

$$M_* = f_* f_b M, \tag{14}$$

where $0 < f_* < 1$ is the star formation efficiency and $f_b \approx \Omega_b/\Omega_m \approx 0.16$ is the fraction of all matter in baryons. For example, taking the CDM halo mass function and a typical optimistic efficiency according to lower redshift observations $f_* = 0.1$ [63], we find that the expected number of halos similar to CEERS-1749 whose stellar mass is $M_* > 10^{9.6} M_\odot$ at redshift $z > 15$ within $f_{\rm sky} = 40$ arcmin^2 is $N \approx 7 \times 10^{-8}$.

To compute the distribution of the heaviest halo expected to be seen in an observed light-cone volume we utilize EVS [64,65]. The EVS probability density function (PDF) in the full mass range $0 < M < \infty$ is [64]

$$\Phi(M) = N f(M) F(M)^{N-1}, \tag{15}$$

where

$$f(M) \propto \int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V_c}{{\rm d}z} \frac{{\rm d}n}{{\rm d}M}, \qquad F(M) = \int_0^M {\rm d}M' f(M') \tag{16}$$

are the normalized PDFs of halo masses within the redshift range $z_{\rm min} < z < z_{\rm max}$ and the corresponding cumulative distribution.

In Fig. 3, we show the 1σ and 2σ confidence intervals of the EVS PDF as a function of k_c for $f_{\rm sky} = 2.7 \times 10^{-7}$ and $z > 15$. As expected, the distribution prefers larger stellar masses at smaller k_c. Moreover, the mean of the EVS distribution deviates by less than 10% from the M_* value corresponding to $N(M > M_*/f_* f_b, z > 15) = 1$ within $f_{\rm sky} = 2.7 \times 10^{-7}$.[5] The green shaded region indicates the stellar mass of CEERS-1749, the gray shaded region is excluded by the Lyman-α constraints on the matter power spectrum and the blue line indicates the conservative constraint from the UV luminosity function.[6] We see that with the commonly used efficiency $f_* = 0.1$ the CEERS-1749 observation is in more than 2σ tension with the model even at the smallest allowed value of k_c. With the maximal efficiency $f_* = 1.0$ a good agreement is evident if $k_c \lesssim 10h$ kpc^{-1}.

[5]Even though the widely used condition $N = 1$ provides a reasonably accurate mass scale it does not provide similarly evident treatment for the accompanying stochasticity. This is the main reason to use EVS, which equips us with a complete statistical framework.

[6]The dominant UV luminosity function constraint $a_s = 0.66^{+0.43}_{-0.17}$ in Ref. [24] is given for the relative enhancement over the ΛCDM in the bin 0.5 Mpc$^{-1} \leq k <$ 2.25 Mpc^{-1}. As the constraint depends nontrivially on the shape of the power spectrum and has not been computed for the parametrization (1), we estimate the corresponding constraint on k_c by comparing the amplitude relative to the ΛCDM at the lower edge of the bin, $k = 0.5$ Mpc^{-1}.

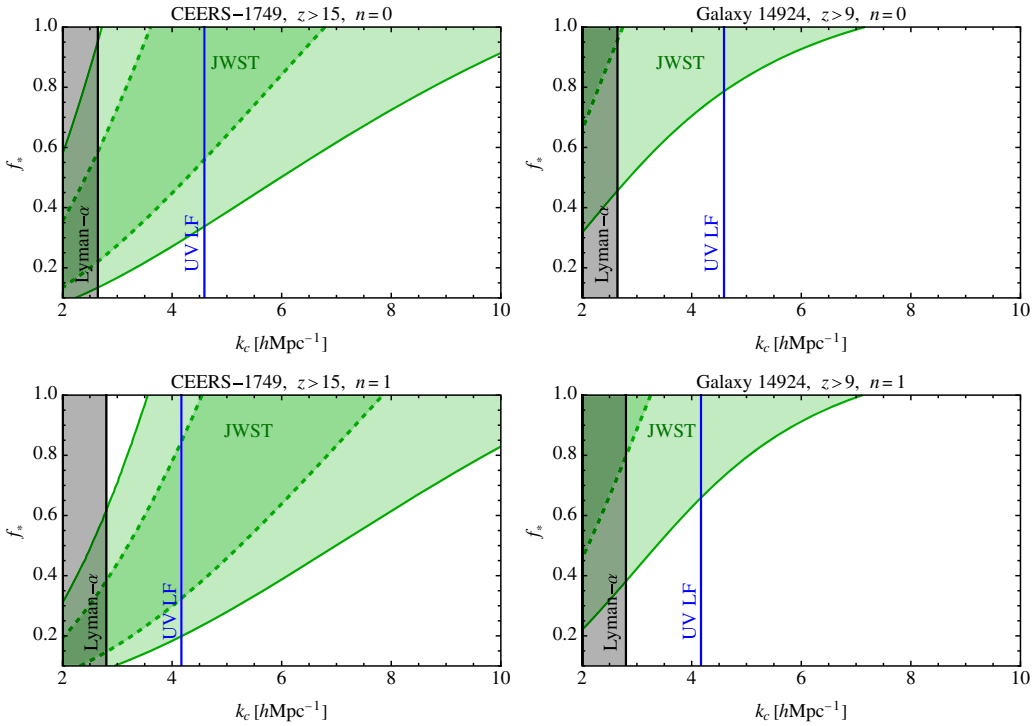


FIG. 4. The green bands indicate the 1σ and 2σ confidence intervals of the likelihood averaged over the EVS PDF. The left panels are for the stellar mass estimate of CEERS-1749 at $z > 15$ and the right panels are for the stellar mass estimate of Galaxy 14924 at $z > 9$.

The final results of our analysis are summarized in Fig. 4. The green bands show the 1σ and 2σ ranges in k_c computed from the likelihood

$$\int \mathrm{d}M\Phi(M)p(M|M_c), \tag{17}$$

assuming a log-normal distribution, $p(M|M_c) = \exp[-\ln(M/M_c)/2\sigma^2]/\sqrt{2\pi\sigma^2}$, for the measured mass M_c. The left panels show these ranges for CEERS-1749 and the right panels for Galaxy 14924. We find that a significantly enhanced star formation efficiency and a modification over the CDM matter power spectrum at scales $k_c = \mathcal{O}(10)h$ Mpc^{-1} would be required to alleviate the tension below the 2σ level. A high star formation efficiency may lead to tensions with the CMB measurements of reionization history [14]. The lowest f_*, still compatible with both the Lyman-α and UV luminosity function constraints at the 2σ confidence level, is obtained for $k_c = 5h$ Mpc^{-1}. This can be reached if

(1) DM consists of axions of mass $m_a \approx 2 \times 10^{-18}$ eV and the Peccei-Quinn symmetry is broken after inflation so that axion miniclusters are formed. The mass range 10^{-19} eV $\lesssim m_a \lesssim 10^{-16}$ eV is, however, potentially excluded by black hole superradiance[7] and the observed spins of superheavy black holes [67–69]. For $m_a > 10^{-16}$ eV we get $k_c > 50h$ Mpc^{-1}, which does not affect the halo mass function at the relevant scales.

(2) A fraction $f_{\mathrm{PBH}} > 0.005$ of DM consists of PBHs or PBH clusters of mass $m_{\mathrm{PBH}} \approx 4 \times 10^6 M_\odot (f_{\mathrm{PBH}}/0.005)^{-1}$. However, accretion constraints [43] exclude the explanation from the Poisson effect from uniformly distributed individual PBHs, but the explanation with heavy clusters of subsolar mass PBHs may be viable.

(3) The curvature power spectrum grows steeply at $4h$ Mpc^{-1} and the growth terminates before the COBE/Firas bound. This case is possible though not well motivated. In particular, it is nontrivial to realize such growth at those scales while simultaneously remaining in agreement with the CMB constraints.

[7]The constraints can be alleviated in self-interacting scenario where the superradiant growth of the axionic cloud may be suppressed [66].

V. CONCLUSIONS

We have performed a careful analysis of whether a modified matter power spectrum could explain the existence of two of the most extreme galaxies observed so far by JWST, CEERS-1749 and Galaxy 14924. Considering power-law modifications to the matter power spectrum and extreme value statistics, we have estimated the distribution of the heaviest galaxies expected to be seen by JWST above a given redshift. We have found that both a high star formation efficiency and a modification of the matter power spectrum at the smallest scales allowed by the Lyman-α and UV luminosity function constraints are needed for this distribution to be compatible with the observed stellar mass, in particular for Galaxy 14924.

We have shown that extreme stellar masses consistent with the JWST observations can be obtained in cosmologies containing axion miniclusters, but the relevant range of axion masses is potentially excluded by superradiance constraints. Explanations relying on isocurvature perturbations from heavy Poisson-distributed PBHs are instead in conflict with existing accretion constraints on PBH abundance. However, such conflicts can be resolved in PBH scenarios with initial spatial clustering.

ACKNOWLEDGMENTS

We thank Diego Blas, Julian Muñoz and Nashwan Sabti for useful discussions on the UV luminosity function. This work was supported by European Regional Development Fund through the CoE program grant No. TK133 and by the Estonian Research Council grant No. PRG803. The work of V. V. has been partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736.

[1] James Webb Space Telescope, https://webb.nasa.gov.
[2] R. P. Naidu *et al.*, arXiv:2208.02794.
[3] I. Labbe *et al.*, arXiv:2207.12446.
[4] H. Atek *et al.*, Mon. Not. R. Astron. Soc. **519**, 1201 (2023).
[5] S. L. Finkelstein *et al.*, Astrophys. J. Lett. **940**, L55 (2022).
[6] Y. Harikane *et al.*, arXiv:2208.01612.
[7] L. D. Bradley *et al.*, arXiv:2210.01777.
[8] Hubble Space Telescope, https://hubblesite.org.
[9] A. Ferrara, A. Pallottini, and P. Dayal, arXiv:2208.00720.
[10] F. Ziparo, A. Ferrara, L. Sommovigo, and M. Kohandel, arXiv:2209.06840.
[11] J. Mirocha and S. R. Furlanetto, Mon. Not. R. Astron. Soc. **519**, 843 (2023).
[12] M. Boylan-Kolchin, arXiv:2208.01611.
[13] C. C. Lovell, I. Harrison, Y. Harikane, S. Tacchella, and S. M. Wilkins, Mon. Not. R. Astron. Soc. **518**, 2511 (2022).
[14] Y. Gong, B. Yue, Y. Cao, and X. Chen, arXiv:2209.13757.
[15] S. R. Furlanetto and J. Mirocha, arXiv:2208.12828.
[16] R. Endsley *et al.*, arXiv:2208.14999.
[17] B. Liu and V. Bromm, Astrophys. J. Lett. **937**, L30 (2022).
[18] M. Biagetti, G. Franciolini, and A. Riotto, arXiv:2210.04812.
[19] N. Menci, M. Castellano, P. Santini, E. Merlin, A. Fontana, and F. Shankar, Astrophys. J. Lett. **938**, L5 (2022).
[20] E. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).
[21] S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer-Verlag, London, 2001).
[22] B. A. Reid *et al.*, Mon. Not. R. Astron. Soc. **404**, 60 (2010).
[23] S. Chabanier, M. Millea, and N. Palanque-Delabrouille, Mon. Not. R. Astron. Soc. **489**, 2247 (2019).
[24] N. Sabti, J. B. Muñoz, and D. Blas, Astrophys. J. Lett. **928**, L20 (2022).
[25] N. Sabti, J. B. Muñoz, and D. Blas, Phys. Rev. D **105**, 043518 (2022).
[26] C. T. Donnan *et al.*, Mon. Not. R. Astron. Soc. **518**, 6011 (2023).
[27] D. J. Eisenstein and W. Hu, Astrophys. J. **496**, 605 (1998).
[28] J. Chluba and D. Grin, Mon. Not. R. Astron. Soc. **434**, 1619 (2013).
[29] E. W. Kolb and I. I. Tkachev, Phys. Rev. Lett. **71**, 3051 (1993).
[30] M. Fairbairn, D. J. E. Marsh, J. Quevillon, and S. Rozier, Phys. Rev. D **97**, 083502 (2018).
[31] M. Feix, J. Frank, A. Pargner, R. Reischke, B. M. Schäfer, and T. Schwetz, J. Cosmol. Astropart. Phys. 05 (2019) 021.
[32] V. Iršič, H. Xiao, and M. McQuinn, Phys. Rev. D **101**, 123518 (2020).
[33] C. A. J. O'Hare, G. Pierobon, J. Redondo, and Y. Y. Y. Wong, Phys. Rev. D **105**, 055025 (2022).
[34] D. Ellis, D. J. E. Marsh, B. Eggemeier, J. Niemeyer, J. Redondo, and K. Dolag, Phys. Rev. D **106**, 103514 (2022).
[35] P. Meszaros, Astron. Astrophys. **38**, 5 (1975).
[36] B. Carr and J. Silk, Mon. Not. R. Astron. Soc. **478**, 3756 (2018).
[37] V. De Luca, G. Franciolini, A. Riotto, and H. Veermäe, Phys. Rev. Lett. **129**, 191302 (2022).
[38] D. Inman and Y. Ali-Haïmoud, Phys. Rev. D **100**, 083528 (2019).
[39] V. De Luca, V. Desjacques, G. Franciolini, and A. Riotto, J. Cosmol. Astropart. Phys. 11 (2020) 028.
[40] K. Freese, R. Price, and D. N. Schramm, Astrophys. J. **275**, 405 (1983).
[41] B. J. Carr and J. Silk, Astrophys. J. **268**, 1 (1983).
[42] N. Cappelluti, G. Hasinger, and P. Natarajan, Astrophys. J. **926**, 205 (2022).
[43] P. D. Serpico, V. Poulin, D. Inman, and K. Kohri, Phys. Rev. Res. **2**, 023204 (2020).
[44] B. Carr, K. Kohri, Y. Sendouda, and J. Yokoyama, Rep. Prog. Phys. **84**, 116902 (2021).

[45] M. Y. Khlopov, R. V. Konoplich, S. G. Rubin, and A. S. Sakharov, arXiv:hep-ph/9807343.
[46] M. Y. Khlopov, S. G. Rubin, and A. S. Sakharov, Astropart. Phys. **23**, 265 (2005).
[47] M. M. Flores and A. Kusenko, Phys. Rev. Lett. **126**, 041101 (2021).
[48] T. H. Jung and T. Okui, arXiv:2110.04271.
[49] V. De Luca, G. Franciolini, and A. Riotto, Phys. Rev. D **104**, 123539 (2021).
[50] J. Silk and M. J. Rees, Astron. Astrophys. **331**, L1 (1998).
[51] J. Garcia-Bellido and E. Ruiz Morales, Phys. Dark Universe **18**, 47 (2017).
[52] K. Kannike, L. Marzola, M. Raidal, and H. Veermäe, J. Cosmol. Astropart. Phys. 09 (2017) 020.
[53] C. Germani and T. Prokopec, Phys. Dark Universe **18**, 6 (2017).
[54] S. Hawking, Mon. Not. R. Astron. Soc. **152**, 75 (1971).
[55] B. J. Carr and S. W. Hawking, Mon. Not. R. Astron. Soc. **168**, 399 (1974).
[56] A. Karam, N. Koivunen, E. Tomberg, V. Vaskonen, and H. Veermäe, arXiv:2205.13540.
[57] C. T. Byrnes, P. S. Cole, and S. P. Patil, J. Cosmol. Astropart. Phys. 06 (2019) 028.
[58] G. Tasinato, Phys. Rev. D **103**, 023535 (2021).
[59] S. Dodelson, *Modern Cosmology* (Academic Press, Amsterdam, 2003).
[60] W. H. Press and P. Schechter, Astrophys. J. **187**, 425 (1974).
[61] R. K. Sheth and G. Tormen, Mon. Not. R. Astron. Soc. **308**, 119 (1999).
[62] A. Cooray and R. K. Sheth, Phys. Rep. **372**, 1 (2002).
[63] P. S. Behroozi, R. H. Wechsler, and C. Conroy, Astrophys. J. **770**, 57 (2013).
[64] I. Harrison and P. Coles, Mon. Not. R. Astron. Soc. **418**, L20 (2011).
[65] I. Harrison and P. Coles, Mon. Not. R. Astron. Soc. **421**, L19 (2012).
[66] M. Baryakhtar, M. Galanis, R. Lasenby, and O. Simon, Phys. Rev. D **103**, 095019 (2021).
[67] M. J. Stott and D. J. E. Marsh, Phys. Rev. D **98**, 083006 (2018).
[68] C. Ünal, F. Pacucci, and A. Loeb, J. Cosmol. Astropart. Phys. 05 (2021) 007.
[69] *The Search for Ultralight Bosonic Dark Matter*, edited by D. F. J. Kimball and K. van Bibber (Springer, Cham, 2023), 10.1007/978-3-030-95852-7.

Appendix 5

V
J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. Starlight from JWST: Implications for star formation and dark matter models. *Astron. Astrophys.*, 702:A109, 2025

A&A, 702, A109 (2025)
https://doi.org/10.1051/0004-6361/202555390
© The Authors 2025

Astronomy & Astrophysics

Starlight from JWST: Implications for star formation and dark matter models

Juan Urrutia[1,2,⋆], John Ellis[3,4], Malcolm Fairbairn[3], and Ville Vaskonen[1,5,6]

[1] Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia
[2] Department of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia
[3] King's College London, Strand, London, WC2R 2LS, United Kingdom
[4] Theoretical Physics Department, CERN, Geneva, Switzerland
[5] Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy
[6] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy

Received 5 May 2025 / Accepted 12 August 2025

ABSTRACT

We compared the star formation rate in different dark matter (DM) models with UV luminosity data from JWST up to $z \simeq 25$ and legacy data from HST. We find that a transition from a Salpeter population to top-heavy Pop III stars is likely at $z \simeq 10$, and that beyond $z = 10-15$ the feedback from supernovae and active galactic nuclei is progressively reduced, so that at $z \simeq 25$ the production of stars is almost free from any feedback. We compared fuzzy and warm DM models that suppress small-scale structures with the CDM paradigm, and find that the fuzzy DM mass $>5.6 \times 10^{-22}$ eV and the warm DM mass >1.5 keV at a 95% CL. The fits of the star formation rate parametrisation do not depend strongly on the DM properties within the allowed range. We find no preference over CDM for enhanced matter perturbations associated with axion mini-clusters or primordial black holes. The scale of the enhancement of the power spectrum should be $>25\,\mathrm{Mpc}^{-1}$ at the 95% CL, excluding axion mini-clusters produced for $m_a < 6.6 \times 10^{-17}$ eV or heavy primordial black holes that constitute a fraction $f_{\rm PBH} > \max[105\,M_\odot/m_{\rm PBH}, 10^{-4}(m_{\rm PBH}/10^4\,M_\odot)^{-0.09}]$ of DM.

Key words. cosmology: theory – dark matter

1. Introduction

An important target for James Webb Space Telescope (JWST) observations is to measure the UV luminosity function at high z (Naidu et al. 2022b; Castellano 2023; Donnan et al. 2023; Morishita & Stiavelli 2023; Bouwens et al. 2023; Pérez-González et al. 2023; Harikane et al. 2023, 2024; Pérez-González et al. 2025; Castellano et al. 2025). These observations complement the more precise lower-z observations from the Hubble Space Telescope (HST) (McLeod et al. 2016; Oesch et al. 2018; Bouwens et al. 2021; Finkelstein et al. 2022; Leethochawalit et al. 2023; Bagley et al. 2024). The high-z observations serve as a probe of galaxy dynamics and the star formation rate (SFR) at early epochs of the universe, constraining the abundance and masses of the conjectured first-generation Pop III stars (Maiolino 2024; Vanzella et al. 2023; Ventura et al. 2024; Fujimoto et al. 2025). The frontier of UV luminosity function measurements has been pushed back to unprecedented early times with recent measurements at $z \sim 25$, opening a window on the epoch before reionisation took place (Pérez-González et al. 2025; Castellano et al. 2025).

The number density of Pop III stars depends on the abundance of early galaxies. This depends in turn on the nature of dark matter (DM) since galaxies populate DM halos. The UV luminosity function is therefore a valuable source of insights into DM models (Sabti et al. 2022b; Hütsi et al. 2023; Parashari & Laha 2023; Liu et al. 2024; Winch et al. 2024; Sipple et al. 2025). As the redshift increases, the halo mass function (HMF) becomes more sensitive to the physical properties of DM since fewer halos are built via hierarchical growth and the imprints from the physics before the CMB are more noticeable.

Our understanding of structure formation and the star formation process in models of DM with suppressed small-scale structure formation at high $z \gtrsim 10$, such as fuzzy dark matter (FDM) and warm dark matter (WDM), or in models with an enhanced matter power spectrum is very limited. Therefore, the very high redshift abundance and distribution of Pop III stars, made of pristine gas from the Big Bang, are topics of hot debate. The roles of external radiation, internal fragmentation of the gas clouds and the dynamics that form the stars are highly uncertain, so the environments where these first stars were born and their mass function are as yet unknown (see Klessen & Glover 2023 for a review). The UV luminosity function measured with JWST has proven to be higher than anticipated, even when considering only the galaxies that are spectroscopically confirmed (Harikane et al. 2024). This suggests either that the feedback mechanisms preventing star formation were less restrictive (Susa & Umemura 2004; Dekel et al. 2023; Fukushima & Yajima 2021) or that the Pop III stars were very top-heavy, and therefore more luminous (Hirano et al. 2015; Steinhardt et al. 2023; Hirano et al. 2014; Chakraborty & Choudhury 2024). This tension has been exacerbated by the measurements of starlight at $z \sim 25$, which are higher than the predictions from many semi-analytical models in the literature (Pérez-González et al. 2025).

The objectives of this work are to parametrise the essential physics of high-z star formation and to understand the features necessary to reproduce the UV luminosity function back to $z \sim 25$ in different DM models. We took into account possible degeneracies and features of modified DM models, feedback,

⋆ Corresponding author: juan.urrutia@kbfi.ee

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

and changes in the stellar population. To this end, we connected the HMF obtained using the excursion set formalism (Bond et al. 1991), with the early Universe microphysics of DM, and combined this with a phenomenological parametrisation of star formation. We also discuss the effects of gravitational lensing and dust attenuation, neither of which seems to be very significant for the interpretation of the $z > 15$ UV data.

To extrapolate the results of DM simulations from low to high z and to obtain reliable bounds, we improved on existing techniques in the literature in two key aspects. First, we derived a new estimate of the DM halo growth rate that accounts for the ellipsoidal collapse correction, and used it consistently for different DM models. To this end, we recomputed the first-crossing probability distributions of the paths with a moving barrier and found a better fit than that originally proposed by Sheth et al. (2001). The conditional mass function in the Sheth-Tormen approach is divergent for very small steps, as has been pointed out by Jiang & van den Bosch (2014) and Zhang et al. (2008). Our fit is divergence-free and reproduces analytical estimations of Zhang et al. (2008). Second, we provide new values for the parameters the smooth-k window function that, unlike those proposed in Leo et al. (2018), consistently match the CDM prediction.

We find that, under the assumption that the star formation rate (SFR) is proportional to the halo growth rate, the observations of the UV luminosity function up to high z indicate two interesting effects that change the luminosity function. First, at $z > 10$, the stars need to be more luminous and/or the SFR needs to be higher. This transition is relatively sharp and might be caused by a top-heavy Pop III-dominated star population. We find that the evidence for this effect is above the 95% CL. In addition, the SFR needs to be further enhanced at $z > 15$. This can be achieved by reducing the feedback from supernovae (SNe) and active galactic nuclei (AGNs) on the SFR. We find that to explain the $z > 15$ observations, this feedback needs to be significantly reduced to match the UV luminosity function, and at $z \simeq 25$ the production of stars should be almost free from feedback effects. In agreement with Yung et al. (2025), with these changes in the stellar population and feedback effects, we find that the JWST data can be explained within the CDM model.

An important focus of this work is whether different DM scenarios require different star formation physics to fit the high-redshift data. Interestingly, we find very little sensitivity to DM properties that differ from CDM, with only a small change in star formation rates favoured for WDM, and no change for FDM.

When marginalizing over the SFR parametrisation, we find a lower bound of $m_{\rm FDM} > 5.6 \times 10^{-22}$eV for the case of FDM and $m_{\rm WDM} > 1.5$ keV for the case of WDM. The new JWST data push the observations to much earlier galaxies, and yet they are still not precise enough to put much stronger bounds than the HST data, as also found in Winch et al. (2024). We also find that enhanced small-scale matter perturbations are not preferred over changes in the star formation, due to the presence of Pop III stars and changes in the feedback in the pre-reionisation era. We are able to put bounds on the scale of white-noise enhanced perturbations at $k_c > 25\,{\rm Mpc}^{-1}$, excluding, for example, the possibility of axion mini-clusters produced for $m_a < 6.6 \times 10^{-17}$ eV or heavy primordial black holes (PBHs) that constitute the fraction $f_{\rm PBH} > \max[105\,M_\odot/m_{\rm PBH}, 10^{-4}(m_{\rm PBH}/10^4\,M_\odot)^{-0.09}]$ of DM. In a forthcoming paper, we will leverage these findings to compute the growth of SMBHs in non-CDM scenarios and obtain, for the first time, even more competitive bounds on deviations from CDM using the measurements of high-z SMBHs by JWST.

Throughout this work, we use the Planck 2018 CMB best-fit values for the cosmological parameters (Aghanim et al. 2020): $\Omega_{\rm M} = 0.315$, $\Omega_{\rm B} = 0.0493$, $z_{\rm eq} = 3402.0$, $\sigma_8 = 0.811$, $h = 0.674$, $T_0 = 2.7255$ K, and $n_s = 0.965$.

2. Halo mass function and growth rate

In the excursion-set formalism (Bond et al. 1991), the conditional and unconditional HMFs are obtained from the first threshold crossing distributions of uncorrelated random walks in the (S, δ) plane, where $S = \sigma(M)^2$ denotes the variance of matter perturbations at mass scale M, and δ is the matter density contrast around a given spatial point. The unconditional mass function is obtained from the first-crossing distribution $p_{\rm FC}(S, z)$ of walks starting from $(0, 0)$

$$\frac{{\rm d}n(z)}{{\rm d}M} = \frac{\rho}{M}\left|\frac{{\rm d}S}{{\rm d}M}\right| p_{\rm FC}(S, z). \tag{1}$$

For spherical collapse in a universe with CDM, the density contrast threshold is $\delta_{\rm sp}(z) = 3(3\pi/2)^{2/3}/(5D_{\rm g}(z))$ (Dodelson 2003), where the CDM linear growth function $D_{\rm g}(z)$ does not depend on S and the first-crossing distribution $p_{\rm FC}(S, z)$ can be computed analytically. In the extension to include ellipsoidal collapse, the threshold becomes S-dependent

$$\delta_{\rm ell}(S, z) = \sqrt{a}\,\delta_{\rm sp}(z)\left[1 + 0.485\left(a\frac{\delta_{\rm sp}(z)^2}{S}\right)^{-0.615}\right], \tag{2}$$

with $a = 0.707$ (Sheth et al. 2001), and $p_{\rm FC}(S, z)$ cannot be calculated analytically. The standard Sheth-Tormen ansatz for the first-crossing distribution with ellipsoidal collapse is (Sheth & Tormen 1999)

$$p_{\rm FC}(S, z) = A\left[1 + (q\nu)^{-p}\right]\sqrt{\frac{q\nu}{2\pi}}\frac{e^{-q\nu/2}}{S},\quad \nu = \frac{\delta_{\rm sp}(z)^2}{S}, \tag{3}$$

where the normalisation factor A is given by $A = [1 + 2^{-p}\Gamma(1/2 - p)/\sqrt{\pi}]^{-1}$, $p = 0.3$, and $q = a = 0.707$[1]. In the left panel of Fig. 1 we show that random walks tend towards a slightly higher value of $q = 0.80$. The histograms shown in Fig. 1 are obtained by generating 2×10^5 random walks with step size $\Delta\nu/\nu = 10^{-4}$. The Seth-Tormen halo mass function has been found to closely match the results of numerical simulations (see e.g. Reed et al. (2003), Zheng et al. (2024)).

For the conditional mass function, we studied random walks starting from $S_0 > 0$ and $\delta_0 = \delta_{\rm ell}(S_0, z_0) > 0$. The simplest ansatz for the first-crossing distribution $p_{\rm FC}(S, z|S_0, z_0)$, as considered by Sheth & Tormen (2002), among others, is obtained by the shift $\delta_{\rm sp}(z) \to \delta_{\rm sp}(z) - \delta_{\rm sp}(z_0)$ and $S \to S - S_0$ in (3). As seen in Fig. 1, comparing the orange dashed curves with the histograms, we find that this ansatz does not reproduce well the results of the random walks. A better estimate, shown by the solid black curves in Fig. 1, is given by

$$\begin{aligned} p_{\rm FC}(S, z|S_0, z_0) &= A'\left[1 + (a\nu')^{-p}\right]\sqrt{\frac{\nu'}{2\pi}}\frac{e^{-\nu'/2}}{S - S_0}, \\ \nu &= \frac{\delta_{\rm sp}(z)^2}{S},\quad \nu' = \frac{(\delta_{\rm ell}(z, S_0) - \delta_{\rm ell}(z_0, S_0))^2}{S - S_0}, \end{aligned} \tag{4}$$

[1] The spherical collapse model corresponds to $p = 0$, $q = 1$, and $A = 1/2$.

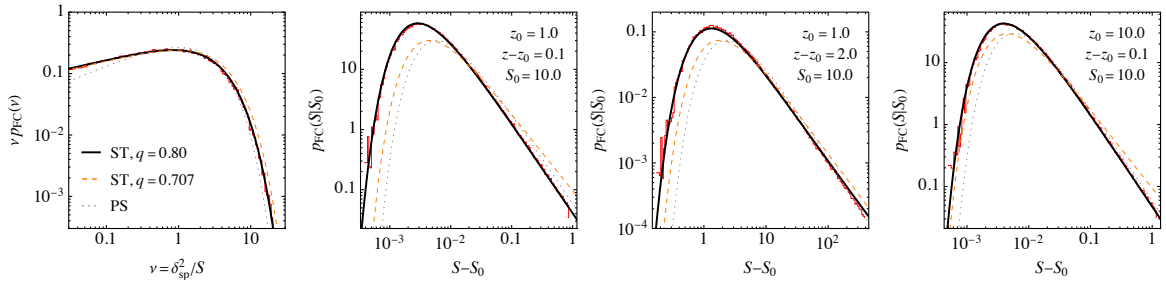


Fig. 1. *Left panel:* Unconditional first-crossing distribution for walks starting from $S = 0$, $\delta = 0$. The histogram shows the results of random walks; the solid black and dashed orange curves show the ansatz (3) with $q = 0.80$ and $q = 0.707$, respectively; and the dotted grey curve shows the spherical collapse result. *Other panels:* Conditional first-crossing distribution at $z > z_0$ for walks starting from $S_0 > 0$ and $\delta_{\rm ell}(z_0, S_0) > 0$ as indicated in the plot. The histogram again shows the results of random walks, the solid black curves show the ansatz (4), the dashed orange curves show the ansatz (3) with $\delta_{\rm sp}(z) \to \delta_{\rm sp}(z) - \delta_{\rm sp}(z_0)$ and $S \to S - S_0$, and the dotted grey curve shows the spherical collapse result.

where, as in Eq. (2), $a = 0.707$ and the normalisation factor A' in the limit $z \to z_0$ is

$$A' = \left[1 + \left(a\frac{\delta_{\rm sp}(z_0)^2}{S_0}\right)^{-p}\right]^{-1}. \tag{5}$$

In the limit $z \to z_0$, our formula (4) agrees well with the analytical estimate derived by Zhang et al. (2008). However, our expression also matches the numerical results for large redshift differences $z - z_0 \gg 0$[2].

From the first-crossing probabilities, we can construct the probability density that a halo whose mass at z' is M' ends up being a part of a halo of mass $M > M'$ at $z < z'$:

$$\frac{{\rm d}\tilde{P}(M, z|M', z')}{{\rm d}M} = p_{\rm FC}(S', z'|S, z)\frac{p_{\rm FC}(S, z)}{p_{\rm FC}(S', z')}\left|\frac{{\rm d}S}{{\rm d}M}\right|. \tag{6}$$

This, in turn, directly gives the halo mass growth rate:

$$\begin{aligned}\frac{{\rm d}M}{{\rm d}z} &= \lim_{\Delta z\to 0}\int_M^{2M} {\rm d}M' \, \frac{M' - M}{\Delta z}\frac{{\rm d}P(M', z|M, z + \Delta z)}{{\rm d}M'} \\ &= \sqrt{\frac{2}{\pi}}\frac{{\rm d}\delta_{\rm ell}(S(M), z)}{{\rm d}z}\left|\frac{{\rm d}S}{{\rm d}M}\right|^{-1}\sqrt{S(M) - S(2M)}.\end{aligned} \tag{7}$$

Unlike the estimate derived by Correa et al. (2015), our estimate (7) accounts for the correction from ellipsoidal collapse. Moreover, our estimate is based on the expected increase in the halo mass, while the estimate of Correa et al. (2015) is based on finding the expected mass of the main progenitor. However, we find that these estimates agree within a factor of two across a broad range of halo masses and redshifts.

3. Dark matter models

3.1. Suppressed small-scale structures

We considered two DM models that predict suppression of small-scale structures: 1) FDM where the DM is provided by coherent waves of an ultra-light bosonic field with mass $m_{\rm FDM} = O(10^{-20}\,{\rm eV})$ and 2) WDM where the DM is provided by semi-relativistic particles with masses $m_{\rm WDM} = O({\rm keV})$. The changes in the matter power spectrum in these cases can be described by transfer functions $T_J(k)^2$, where J labels the model. The resulting matter power spectrum is given by

$$\mathcal{P}_{\rm sup}(k) = T_J(k)^2\,\mathcal{P}_{\rm CDM}(k), \tag{8}$$

where $\mathcal{P}_{\rm CDM}(k)$ denotes the CDM matter power spectrum that we compute with the transfer function from Eisenstein & Hu (1998).

In the case of FDM, the speed of sound is modified by quantum pressure, so that $c_s \sim k/(4a^2 m_{\rm FDM}^2)$. Consequently, the Jeans scale is non-zero (see e.g. Marsh (2016)),

$$k_{\rm J} = \frac{66.5\,{\rm Mpc}^{-1}}{(1 + z_{\rm eq})^{1/4}}\left[\frac{\Omega_{\rm FDM}\,h^2}{0.12}\right]^{\frac{1}{4}}\left[\frac{m_{\rm FDM}}{10^{-22}\,{\rm eV}}\right]^{\frac{1}{2}}, \tag{9}$$

and perturbations at small scales $k \gtrsim k_{\rm J}$ are suppressed. This suppression can be estimated using an effective fluid approximation that is accurately matched to the solution of the Klein-Gordon equation (Hu et al. 2000; Passaglia & Hu 2022). We use the fitting formula provided by Passaglia & Hu (2022):

$$T_{\rm F} = \frac{\sin x^n}{x^n(1 + Bx^{6-n})}, \quad x = A\frac{k}{k_{\rm J}}, \tag{10}$$

where

$$A = 2.22 m_{22}^{1/25 + 1/1000\ln m_{22}} \text{ and } B = 0.16 m_{22}^{-1/20} \tag{11}$$

with $m_{22} \equiv m_{\rm FDM}/10^{-22}\,{\rm eV}$.

In the case of WDM, a thermal production mechanism is typically assumed, and the particles are lighter than typical CDM particles and remain relativistic for a longer time. This implies that the free streaming length of WDM particles can be long, $\lambda_{\rm fs} \sim (m_{\rm WDM}/{\rm keV})^{-4/3}{\rm Mpc}$. In order to compute the resultant damping of the small-scale structures, it is necessary to track accurately the transition between relativistic and non-relativistic behaviour by solving the Boltzmann equations. This yields the fit (Bode et al. 2001; Hansen et al. 2002; Viel et al. 2005)

$$T_{\rm W} = \left[1 + (\alpha k)^{2\mu}\right]^{-5/\mu}, \tag{12}$$

with $\mu = 1.12$ and

$$\alpha = 0.049\left[\frac{m_{\rm WDM}}{\rm keV}\right]^{-1.11}\left[\frac{\Omega_{\rm DM}}{0.25}\right]^{0.11}\left[\frac{h}{0.7}\right]^{1.22} h^{-1}{\rm Mpc}. \tag{13}$$

[2] The ellipsoidal collapse thresholds at different redshifts can intersect; however, this occurs only at very small halo masses, $M < 10^4 M_\odot$, which are not relevant for our study.

3.2. Enhanced small-scale structures

For the enhancement of small-scale structures, we also considered two cases: 1) an axion DM model that includes axion mini-clusters and 2) a scenario where heavy primordial black holes (PBHs) constitute a fraction of the DM density. In both cases, we model the enhancement of the matter power spectrum as (Hütsi et al. 2023)

$$\mathcal{P}_{\rm enh}(k) = \mathcal{P}_{\rm CDM}(k) + \mathcal{P}_{\rm CDM}(k_c). \tag{14}$$

This parametrisation includes one parameter, the scale k_c above which the power spectrum is dominated by a white noise contribution. We performed the analysis in terms of k_c, which is related to the parameters of the underlying DM models as discussed below.

In axion-like particle DM models, where the global $U(1)$ symmetry is broken after cosmic inflation, significantly small-scale density fluctuations arise due to the Kibble mechanism. These lead to the formation of axion mini-clusters (Kolb & Tkachev 1993). This scenario results in a white noise contribution to the matter power spectrum (Fairbairn et al. 2018; Feix et al. 2019; Ellis et al. 2022). The resulting power spectrum can be approximated by Eq. (14) with

$$k_c \approx 3h\,{\rm Mpc}^{-1}\left[\frac{m_a}{10^{-18}\,{\rm eV}}\right]^{0.6}, \tag{15}$$

where m_a denotes the axion-like particle mass. The white noise contribution is cut at scales that re-entered the horizon when the axion began oscillating:

$$k_{\rm cut} \simeq 300\,{\rm Mpc}^{-1}\sqrt{\frac{m_a}{10^{-18}\,{\rm eV}}}. \tag{16}$$

However, in the relevant part of the parameter space, the cut is far beyond the enhancement scale, $k_c \ll k_{\rm cut}$. We also note that the Jeans scale is larger than the cutoff scale, $k_J > k_{\rm cut}$.

In scenarios involving heavy PBHs there is a white noise contribution to the matter power spectrum (Inman & Ali-Haïmoud 2019; De Luca et al. 2020) with

$$k_c \approx 6h\,{\rm Mpc}^{-1}\left[\frac{f_{\rm PBH}m_{\rm PBH}}{10^4 M_\odot}\right]^{-0.4}, \tag{17}$$

where $f_{\rm PBH}$ is the DM fraction in PBHs, $m_{\rm PBH}$ is the PBH mass, and $\rho_{\rm DM}$ is the dark matter density. The cutoff is set by the average separation of PBHs (Hütsi et al. 2023):

$$k_{\rm cut} = 900h\,{\rm Mpc}^{-1}\left[f_{\rm PBH}\frac{10^4 M_\odot}{m_{\rm PBH}}\right]^{1/3}. \tag{18}$$

Below this scale, one expects roughly one PBH per corresponding comoving volume, and the seed effect (Carr & Silk 1983, 2018; Cappelluti et al. 2022) becomes dominant. The condition $k_c < k_{\rm cut}$ is satisfied for $f_{\rm PBH} > 10^{-4}(m_{\rm PBH}/10^4 M_\odot)^{-0.09}$.

3.3. Window function

In order to convert the matter power spectrum into the HMF, it is necessary to specify a window function to be used in conjunction with the excursion-set formalism. The window function is used to compute the variance of the matter perturbations,

$$\sigma^2(R) = A\int {\rm d}k\, k^2 P(k)\tilde{W}^2(k|R), \tag{19}$$

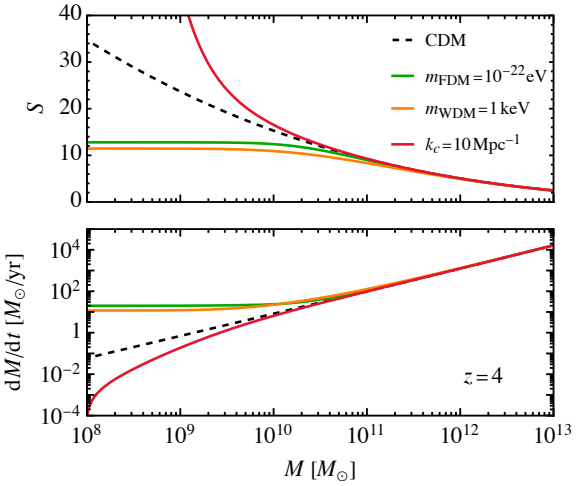


Fig. 2. Variance $S = \sigma^2$ of matter perturbations (upper panel) and halo growth rate (lower panel) in the models considered in this work: CDM (dashed black line), FDM (green line), WDM (orange line), and an enhanced matter power spectrum (red line).

where the normalisation factor A is chosen to match the Planck measurement of $\sigma_8 = \sigma(R = 8/h\,{\rm Mpc}) = 0.811$ (Aghanim et al. 2020). We used a window function of the form (Leo et al. 2018; Verwohlt et al. 2024):

$$\tilde{W}(k|R) = \frac{1}{1+(c_1 kR)^{c_2}}, \tag{20}$$

where R is related to the mass of the overdensity by $M = 4\pi\bar{\rho}R^3/3$. In the case of a suppressed matter power spectrum, the HMF scales as ${\rm d}n/{\rm d}\ln M \propto M^{c_2/6}$ and the halo growth rate as ${\rm d}M/{\rm d}z \propto M^{1-c_2/6}$ at scales far below the suppression scale. For all models, we used $c_2 = 6$ so that also for WDM and FDM models the growth rate increases with the halo mass, and fixed $c_1 = 0.43$ so that the resulting variance in CDM closely matches that computed using the standard real space top hat window function, in agreement with simulation results. We find that this choice also reproduces well the simulation results in the FDM model (Schive et al. 2016), but it underestimates the halo mass where the HMF found in WDM simulations deviates from the CDM prediction (Lovell et al. 2014). On the other hand, the choice $c_1 = 1/3.3$ suggested by Leo et al. (2018) reproduces the low-mass part, but it does not match the CDM prediction at high masses. To reproduce the simulation results in WDM and to match CDM predictions at masses much above the suppression scale, we used $c_1 = 0.43$ in the window function and multiplied α in the WDM transfer function by $3.3c_1 \approx 1.4$.

We show in the upper panel of Fig. 2 how the variance of the density perturbations in the models described above differs from that in the CDM model. For FDM and WDM, the variance asymptotes to a constant value at low masses. The break from the CDM-like behaviour at high masses is sharper for FDM than for WDM. This reflects the fact that the free streaming length does not make as abrupt a transition as the Jeans scale. For the case of enhanced perturbations, we see the opposite effect and the variance of the perturbation grows rapidly at low masses. In the lower panel of Fig. 2, we show the change in the DM halo growth rate for the different scenarios. In cases with suppressed small-scale structures, we find that the growth rate exceeds that

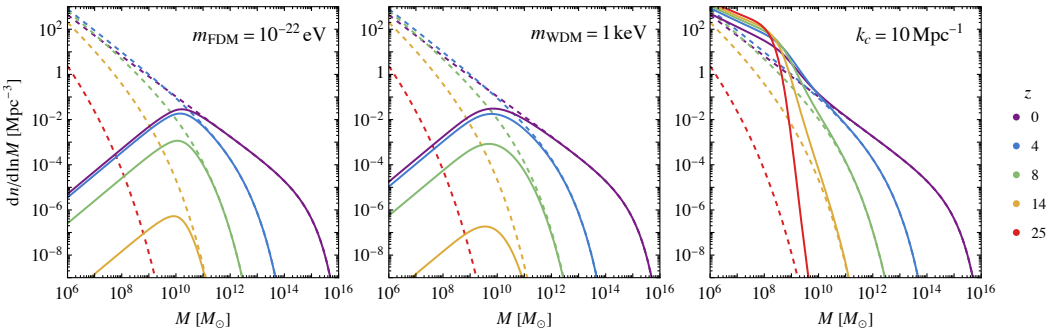


Fig. 3. HMF at various redshifts z in the FDM (left) and WDM (middle) models, and in models where white noise dominates the matter perturbations at $k > k_c$ (right). The dashed contours show the HMF in the CDM model for comparison.

of CDM, while the halos grow more slowly in the case with enhanced small-scale structures.

The HMFs are shown in Fig. 3, where they are compared with the CDM predictions. For the case with suppressed small-scale structures, the HMF grows $\propto M$ at low masses. This power-law behaviour does not significantly affect the fits to the UV luminosity observations discussed below. For the enhanced perturbations, we can see how the enhancement gets weaker at low z as the mini-clusters merge with heavier halos and their original imprint is diluted.

4. UV luminosity functions

After discussing the DM structures, which provide the potential wells that trap baryons, we now turn to the source of the UV light, namely the stars that populate galaxies. The stellar mass in a galaxy increases not only via mergers, but mainly via the formation of new stars from cold gas. At high z, the star formation rate (SFR) can be probed using the UV luminosity function, which quantifies the number density of galaxies as a function of their UV luminosity. The UV luminosity of a galaxy is directly proportional to the SFR because the sources that dominate this emission are massive $O(10\,M_\odot)$ stars that are short-lived at cosmological scales, with $\tau \lesssim 10\,{\rm Myr}$[3].

The observed UV luminosity function can be computed from the HMF as

$$\Phi_{\rm UV}(L_{\rm obs}) = \int {\rm d}L {\rm d}\ln M \frac{{\rm d}P(L_{\rm obs}|L)}{{\rm d}L_{\rm obs}} \frac{{\rm d}P(L|M)}{{\rm d}L} \frac{{\rm d}n}{{\rm d}\ln M}, \tag{21}$$

where ${\rm d}P(L|M)/{\rm d}L$ denotes the distribution of luminosities emitted by galaxies in halos of mass M and ${\rm d}P(L_{\rm obs}|L)/{\rm d}L_{\rm obs}$ is the distribution of observed luminosities if the emitted luminosity is L. The latter is affected by gravitational lensing magnification (Takahashi et al. 2011) and dust attenuation (Wang & Heckman 1996):

$$\frac{{\rm d}P(L_{\rm obs}|L)}{{\rm d}L_{\rm obs}} \approx \int {\rm d}\mu {\rm d}B \frac{{\rm d}P(\mu)}{{\rm d}\mu} \frac{{\rm d}P(B)}{{\rm d}B} \delta\Big(L_{\rm obs} - \frac{\mu}{B} L\Big). \tag{22}$$

Here ${\rm d}P(B)/{\rm d}B$ denotes the distribution of the dust attenuation B, and ${\rm d}P(\mu)/{\rm d}\mu$ denotes the distribution of the lensing magnification μ.

[3] The timescale is further reduced if the Pop III mass function is more top-heavy.

4.1. Lensing magnification

The first correction to the observed luminosity that we consider is gravitational lensing. This correction is relevant for the bright end of the UV luminosity function, which receives contributions from magnified distant galaxies lensed by intervening structures. As shown by Takahashi et al. (2011), the magnification can be approximated as[4]

$$\mu \approx (1 - \kappa)^{-2}, \tag{23}$$

where κ denotes the convergence given by the Laplacian of the lens potential. For multiple lenses, the convergence is obtained by summing the contributions from the different lenses j and subtracting the empty beam convergence $\kappa_{\rm E}$,

$$\kappa = \kappa_{\rm E} + \sum_j \kappa^{(1)}(\boldsymbol{\theta}_j), \quad \kappa^{(1)}(\boldsymbol{\theta}_j) = \frac{8\pi D_j D_{j,s}}{D_s} \Sigma_j(r_j), \tag{24}$$

where $\Sigma_j(r_j)$ denotes the projected surface mass density of the jth lens, D_j the angular diameter distance of the lens, D_s the angular diameter distance of the source, and $D_{j,s}$ the angular diameter distance between the lens and the source. The convergence of the jth lens depends on the mass of the lens M_j, the lens redshift z_j, and the distance of the lens from the line of sight to the source r_j, $\boldsymbol{\theta}_j = \{M_j, z_j, r_j\}$. As the mean convergence is $\langle\kappa\rangle = 0$, the empty beam convergence can be written as $\kappa_{\rm E} = -\langle\sum_j \kappa^{(1)}(\boldsymbol{\theta}_j)\rangle$.

We compute the distribution of μ by generating numerically the distribution of κ, converting it to ${\rm d}P(\mu)/{\rm d}\mu$ using Eq. (23) and applying a $1/\mu$ factor to transform the image-plane distribution into the source-plane distribution. We model the lenses by NFW halos whose spatial distribution follows a uniform distribution and use the halo mass-concentration/redshift relation derived by Ludlow et al. (2016). We account for the finite source sizes by averaging the surface density over a sphere of radius R_s that we obtained by projecting the source size to the lens plane. We fix the source size to 10 kpc (see in Takahashi et al. (2011), Ferrami & Wyithe (2023) how the result is affected by the source size). The distribution of lenses is given by

$$\frac{{\rm d}N(\boldsymbol{\theta})}{{\rm d}\ln M} = \int_0^{z_s} {\rm d}z_l \int_0^{r_{\rm max}} {\rm d}r \frac{2\pi(1+z_l)^2 r}{H(z_l)} \frac{{\rm d}n(z_l)}{{\rm d}\ln M}. \tag{25}$$

To limit the number of lenses N, we impose an upper limit on the distance r of the halo from the line of sight, $r < r_{\rm max}(M, z_l)$,

[4] This approximation underestimates the high-μ tail of the distribution. However, this tail is damped by the finite sizes of the lenses.

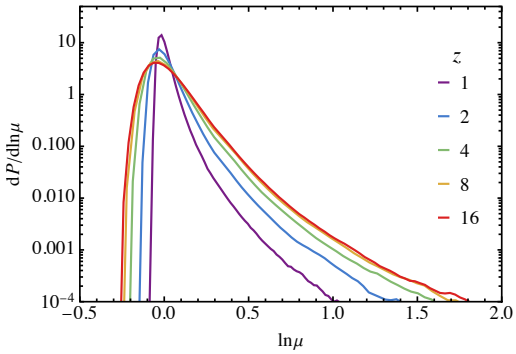


Fig. 4. Distributions of lensing magnifications at different source redshifts.

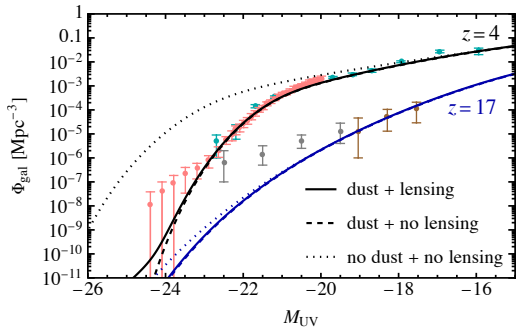


Fig. 5. UV luminosity function at $z = 4$ (black) and $z = 17$ (blue) with and without corrections due to lensing magnification and dust attenuation. The cyan and pink points show respectively the HST $z = 4$ data from Bouwens et al. (2021) and Harikane et al. (2022), and the brown and grey points the JWST $z = 17$ data from Pérez-González et al. (2025) and Castellano et al. (2025).

so that $\kappa^{(1)} > \kappa_{\rm thr}$. We checked that our choice of $\kappa_{\rm thr}$ is sufficiently small that our results are insensitive to it. The distribution of contributions $\kappa^{(1)}$ is given by

$$\begin{aligned}\frac{\mathrm{d}P^{(1)}(\kappa)}{\mathrm{d}\kappa} &= \frac{1}{N}\int \mathrm{d}N(\boldsymbol{\theta})\,\delta(\kappa-\kappa^{(1)}(\boldsymbol{\theta}))\\ &= \frac{1}{N}\int \mathrm{d}z_l\mathrm{d}r\left|\frac{\mathrm{d}\kappa^{(1)}}{\mathrm{d}\ln M}\right|^{-1}\frac{\mathrm{d}N(\boldsymbol{\theta})}{\mathrm{d}\ln M\mathrm{d}z_l\mathrm{d}r}\bigg|_{M:\,\kappa^{(1)}=\kappa}.\end{aligned} \tag{26}$$

We compute the distribution of the convergence by picking random numbers from $\mathrm{d}P^{(1)}(\kappa)/\mathrm{d}\kappa$ to generate multiple realisations of $\sum_j \kappa^{(1)}(\boldsymbol{\theta}_j)$ and shifting the resulting distribution by $\kappa_{\rm E}$ so that its mean is zero[5].

The resulting distributions of lensing magnifications at different source redshifts are shown in Fig. 4. The distributions have a characteristic long tail towards high magnifications. This amplifies the high-luminosity (low-magnitude) tail of the UV luminosity functions, as seen in Fig. 5 by comparing the dashed and solid curves. We note that the lensing amplification is dominated by relatively heavy structures, $M \gtrsim 10^{10}\,M_\odot$, and is not significantly affected by the changes in the small-scale structures that we consider in this work.

4.2. Dust attenuation

Following the literature (see e.g. Trenti et al. (2015), Vogelsberger et al. (2020)), we neglect the variance in the dust attenuation B and approximate $\mathrm{d}P(B)/\mathrm{d}B \approx \delta(B-\bar{B})$. This gives

$$\Phi_{\rm UV}(L_{\rm obs}) = \int \mathrm{d}\mu\,\mathrm{d}\ln M\frac{\bar{B}}{\mu}\frac{\mathrm{d}P(\mu)}{\mathrm{d}\mu}\frac{\mathrm{d}P(L|M)}{\mathrm{d}L}\frac{\mathrm{d}n}{\mathrm{d}\ln M}, \tag{27}$$

where $L = \bar{B}L_{\rm obs}/\mu$.

The UV luminosity functions are commonly presented in terms of the absolute magnitude instead of the luminosity. The associated absolute magnitude $M_{\rm UV}$ is defined by (Oke & Gunn 1983)

$$\log_{10}\left[\frac{L_{\rm UV}}{\mathrm{erg\,s^{-1}}}\right] = 0.4\,(51.63 - M_{\rm UV}), \tag{28}$$

[5] A similar approach of finding the distribution of a sum of random variables was introduced by Ellis et al. (2024) for the computation of the gravitational wave background from a population of supermassive black hole binaries.

where the normalisation is chosen to match the luminosity of the star Vega. Dust attenuation shifts the absolute magnitude by $A_{\rm UV} = 2.5\log_{10}\bar{B}$, which we estimate as

$$A_{\rm UV} = \frac{1}{s}\ln\left[e^{s(C_0+0.2\ln 10(C_1\sigma_\beta)^2+C_1\bar{\beta}_{\rm UV})}+1\right], \tag{29}$$

with $s = 3$, $C_0 = 4.4$, $C_1 = 2.0$, $\sigma_\beta = 0.34$, and

$$\bar{\beta}_{\rm UV} = -\exp\left[\frac{0.17(19.5+M_{\rm UV})}{1.54+0.075z}\right](1.54+0.075z). \tag{30}$$

The last is obtained by fitting the results shown in Table 3 of Bouwens et al. (2014), assuming that $\bar{\beta}_{\rm UV}$ scales exponentially with $M_{\rm UV}$. We note that, in Eq. (29), instead a piecewise function, we use a softplus function with sharpness parameter $s = 3$ chosen so that the resulting dust attenuation roughly matches that obtained by numerical sampling shown in Fig. 1 of Vogelsberger et al. (2020).

In terms of absolute magnitude, the UV luminosity function is given by

$$\begin{aligned}\Phi_{\rm UV}(M_{\rm obs}) &= \Phi_{\rm UV}(L_{\rm obs})\frac{\mathrm{d}L_{\rm obs}}{\mathrm{d}M_{\rm obs}}\\ &= \int \mathrm{d}\mu\frac{\mathrm{d}P(\mu)}{\mathrm{d}\mu}\int \mathrm{d}\ln M\frac{\mathrm{d}n}{\mathrm{d}\ln M}\frac{\mathrm{d}P(M_{\rm UV}|M)}{\mathrm{d}M_{\rm UV}},\end{aligned} \tag{31}$$

where $M_{\rm UV} = M_{\rm obs} - A_{\rm UV} + 1.086\ln\mu$. Comparing the short- and long-dashed curves in Fig. 5, we see that dust attenuation suppresses strongly the high-luminosity tail of the luminosity function at low redshifts, but at very high redshifts its effect is almost negligible. However, we note that the dust attenuation is calibrated to observations at $z \lesssim 8$ (Bouwens et al. 2014) and we have simply extrapolated these results to $z \gg 8$.

4.3. Star formation rate

We assume that the emitted UV magnitudes $M_{\rm UV}$ in a given halo of mass M follow a Gaussian distribution:

$$\frac{\mathrm{d}P(M_{\rm UV}|M)}{\mathrm{d}M_{\rm UV}} = \frac{1}{\sqrt{2\pi}\sigma_{\rm UV}}\exp\left[-\frac{(M_{\rm UV}-\bar{M}_{\rm UV})^2}{2\sigma_{\rm UV}^2}\right]. \tag{32}$$

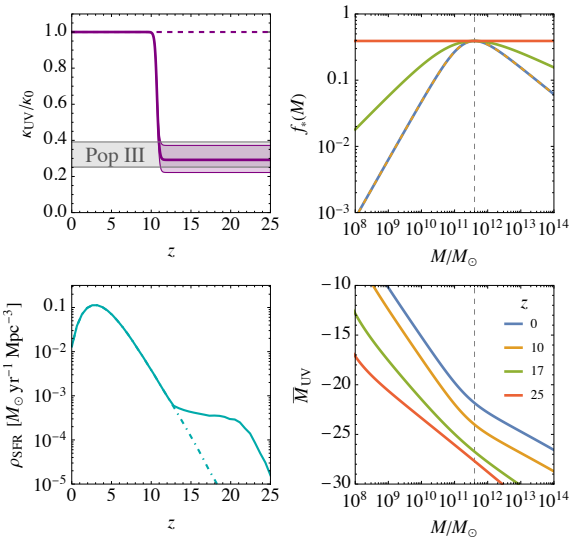


Fig. 6. *Upper left:* Best-fit UV conversion factor (35) as a function of redshift (thick purple curve) and the 68% range in f_κ (purple band). *Upper right:* Best-fit SFR (34) as a function of halo mass at different redshifts. *Lower left:* SFR density with (solid) and without (dot-dashed) the suppression (36) of the feedback effects. *Lower right:* Best-fit mean UV magnitude as a function of the halo mass at different redshifts.

The scatter in the $M_{\rm UV} - M$ relation, $\sigma_{\rm UV}$, is taken as a free parameter. The mean emitted UV luminosity, $\bar{M}_{\rm UV}$, is directly proportional to the SFR:

$$\bar{L}_{\rm UV} = \frac{\dot{M}_*}{\kappa_{\rm UV}}. \tag{33}$$

As mentioned in the Introduction, we assume that the SFR is proportional to the halo growth rate (Bian et al. 2013), $\dot{M}_* = f_B f_*(M)\,\dot{M}$, where $f_B = \Omega_B/\Omega_M \approx 0.16$. We parametrise the proportionality coefficient $f_*(M)$, which models the feedback effects from SNe and AGNs, as a broken power law around $M = M_c$ with an exponential suppression in halos with $M < M_t < M_c$,

$$f_*(M) = \epsilon\,\frac{\alpha+\beta}{\beta(M/M_c)^{-\alpha} + \alpha(M/M_c)^{\beta}}\,e^{-M_t/M}, \tag{34}$$

where $\alpha, \beta, \epsilon > 0$. A similar parametrisation is used in the GALLUMI code (Sabti et al. 2022a), for example. The difference is that in our parametrisation the maximum of $f_*(M)$ is at $M = M_c$ for $M_t \ll M_c$. We find that z-independent values of α, β, ϵ, and $\kappa_{\rm UV}$ give a good fit to the observations up to $z \approx 10$, but that the data prefer higher luminosities at $z \gtrsim 10$.

To accommodate this enhancement in the luminosities of the early stars, we introduce a parametrisation in which the conversion factor $\kappa_{\rm UV}$ changes around $z = z_\kappa$,

$$\kappa_{\rm UV} = \kappa_0\left[\frac{1+f_\kappa}{2} - \frac{1-f_\kappa}{2}\tanh\left(\frac{z-z_\kappa}{\gamma}\right)\right], \tag{35}$$

where $\kappa_0 = 1.15\times10^{-22} M_\odot\,{\rm s\,erg^{-1}\,Myr^{-1}}$, $0 < f_\kappa \leq 1$, and $\gamma > 0$. The conversion factor depends on the initial stellar mass function and the value of κ_0 corresponds to the Salpeter initial mass function (Madau & Dickinson 2014). We allow for a wide range of possibilities in the transition to Pop III stars: z_κ determines the redshift at which the change of $\kappa_{\rm UV}$ starts, γ parametrises how sharp the transition is, and f_κ is the fractional change in κ, such that at $z \gg z_\kappa$, $\kappa_{\rm UV} = f_\kappa\kappa_0$. We show $\kappa_{\rm UV}$ as a function of z for the best-fit values in the CDM model in the upper left panel of Fig. 6.

In order to fit the highest-z UV data, we find that the feedback effects, which we parametrise with $f_*(M)$, need to be reduced. To allow for such a reduction, we parametrised the powers $X = \alpha, \beta$ so that they can approach zero linearly with z above some redshift $z_{\rm fb}$:

$$X_{z>z_{\rm fb}} = X\max\left[0, \frac{z_* - z}{z_* - z_{\rm fb}}\right]. \tag{36}$$

In this parametrisation, α and β become zero at $z > z_*$, and $f_*(M)$ is mass independent at $M \gg M_t$, as shown in the upper right panel of Fig. 6 for the best-fit values. This implies that in all halos with $M \gg M_t$, the baryons that the halo accretes are converted into stars with efficiency ϵ.

In total, our parametrisation of the SFR, the UV conversion factor, and the distribution of the UV magnitudes includes 11 free parameters: $\{M_t, M_c, \epsilon, \alpha, \beta, \gamma, f_\kappa, z_\kappa, z_{\rm fb}, z_*, \sigma_{\rm UV}\}$.

5. Results

We considered UV luminosity function measurements derived from both HST observations (Bouwens et al. 2021; Harikane et al. 2022) and JWST observations (Donnan et al. 2024; Pérez-González et al. 2025; Castellano et al. 2025). The latter are based on high-z photometric measurements, most of which are not spectroscopically confirmed. This can lead to errors in the estimation of redshift. Most notably, a photometric measurement of CEERS-93316 indicated that this galaxy was at $z \approx 16$ (Naidu et al. 2022a), but a later spectroscopic measurement revealed it to be at $z \approx 5$ (Arrabal Haro et al. 2023). However, the redshifts of some of the high-z galaxies have been confirmed spectroscopically by Donnan et al. (2024).

We performed a Markov chain Monte Carlo (MCMC) analysis of the models. For each scan over the model parameters, we generated eight MCMC chains, each consisting of 12000 samples of which the first 2000 are discarded as burn-in. We used the Metropolis-Hastings sampler with Gaussian proposal distributions whose widths we chose so that the acceptance rate is around 10%. The model parameter priors are given in Table 1. We assumed that the data follow a split normal distribution $\mathcal{N}(\mu, \sigma_1, \sigma_2)$, where μ denotes the mode and $\sigma_{1,2}$ the left- and right-hand side standard deviations. The likelihood is given by

$$\mathcal{L} = \prod_j \mathcal{N}(\Phi(M_{\rm UV}^{(j)}, z^{(j)}) - \Phi^{(j)}, \sigma_1^{(j)}, \sigma_2^{(j)}), \tag{37}$$

where j represents the measurements and the function $\Phi(M_{\rm UV}, z)$ is the model prediction. The code used in the analysis is available at GitHub.

5.1. Cold dark matter

The full posteriors for our CDM fit parameters are shown in blue in Fig. 7 and the best-fit values are given in Table 1. We checked that the Gelman-Rubin statistic R (Gelman & Rubin 1992) is very close to 1 for each of the parameters. This indicates good convergence of the MCMC chains. The highest value, $R \approx 1.09$, was obtained for $\log_{10}(M_t/M_\odot)$.

Most of the parameters, in particular those parametrising the broken power-law shape of $f_*(M)$, are well constrained by the

Table 1. Parameter range and best-fit values.

Parameter	Prior	Best fit
$M_t/M_\odot$	LogU(6, 10)	$10^{7.9}$
$M_c/10^{11}M_\odot$	U(3, 5)	4.0
ϵ	U(0.36, 0.42)	0.39
α	U(0.6, 1.1)	0.88
β	U(0.2, 0.6)	0.40
γ	U(0.05, 0.6)	0.27
f_κ	U(0.05, 0.7)	0.29
z_κ	U(9.8, 11.4)	10.7
$z_{\rm fb}$	U(7, 25)	12.7
z_*	U(20, 36)	22.9
$\sigma_{\rm UV}$	U(0.05, 0.2)	0.068

Notes. Prior ranges and best-fit values of the parameters in the CDM model. The parametrisation is given by Eqs. (32), (34), (35) and (36).

data, and the posteriors show only a mild negative correlation between α and β, and between α and M_c. The fits of α, β, and M_c are similar to those found by Harikane et al. (2022), but we find about a factor of two higher SFR. Compared to Sabti et al. (2022a), we fixed the cosmological parameters to the Planck CMB values, and we are able to constrain M_c and β better because the HST data we consider extend to higher luminosities.

From the posteriors of γ, f_κ, and z_κ, we see a clear (>95%) preference for a relatively rapid change in $\kappa_{\rm UV}$ around $z_\kappa = 10-11$ by factor of ∼3. The posterior of $z_{\rm fb}$ shows that the preference for a high-z change in the powers α and β is less significant, <95%. This is expected, since the change is driven by the measurements at $z = 17$ and $z = 25$ that have relatively large uncertainties and $\Phi = 0$ is within the 95% range. Moreover, we see a correlation between f_κ and $z_{\rm bf}$, indicating that the enhancement of the UV luminosity can be partly reduced by changing the powers α and β. The data prefer a narrow scatter in the UV magnitude, $\sigma_{\rm UV} < 0.15$ at the 95% CL, and give an upper bound on the scale of the exponential SFR suppression, $M_t < 10^9\,M_\odot$, at the 95% CL. This implies that the explanation for the UV excess as a consequence of a large constant scatter in the $(M_{\rm UV}, M)$ is disfavoured when compared to the Pop III star hypothesis[6]. This result should be treated with caution, however, since selection effects, which are not considered in the likelihood analysis, could modify the conclusion.

The evolution of $\kappa_{\rm UV}$ for the best fit is shown in the upper left panel of Fig. 6. The range shown in grey corresponds to estimates of the conversion factor for Pop III stars with top-heavy initial mass functions (Harikane et al. 2023). We see that the best fit prefers a sharp transition from Pop I and Pop II stars with standard Salpeter profiles up to $z \sim 10$ to a top-heavy Pop III population at $z > 10.8$. The upper right panel of Fig. 6 illustrates how the feedback effects are suppressed for the best fit. This strongly enhances the SFR density,

$$\rho_{\rm SFR} = \int {\rm d}M \frac{{\rm d}n}{{\rm d}M} \dot{M}_*(M), \tag{38}$$

at high z, as shown in the lower left panel of Fig. 6. The lower right panel of Fig. 6 shows the relation between the mean UV magnitude and the halo mass at different redhifts.

In Fig. 8 we show the best-fit UV luminosity function in the CDM model (in black) compared with the HST and JWST observations at different redshifts. The change in the tail shape at $M_{\rm UV} < -22$ is primarily driven by the lensing magnification. The slight turn seen in the last panel at $M_{\rm UV} > -18$ is due to the exponential term in (34) that becomes relevant in the shown $M_{\rm UV}$ range only at the highest z, as seen from the lower right panel of Fig. 6. The fit without the suppression of the feedback effects is shown by the dashed lines, and undershoots the measurements at $z = 17$ and 25. The dot-dashed curve shows the fit when the high-z enhancement in the luminosities is also removed. We see that this case undershoots the observed luminosity function at redshifts $z > 10$. Moreover, we see that our model does not provide a good fit of the brightest objects at $z = 17$. This may reflect a limitation of our parametrisation, but it is also possible that the luminosities of some of these objects are contaminated by AGNs (Castellano et al. 2025).

5.2. Suppressed small-scale perturbations

We also see in Fig. 8 the comparison between the UV data and the predictions with suppressed small-scale structures in the FDM (green) and WDM (orange) models. For FDM we show the case where the model parameters correspond to the CDM best fit and the FDM mass is at the 95% CL lower limit, while for WDM we show the best fit case, which is slightly better than in the CDM model. We see that the predictions are indistinguishable for $z \lesssim 10$, and up to $z \sim 15$ the suppression is at lighter halos (bigger $M_{\rm UV}$) than the observations can probe. At the highest redshifts, $z > 15$, we see that the model predictions undershoot the observations. However, since the uncertainties in the data are large, the observations at high z are compatible with a strongly suppressed UV luminosity function.

For the FDM model, the parameters for the star formation rate, the enhancement of the UV luminosity and the suppression of the feedback effects remain the same as in the CDM model, as seen in the posterior plots in Fig. 7. This reflects the fact that the FDM prediction deviates sharply from the CDM prediction at the suppression scale. The main difference between FDM and CDM models is seen in z_*, which determines when the feedback effects are fully removed. For low FDM masses the feedback effects can be fully removed at lower z because the luminosity function is suppressed by the FDM quantum pressure, which causes the peak in the posterior of the FDM mass shown in the top panels of Fig. 9. The posterior also shows a lower bound on the FDM mass, $m_{\rm FDM} > 5.6 \times 10^{-22}$ eV at 95% CL. The UV luminosity function has been previously studied in the FDM model (Bozek et al. 2015; Schive et al. 2016; Corasaniti et al. 2017; Winch et al. 2024). Our constraint on the FDM mass is somewhat stronger than that obtained by Winch et al. (2024) using the HST data and JWST data from Harikane et al. (2024).

The deviation from the CDM prediction is smoother in the WDM case, and starts far above the half-mode mass. Consequently, the WDM fit at small $m_{\rm WDM}$ moves to smaller values of α and larger values of β than in the CDM model, as seen in the posterior plots in Figs. 7 and 9. Due to the degeneracy between α and $m_{\rm WDM}$, in particular, the WDM model favours a narrow mass range around 2 keV. The maximum likelihood in the WDM model is slightly higher than in CDM, $\Delta \ln \mathcal{L} \approx 2.7$, and the Bayesian information criterion (BIC), which accounts for the increase in the number of parameters, indicates a mild (insignificant) preference for WDM over CDM with $\Delta{\rm BIC} \approx -0.4$. As seen from Fig. 8, the fit improves in particular at $z = 17$.

The posteriors of the WDM mass shown in the middle panels of Fig. 9 show a lower bound $m_{\rm WDM} > 1.5$ keV at 95% CL. Several previous works have derived constraints on WDM

[6] However, a mass-dependent scatter that increases at small halo masses could enhance the UV luminosities at high z (Gelli et al. 2024).

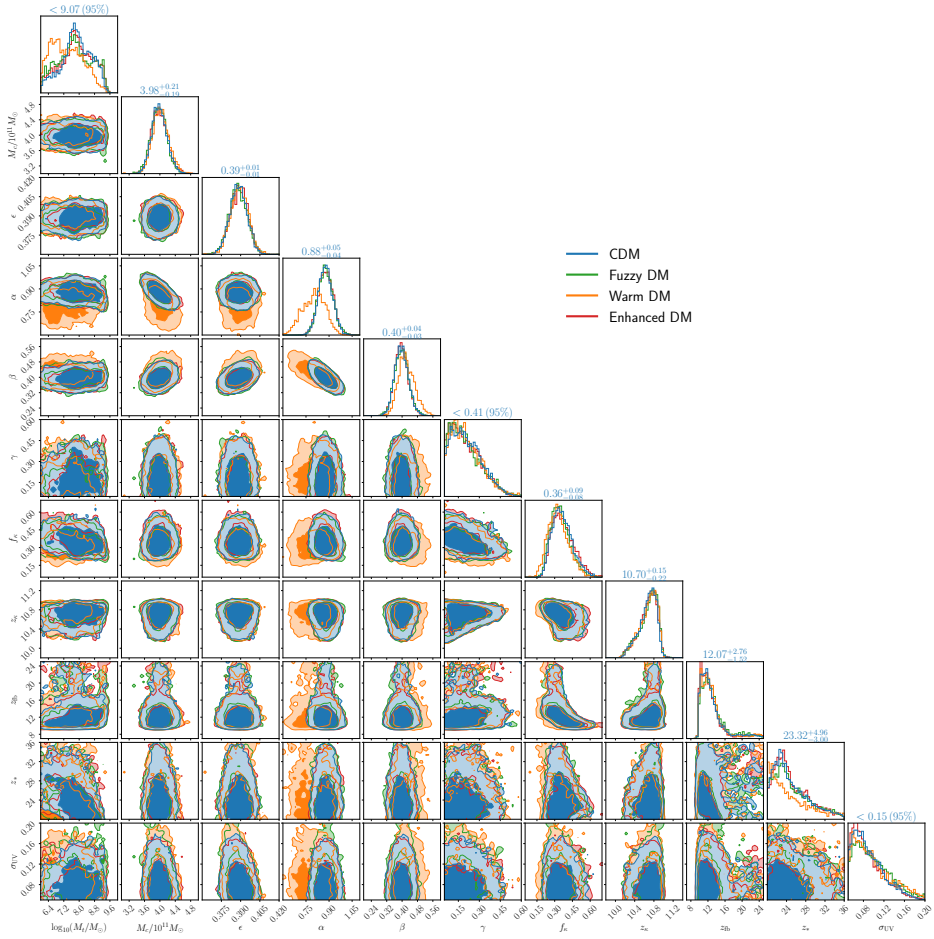


Fig. 7. Posteriors of SFR fits to the UV luminosity function observations in different models. The contours indicate the 68% and 95% credible regions. The marginalised 68% credible ranges and the 95% credible constraints are shown for CDM at the top of each column.

from UV luminosity function observations (Menci et al. 2016; Corasaniti et al. 2017; Rudakovskyi et al. 2021; Hibbard et al. 2022; Maio & Viel 2023; Liu et al. 2024). Although we used a larger set of JWST data, the bound on the WDM mass is weaker than that found by Liu et al. (2024). This difference arises because Liu et al. (2024) uses a low-z fit of the HMF to numerical simulations derived in Stücker et al. (2021), while we estimate the WDM HMF using the transfer function and the smooth-k window function.

5.3. Enhanced small-scale perturbations

Enhancing the small-scale structure has the opposite effect to WDM or FDM, but the conclusions are very similar. We see in Fig. 8 that for $z \lesssim 10$ the predictions resemble those of CDM, and only at the highest redshifts do we see a small enhancement of the UV luminosity function that grows at lower scales, $M_{\rm UV} \lesssim -20$. This reflects the fact that the higher the redshift, the more deviations from CDM one can expect to see, especially for low-mass halos. When fitting the star formation parameters, the results are independent of the enhancement of the scales and very similar to CDM, as seen in Fig. 7. We conclude that enhancing the matter power spectrum does not help to explain the excess in the UV radiation at high z when other possibilities (e.g. Pop III stars) are taken into account. The fit favours a sharp transition to Pop III stars and such a feature cannot be reproduced by enhancing the matter power spectrum. The fit gives a lower bound of $k_c > 25\,{\rm Mpc}^{-1}$ at 95% CL, as seen in the bottom panels of Fig. 9. The excluded region corresponds to axion mini-clusters produced with $m_a < 6.6 \times 10^{-17}$ eV or PBHs that obey $f_{\rm PBH} > \max[105 M_\odot/m_{\rm PBH}, 10^{-4}(m_{\rm PBH}/10^4 M_\odot)^{-0.09}]$, where the second bound comes from the requirement that $k_c < k_{\rm cut}$. Our constraint on k_c is significantly stronger than that implied by the measurements of the matter power spectrum at scales $0.5 < k/{\rm Mpc}^{-1} < 10$ derived by Sabti et al. (2022b) and that implied by the analysis of Iršič et al. (2020), both using the HST data alone. For PBHs, our constraint is stronger than that derived by Murgia et al. (2019), which considered the limit on an enhanced small-scale power spectrum imposed by the Lyman-α forest observations. In Fig. 10, we show a compilation of previous PBH constraints together with our new constraint in red.

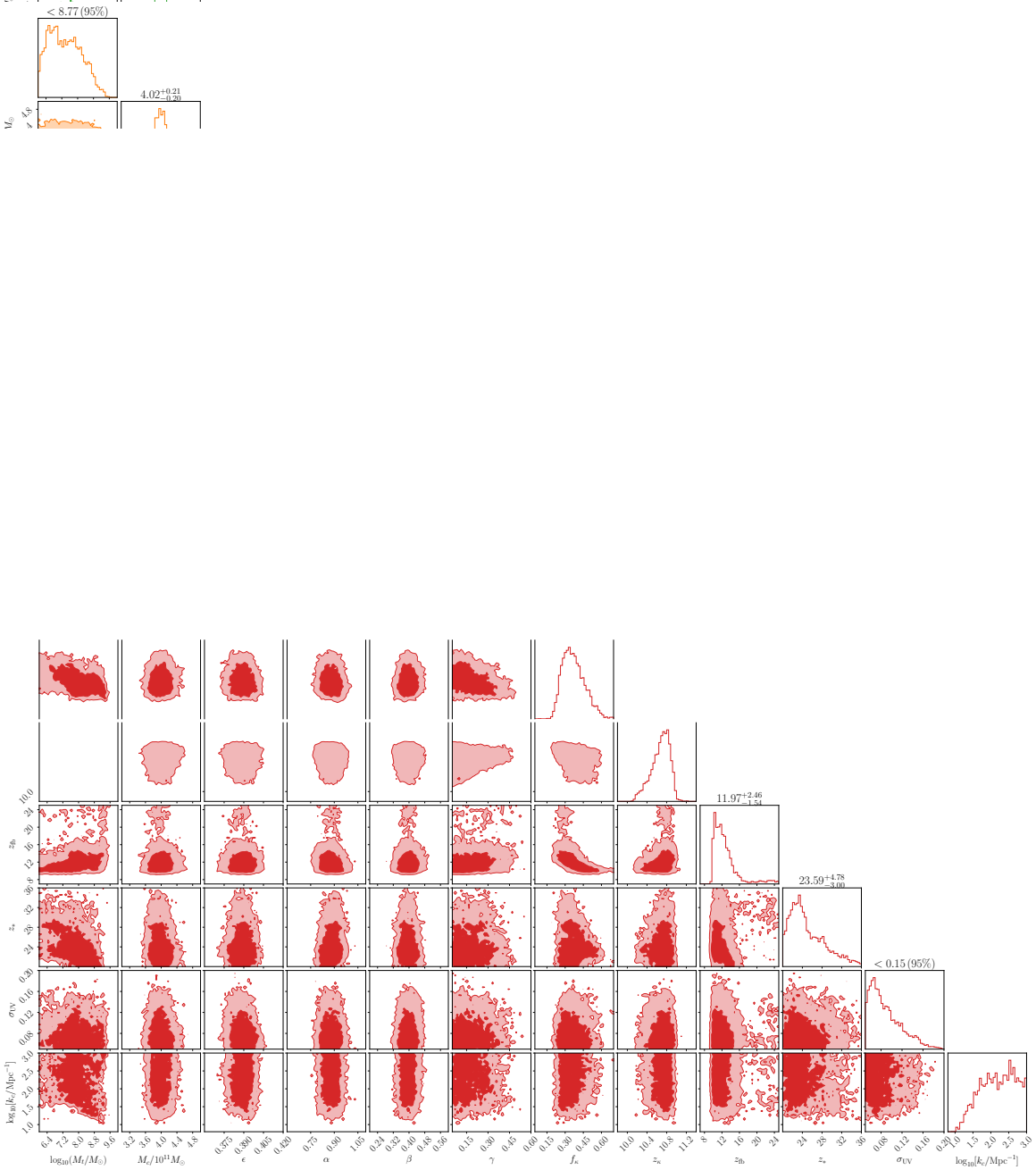


Fig. 9. Posteriors of SFR fits to the UV luminosity function observations as a function of FDM mass (top panels), WDM mass (middle panels), and the enhancement scale (bottom panels). The contours indicate the 68% and 95% credible regions.

6. Conclusions

We have investigated in this work the implications of the high-z UV luminosity function observations with HST and JWST for star formation and different DM models. We have revisited the excursion-set formalism to compute the unconditional and conditional halo mass functions and derived a new estimate of the halo growth rate. We have also proposed a new window function, inspired by the smooth-k Leo et al. (2018) window function, which reproduces well the numerical results and matches better the CDM predictions at high masses.

We have found that, assuming that the SFR is proportional to the halo growth rate, the observations of the UV luminosity function indicate two interesting effects that change the luminosity function at high z. First, we find evidence at more than 95% CL for a sharp enhancement in the luminosity function around $z \simeq 10$. This can be explained by a transition from a Salpeter mass function to a top-heavy Pop III mass function. Beyond $z = 10 - 15$ the luminosity function needs to be further enhanced, in particular to match the JWST observations at $z = 17$ and 25. This enhancement can be achieved by reducing the feedback from SNe and AGNs. The evidence for this second effect is, however, less than 95% CL. We have also considered the effect of dust attenuation and lensing, and shown that they are relevant for the bright end of the luminosity function at $z < 10$, but do not play a significant role in fitting the higher-redshift data.

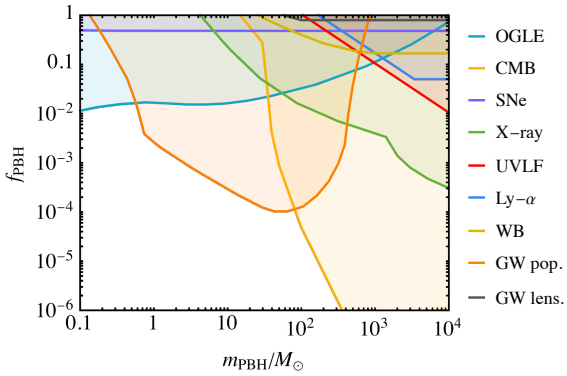


Fig. 10. Constraints on the fraction of the DM density provided by PBHs from the OGLE microlensing survey (Mróz et al. 2024), accretion limits from CMB observations (Serpico et al. 2020), SNe lensing (Zumalacarregui & Seljak 2018), the population of X-ray sources (Inoue & Kusenko 2017), Lyman-α data (Murgia et al. 2019), survival of wide binaries (Monroy-Rodríguez & Allen 2014), the population of gravitational wave (GW) sources (Andrés-Carcasona et al. 2024), and GW lensing (Urrutia et al. 2023). The red line shows the upper limit derived in this work from measurements of the UV luminosity function.

We have considered departures from CDM in two directions, suppressing the small-scale structures via either warm or fuzzy DM and enhancing the matter power spectrum with white noise, which could happen in scenarios that include axion mini-clusters or PBHs. When marginalising over the SFR parametrisation, we find a lower bound of 5.6×10^{-22}eV for the FDM mass and 1.5 keV for the WDM mass, at 95% CL. Because the new JWST data at high z still have significant uncertainties, we find that the DM bounds are not much stronger than those obtained using the HST data alone. Another important conclusion is that enhanced matter perturbations are not preferred over changes in the star formation rate due to the presence of Pop III stars and changes in the feedback in the pre-reionisation era. We have found that the data constrain the scale of the enhancement to $k_c > 25\,\mathrm{Mpc}^{-1}$ at 95% CL, excluding, for example, the possibility of axion mini-clusters produced for $m_a < 6.6 \times 10^{-17}$ eV or heavy PBHs that constitute a fraction $f_{\rm PBH} > \max[105 M_\odot/m_{\rm PBH}, 10^{-4}(m_{\rm PBH}/10^4 M_\odot)^{-0.09}]$ of DM. As JWST gathers more data and the uncertainties are reduced, the bounds on DM will become significantly stronger and our knowledge of star formation at high z will improve. In a forthcoming paper we will leverage our findings to compute the growth of SMBHs in non-CDM scenarios and obtain more competitive bounds on deviations from CDM using the measurements of high-z SMBHs by JWST.

Acknowledgements. This work was supported by the Estonian Research Council grants PRG803, PSG869, RVTT3 and RVTT7 and the Center of Excellence program TK202. The work of J.E. and M.F. was supported by the United Kingdom STFC Grants ST/T000759/1 and ST/X000753/1. The work of V.V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736.

References

Aghanim, N., Akrami, Y., Ashdown, M., et al. 2020, A&A, 641, A6 [Erratum: Astron. Astrophys. 652, C4 (2021)]
Andrés-Carcasona, M., Iovino, A. J., Vaskonen, V., et al. 2024, Phys. Rev. D, 110, 023040
Arrabal Haro, P., Dickinson, M., Finkelstein, S. L., et al. 2023, Nature, 622, 707
Bagley, M. B., Finkelstein, S. L., Rojas-Ruiz, S., et al. 2024, ApJ, 961, 209
Bian, F., Fan, X., Jiang, L., et al. 2013, ApJ, 774, 28
Bode, P., Ostriker, J. P., & Turok, N. 2001, ApJ, 556, 93
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2014, ApJ, 793, 115
Bouwens, R. J., Oesch, P. A., Stefanon, M., et al. 2021, AJ, 162, 47
Bouwens, R. J., Stefanon, M., Brammer, G., et al. 2023, MNRAS, 523, 1036
Bozek, B., Marsh, D. J. E., Silk, J., & Wyse, R. F. G. 2015, MNRAS, 450, 209
Cappelluti, N., Hasinger, G., & Natarajan, P. 2022, ApJ, 926, 205
Carr, B. J., & Silk, J. 1983, ApJ, 268, 1
Carr, B., & Silk, J. 2018, MNRAS, 478, 3756
Castellano, M., Fontana, A., Treu, T., et al. 2023, ApJ, 948, L14
Castellano, M., Fontana, A., Merlin, E., et al. 2025, A&A, submitted [arXiv:2504.05893]
Chakraborty, A., & Choudhury, T. R. 2024, JCAP, 07, 078
Corasaniti, P. S., Agarwal, S., Marsh, D. J. E., & Das, S. 2017, Phys. Rev. D, 95, 083512
Correa, C. A., Wyithe, J. S. B., Schaye, J., & Duffy, A. R. 2015, MNRAS, 450, 1514
De Luca, V., Desjacques, V., Franciolini, G., & Riotto, A. 2020, JCAP, 11, 028
Dekel, A., Sarkar, K. C., Birnboim, Y., Mandelker, N., & Li, Z. 2023, MNRAS, 523, 3201
Dodelson, S. 2003, Modern Cosmology (Amsterdam: Academic Press)
Donnan, C. T., McLeod, D. J., Dunlop, J. S., et al. 2023, MNRAS, 518, 6011
Donnan, C. T., McLure, R. J., Dunlop, J. S., et al. 2024, MNRAS, 533, 3222
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
Ellis, D., Marsh, D. J. E., Eggemeier, B., et al. 2022, Phys. Rev. D, 106, 103514
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2024, Phys. Rev. D, 109, L021302
Fairbairn, M., Marsh, D. J. E., Quevillon, J., & Rozier, S. 2018, Phys. Rev. D, 97, 083502
Feix, M., Frank, J., Pargner, A., et al. 2019, JCAP, 05, 021
Ferrami, G., & Wyithe, J. S. B. 2023, MNRAS, 523, L21
Finkelstein, S. L., Bagley, M., Song, M., et al. 2022, ApJ, 928, 52
Fujimoto, S., Naidu, R. P., Chisholm, J., et al. 2025, ApJ, 989, 46
Fukushima, H., & Yajima, H. 2021, MNRAS, 506, 5512
Gelli, V., Mason, C., & Hayward, C. C. 2024, ApJ, 975, 192
Gelman, A., & Rubin, D. B. 1992, Statist. Sci., 7, 457
Hansen, S. H., Lesgourgues, J., Pastor, S., & Silk, J. 2002, MNRAS, 333, 544
Harikane, Y., Ono, Y., Ouchi, M., et al. 2022, ApJS, 259, 20
Harikane, Y., Ouchi, M., Oguri, M., et al. 2023, ApJS, 265, 5
Harikane, Y., Nakajima, K., Ouchi, M., et al. 2024, ApJ, 960, 56
Hibbard, J. J., Mirocha, J., Rapetti, D., et al. 2022, ApJ, 929, 151
Hirano, S., Hosokawa, T., Yoshida, N., et al. 2014, ApJ, 781, 60
Hirano, S., Hosokawa, T., Yoshida, N., Omukai, K., & Yorke, H. W. 2015, MNRAS, 448, 568
Hu, W., Barkana, R., & Gruzinov, A. 2000, Phys. Rev. Lett., 85, 1158
Hütsi, G., Raidal, M., Urrutia, J., Vaskonen, V., & Veermäe, H. 2023, Phys. Rev. D, 107, 043502
Inman, D., & Ali-Haïmoud, Y. 2019, Phys. Rev. D, 100, 083528
Inoue, Y., & Kusenko, A. 2017, JCAP, 10, 034
Iršič, V., Xiao, H., & McQuinn, M. 2020, Phys. Rev. D, 101, 123518
Jiang, F., & van den Bosch, F. C. 2014, MNRAS, 440, 193
Klessen, R. S., & Glover, S. C. O. 2023, ARA&A, 61, 65
Kolb, E. W., & Tkachev, I. I. 1993, Phys. Rev. Lett., 71, 3051
Leethochawalit, N., Roberts-Borsani, G., Morishita, T., Trenti, M., & Treu, T. 2023, MNRAS, 524, 5454
Leo, M., Baugh, C. M., Li, B., & Pascoli, S. 2018, JCAP, 04, 010
Liu, B., Shan, H., & Zhang, J. 2024, ApJ, 968, 79
Lovell, M. R., Frenk, C. S., Eke, V. R., et al. 2014, MNRAS, 439, 300
Ludlow, A. D., Bose, S., Angulo, R. E., et al. 2016, MNRAS, 460, 1214
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Maio, U., & Viel, M. 2023, A&A, 672, A71
Maiolino, R., Übler, H., Perna, M., et al. 2024, A&A, 687, A67
Marsh, D. J. E. 2016, Phys. Rept., 643, 1
McLeod, D. J., McLure, R. J., & Dunlop, J. S. 2016, MNRAS, 459, 3812
Menci, N., Grazian, A., Castellano, M., & Sanchez, N. G. 2016, ApJ, 825, L1
Monroy-Rodríguez, M. A., & Allen, C. 2014, ApJ, 790, 159
Morishita, T., & Stiavelli, M. 2023, ApJ, 946, L35
Mróz, P., Udalski, A., Szymanski, M. K., et al. 2024, Nature, 632, 749
Murgia, R., Scelfo, G., Viel, M., & Raccanelli, A. 2019, Phys. Rev. Lett., 123, 071102
Naidu, R. P., Oesch, P. A., Setton, D. J., et al. 2022a, arXiv e-prints [arXiv:2208.02794]
Naidu, R. P., Oesch, P. A., van Dokkum, P., et al. 2022b, ApJ, 940, L14
Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. 2018, ApJ, 855, 105

Oke, J. B., & Gunn, J. E. 1983, ApJ, 266, 713
Parashari, P., & Laha, R. 2023, MNRAS, 526, L63
Passaglia, S., & Hu, W. 2022, Phys. Rev. D, 105, 123529
Pérez-González, P. G., et al. 2025, arXiv e-prints [arXiv:2503.15594]
Pérez-González, P. G., Costantin, L., Langeroodi, D., et al. 2023, ApJ, 951, L1
Reed, D., Gardner, J., Quinn, T., et al. 2003, MNRAS, 346, 565
Rudakovskyi, A., Mesinger, A., Savchenko, D., & Gillet, N. 2021, MNRAS, 507, 3046
Sabti, N., Muñoz, J. B., & Blas, D. 2022a, Phys. Rev. D, 105, 043518
Sabti, N., Muñoz, J. B., & Blas, D. 2022b, ApJ, 928, L20
Schive, H.-Y., Chiueh, T., Broadhurst, T., & Huang, K.-W. 2016, ApJ, 818, 89
Serpico, P. D., Poulin, V., Inman, D., & Kohri, K. 2020, Phys. Rev. Res., 2, 023204
Sheth, R. K., & Tormen, G. 1999, MNRAS, 308, 119
Sheth, R. K., & Tormen, G. 2002, MNRAS, 329, 61
Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1
Sipple, J., Lidz, A., Grin, D., & Sun, G. 2025, MNRAS, 538, 1830
Steinhardt, C. L., Kokorev, V., Rusakov, V., Garcia, E., & Sneppen, A. 2023, ApJ, 951, L40
Stücker, J., Angulo, R. E., Hahn, O., & White, S. D. M. 2021, MNRAS, 509, 1703
Susa, H., & Umemura, M. 2004, ApJ, 600, 1
Takahashi, R., Oguri, M., Sato, M., & Hamana, T. 2011, ApJ, 742, 15
Trenti, M., Perna, R., & Jimenez, R. 2015, ApJ, 802, 103
Urrutia, J., Vaskonen, V., & Veermäe, H. 2023, Phys. Rev. D, 108, 023507
Vanzella, E., Loiacono, F., Bergamini, P., et al. 2023, A&A, 678, A173
Ventura, E. M., Qin, Y., Balu, S., & Wyithe, J. S. B. 2024, MNRAS, 529, 628
Verwohlt, J., Mason, C. A., Muñoz, J. B., et al. 2024, Phys. Rev. D, 110, 103533
Viel, M., Lesgourgues, J., Haehnelt, M. G., Matarrese, S., & Riotto, A. 2005, Phys. Rev. D, 71, 063534
Vogelsberger, M., Nelson, D., Pillepich, A., et al. 2020, MNRAS, 492, 5167
Wang, B., & Heckman, T. M. 1996, ApJ, 457, 645
Winch, H., Rogers, K. K., Hložek, R., & Marsh, D. J. E. 2024, ApJ, 976, 40
Yung, L. Y. A., Somerville, R. S., & Iyer, K. G. 2025, arXiv e-prints [arXiv:2504.18618]
Zhang, J., Ma, C.-P., & Fakhouri, O. 2008, MNRAS, 387, L13
Zheng, H., Bose, S., Frenk, C. S., et al. 2024, MNRAS, 528, 7300
Zumalacarregui, M., & Seljak, U. 2018, Phys. Rev. Lett., 121, 141101

Appendix 6

VI
J. Ellis, M. Fairbairn, G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Prospects for future binary black hole gravitational wave studies in light of PTA measurements. *Astron. Astrophys.*, 676:A38, 2023

A&A 676, A38 (2023)
https://doi.org/10.1051/0004-6361/202346268
© The Authors 2023

Astronomy & Astrophysics

Prospects for future binary black hole gravitational wave studies in light of PTA measurements

John Ellis[1,2,3], Malcolm Fairbairn[2], Gert Hütsi[3], Martti Raidal[3], Juan Urrutia[3,4], Ville Vaskonen[3,5,6], and Hardi Veermäe[3]

[1] Theoretical Physics Department, CERN, Geneva, Switzerland
[2] King's College London, Strand, London WC2R 2LS, UK
[3] Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia
e-mail: juan.urrutia@kbfi.ee
[4] Departament of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia
[5] Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy
[6] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy

Received 28 February 2023 / Accepted 3 June 2023

ABSTRACT

NANOGrav and other Pulsar Timing Arrays (PTAs) have discovered a common-spectrum process in the nHz range that may be due to gravitational waves (GWs): if so, they are likely to have been generated by black hole (BH) binaries with total masses $>10^9\,M_\odot$. Using the Extended Press-Schechter formalism to model the galactic halo mass function and a simple relation between the halo and BH masses suggests that these binaries have redshifts $z = \mathcal{O}(1)$ and mass ratios $\gtrsim 10$, and that the GW signal at frequencies above $\mathcal{O}(10)$ nHz may be dominated by relatively few binaries that could be distinguished experimentally and would yield observable circular polarization. Extrapolating the model to higher frequencies indicates that future GW detectors such as LISA and AEDGE could extend the PTA observations to lower BH masses $\gtrsim 10^3\,M_\odot$.

Key words. black hole physics – gravitational waves – cosmology: theory

1. Introduction

The discovery (LIGO Scientific & Virgo Collaborations 2016) of gravitational waves (GWs) by the LIGO and Virgo experiments, in the $\mathcal{O}(100)$ Hz range of frequencies generated by mergers of black holes (BHs) with masses of $\mathcal{O}(10{-}100)\,M_\odot$ (LIGO Scientific, Virgo & KAGRA Collaborations 2021), has opened a new window onto both astrophysics and cosmology. Supermassive black holes (SMBHs) with masses of $\mathcal{O}(10^6{-}10^{10})\,M_\odot$ are known to be present in galactic nuclei, and the immediate surroundings of two of them have recently been imaged by the Event Horizon Telescope (Event Horizon Telescope Collaboration 2019, 2022). However, information on intermediate-mass black holes (IMBHs) with masses in the range $\mathcal{O}(10^3{-}10^5)\,M_\odot$ (Greene et al. 2020) is less complete. Observations of IMBHs and their mergers would cast light on the uncharted mechanisms that must be presented for SMBHs to form.

A prerequisite for BH mergers is the formation of tightly bound binaries that can radiate GWs efficiently. While there are several well-understood paths for forming compact stellar-mass BHs binaries (e.g., common envelope evolution or dynamical capture in dense stellar environments), the final stage of SMBH binary formation in galaxy mergers is still not fully understood and is commonly known as "the final parsec problem" (Begelman et al. 1980). However, observations of quasar optical variability provide indirect evidence for the existence of some tight binary SMBHs in the regime where the emission of GWs must already have a noticeable effects on their orbital evolution (Rieger & Mannheim 2000; De Paolis et al. 2002, 2004; Valtonen 2008; Boroson & Lauer 2009; Iguchi et al. 2010; Graham et al. 2015; O'Neill et al. 2022; Kovačević et al. 2022)[1]. The GW data will be essential for a better understanding how SMBHs overcome the "final parsec" obstacle, and how the assembly of the SMBH population proceeds in general.

Pulsar timing arrays (PTAs) are potentially sensitive to GWs in the nHz range, while NANOGrav and other PTAs have recently reported evidence for a common-spectrum stochastic process (Arzoumanian et al. 2020; Goncharov et al. 2021; Chen et al. 2021; Antoniadis et al. 2022). Their signals have power spectra that are consistent with predictions based on inspiralling binary SMBH models (Phinney 2001)[2], but they have not (yet) detected the Hellings-Downs quadrupolar signature that is characteristic of GWs (Hellings & Downs 1983)[3].

The BH merger interpretation of the PTA measurements does not provide direct information on the masses of the infalling BHs, which requires the modelling and measurements of this stochastic GW background in different frequency

[1] However, alternative scenarios for quasar optical variability may also be considered, such as intrinsic variability in the accretion disc (King et al. 2013).
[2] A similar signal could be generated by primordial BHs (Vaskonen & Veermäe 2021; De Luca et al. 2021; Kohri & Terada 2021; Ashoorioon et al. 2022), but might require modifications of models based on simple cosmic string networks (Ellis & Lewicki 2021; Blasi et al. 2021; Buchmuller et al. 2020): see, e.g., Blanco-Pillado et al. (2021).
[3] We note in passing that binaries capable of generating the PTA nHz background could not explain the year-like quasi-periodicity seen in blazars (Holgado et al. 2018).

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

ranges (Sesana et al. 2008). The purpose of this paper is to illustrate how future measurements of the GW spectrum due to unresolved infall sources and individual binary mergers at frequencies between the PTAs and LIGO/Virgo will be able to extend the PTA measurements to lower black-hole mass ranges, probing models for the assembly of SMBHs, and perhaps obtaining direct evidence for the mergers of IMBHs. We take as the amplitude of the common-spectrum process measured by the PTAs $A = 2.8^{+1.2}_{-0.8} \times 10^{-15}$ at a reference frequency of 1/yr, as found by the International Pulsar Timing Array (IPTA; Antoniadis et al. 2022) assuming the spectral index of $\alpha = -2/3$ expected from inspiralling SMBHs, which is consistent with other measurements.

We use examples of next-generation GW detectors the planned LISA laser interferometer (Berti et al. 2006), whose peak sensitivity is at frequencies $O(10^{-4}-10^{-2})$ Hz, the proposed ET laser interferometer (Sathyaprakash et al. 2012), whose peak sensitivity is at frequencies $O(1-10^2)$ Hz and the projected AION-km and AEDGE atom interferometers (Badurina et al. 2020; El-Neaj et al. 2020), whose peak sensitivities would be in the range of $O(10^{-2}-1)$ Hz. We show that, while none of these detectors would observe a signal from the binary black holes weighing $>O(10^9)\,M_\odot$ that would probably be responsible for most of the PTA signal, LISA would observe a merger signal from binary black holes weighing $O(10^3-10^9)\,M_\odot$ that might contribute part of the PTA signal, while AEDGE would observe a merger signal if there is a population of binary black holes weighing $O(10^3-10^6)\,M_\odot$ with a formation history similar to the heavier BHs responsible for the PTA signal.

Throughout this paper, we use natural units with $c = 1$, $G_N = 1$.

2. BH merger rate

As a first step towards estimating the BH merger rate, we use the Extended Press-Schechter (EPS) formalism (Press & Schechter 1974; Bond et al. 1991) to calculate the galactic halo mass function,

$$\frac{\mathrm{d}n(M,t)}{\mathrm{d}\ln M} = \frac{\rho_0}{M}\sqrt{\frac{2}{\pi}}\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\frac{\delta_c(z)}{\sigma(M)}e^{-\frac{\delta_c(z)^2}{2\sigma^2(M)}}, \tag{1}$$

where ρ_0 is the background matter density today, $\sigma^2(M)$ is the variance of the matter fluctuations and $\delta_c(z)$ is the critical overdensity for collapse. The latter is given by $\delta_c(z) \approx 1.686/D(z)$ where $D(z)$ is the linear growth function (Dodelson 2003). We calculate $\sigma^2(M)$ from the cold dark matter power spectrum with the *Planck* 2018 cosmological parameters (Planck Collaboration VI 2020) and the transfer function derived in Eisenstein & Hu (1998). The EPS formalism also gives an estimate for the probability per unit time for a halo of mass M_1 to merge with another one of mass, M_2, at some redshift, z, and become a halo of mass $M_f = M_1+M_2$ (Lacey & Cole 1993):

$$\frac{\mathrm{d}p(M_1,M_2,t)}{\mathrm{d}t\mathrm{d}M_2} = \frac{1}{M_f}\sqrt{\frac{2}{\pi}}\left|\frac{\dot\delta_c}{\delta_c}\right|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M_f}\left[1-\frac{\sigma^2(M_f)}{\sigma^2(M_1)}\right]^{-\frac{3}{2}} \times\frac{\delta_c}{\sigma(M_f)}\exp\left[-\frac{\delta_c^2}{2}\left(\frac{1}{\sigma^2(M_f)}-\frac{1}{\sigma^2(M_1)}\right)\right]. \tag{2}$$

Defining the halo merger rate kernel:

$$Q(M_1,M_2,t) \equiv \frac{\mathrm{d}p(M_1,M_2,t)}{\mathrm{d}t\mathrm{d}M_2}\left[\frac{\mathrm{d}n(M_2,t)}{\mathrm{d}M_2}\right]^{-1}, \tag{3}$$

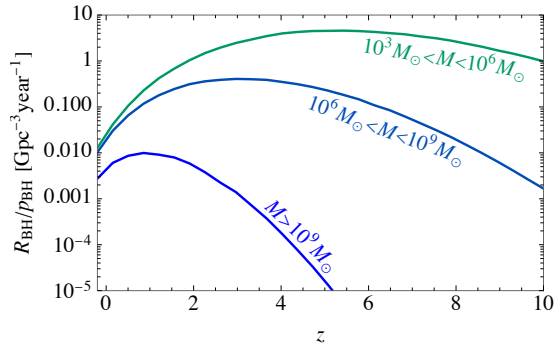


Fig. 1. BH merger rates calculated as functions of the redshift z in three ranges of the total merging BH mass M and normalized relative to the merger probability $p_{\rm BH}$ discussed in the text.

the halo merger rate is

$$\frac{\mathrm{d}R_h}{\mathrm{d}M_1\mathrm{d}M_2} = \frac{\mathrm{d}n(M_1,t)}{\mathrm{d}M_1}\frac{\mathrm{d}n(M_2,t)}{\mathrm{d}M_2}Q(M_1,M_2,t). \tag{4}$$

As discussed by Benson et al. (2005), the merger rate kernel $Q(M_1,M_2,t)$ in the EPS formalism is not symmetric in the exchange of M_1 and M_2. However, it was shown by Erickcek et al. (2006) that taking $M_1 < M_2$ in the merger kernel aptly approximates the result of the Benson-Kamionkowski-Hassani (BKH) merger theory (Benson et al. 2005) where the merger rate kernel is computed from the Smoluchowski coagulation equation and preserves the EPS halo mass function. We will, therefore, require that the first argument of Q is smaller than the second.

In order to estimate the merger rate of central BHs, we must estimate the probability $p_{\rm occ}(m|M,z)$ that a BH of mass m occupies a halo of mass M, as well as the probability $p_{\rm merg}(m_1,m_2)$ that the galactic merger leads to a merger of their central BHs. The resulting BH merger rate can be expressed as

$$\frac{\mathrm{d}R_{\rm BH}}{\mathrm{d}m_1\mathrm{d}m_2} = \int \mathrm{d}M_1\mathrm{d}M_2\, p_{\rm merg}(m_1,m_2) \times p_{\rm occ}(m_1|M_1,z)p_{\rm occ}(m_2|M_2,z)\frac{\mathrm{d}R_h}{\mathrm{d}M_1\mathrm{d}M_2}. \tag{5}$$

We assume the following simple redshift-dependent relation between the halo mass and the BH mass (Barkana & Loeb 2001; Wyithe & Loeb 2003):

$$\frac{M_v}{10^{12}M_\odot} = 10.5\left[\frac{\Omega_M(0)}{\Omega_M(z)}\frac{\Delta_c(z)}{18\pi^2}\right]^{-\frac{1}{2}}(1+z)^{-\frac{3}{2}}\left[\frac{m_{\rm BH}}{10^8 M_\odot}\right]^{\frac{3}{5}}, \tag{6}$$

where $\Delta_c(z) = 18\pi^2 + 82[\Omega_M(z)-1] - 39[\Omega_M(z)-1]^2$ is the critical overdensity at virialization, which corresponds to $p_{\rm occ}(m|M) = p_{\rm occ}(m)\delta(m - m_{\rm BH}(M))$. It is generally expected that SMBHs inhabit most large galaxies, and X-ray observations (Miller et al. 2015) constrain the SMBH occupation fraction $p_{\rm occ}(m)$ to be $> 20\%$ for early galaxies with lower stellar masses $10^7 < M_*/M_\odot < 10^{10}$. It has been estimated that an initially small occupation fraction may grow at low redshifts (see, e.g., Lippai et al. 2009), which we neglect in this analysis. As we show later, the PTA GW signal is dominated by events in a range $1 \lesssim z \lesssim 3$. The estimated z dependence of $p_{\rm occ}$ would shift this range towards lower redshifts when compared to a constant $p_{\rm occ}$, increasing the observability of BH mergers.

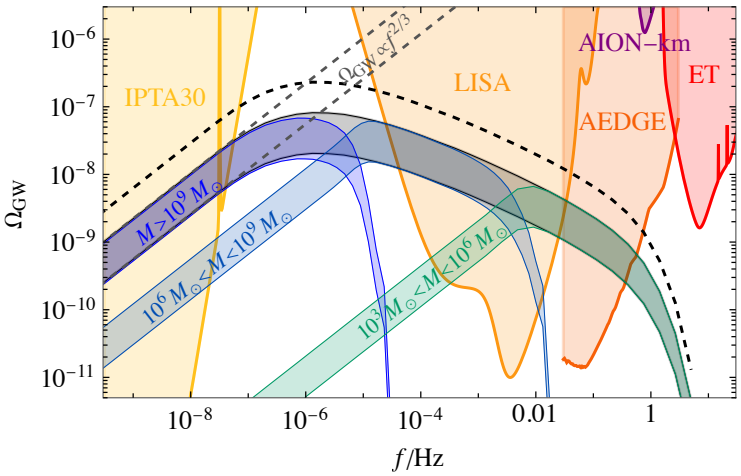


Fig. 2. Mean GW energy density spectrum from massive BH mergers compared with the sensitivities of different experiments. The black dashed curve shows the case where all mergers of galaxies produce a GW signal for BHs heavier than $10^3\,M_\odot$ ($p_{\rm BH} = 1$). The colored bands show the spectra from SMBHs heavier than $10^9\,M_\odot$ (dark blue), from SMBHs in the range ($10^6\,M_\odot, 10^9\,M_\odot$); light blue), and from IMBHs with masses in the range 10^3–$10^6\,M_\odot$ (green), assuming a universal efficiency factor $p_{\rm BH} = 0.17^{+0.18}_{-0.08}$. The shaded regions show the prospective sensitivities of IPTA30 (Chen et al. 2017; Kaiser & McWilliams 2021), LISA (LISA Collaboration 2017), AEDGE (El-Neaj et al. 2020; Badurina et al. 2021), AION-km (Badurina et al. 2020, 2021) and ET (Sathyaprakash et al. 2012).

In view of the paucity of information about the merger probability $p_{\rm merg}(m_1, m_2)$, we model $p_{\rm merg}$ by a constant for the sake of simplicity. After the galactic merger, various dynamical mechanisms must decrease the size of the SMBH binary below subparsec scales in order for a GW emission-driven merger to take place (Begelman et al. 1980). The crossing of the final parsec is determined mostly by binary hardening via stellar loss-cone scattering (Merritt 2013), which can currently be considered the largest source of uncertainty in $p_{\rm merg}$. The numerical study of Kelley et al. (2017a), based on a population of 10^6–$10^{10}\,M_\odot$ BH binaries in the `Illustris` simulation (Vogelsberger et al. 2014), found for a wide range of model parameters that the coalescing fraction is nearly independent of the total mass of the binary but decreases with the mass ratio. Neglecting the latter dependence should not significantly affect our results, since extreme mass ratio inspirals contribute subdominantly to the GW background. With these assumptions, we obtain:

$$\frac{{\rm d}R_{\rm BH}}{{\rm d}m_1 {\rm d}m_2} \approx p_{\rm BH} \frac{{\rm d}M_1}{{\rm d}m_1} \frac{{\rm d}M_2}{{\rm d}m_2} \frac{{\rm d}R_h}{{\rm d}M_1 {\rm d}M_2}, \tag{7}$$

where $p_{\rm BH} \equiv p_{\rm occ}(m_1) p_{\rm occ}(m_2) p_{\rm merg}$ denotes the probability that the halo merger leads to a merger of the central SMBHs.

We show in Fig. 1 the BH merger rates as functions of the redshift z, calculated using this approach in three ranges of the total merging BH mass M and normalized relative to the merger probability $p_{\rm BH}$ discussed above. We see that mergers with total masses $M > 10^9\,M_\odot$ occur typically at $z = O(1)$, those with total masses $M \in (10^6, 10^9)\,M_\odot$ occur typically at $z = O(3)$, and those with total masses $M \in (10^3, 10^6)\,M_\odot$ occur typically at $z = O(5)$.

In this paper, we use the IPTA measurement (Antoniadis et al. 2022) to normalize $p_{\rm BH}$ for large BH masses, which circumvents the astrophysical uncertainties related to $p_{\rm occ}$ and $p_{\rm merg}$, and extrapolate it to smaller BH masses. We also comment on the potential mass dependence of $p_{\rm BH}$, which is relevant for the GW phenomenology when we extrapolate from the frequency band relevant for PTAs to those to be explored by next-generation GW detectors.

3. Analysis

3.1. The GW energy spectrum

The mean GW energy density spectrum from the SMBH binary population can be estimated as (Phinney 2001):

$$\Omega_{\rm GW}(f) \equiv \frac{1}{\rho_{\rm c}} \frac{{\rm d}\rho_{\rm GW}}{{\rm d}\ln f} = \int {\rm d}\lambda \, \frac{2\pi}{5} \frac{f^3 |\tilde{h}(f)|^2}{\rho_c}, \tag{8}$$

where $|\tilde{h}(f)|$ denotes the optimal amplitude of the Fourier transform of the GW strain, $\rho_{\rm c} = 3H_0^2/8\pi$,

$${\rm d}\lambda = {\rm d}m_1 {\rm d}m_2 \, \frac{{\rm d}z}{1+z} \frac{{\rm d}V_c}{{\rm d}z} \frac{{\rm d}R_{\rm BH}(t)}{{\rm d}m_1 {\rm d}m_2}, \tag{9}$$

is the differential merger rate and V_c denotes the comoving volume available at redshift z. We compute $|\tilde{h}(f)|$ using the inspiral-merger-ringdown template (Ajith et al. 2008):

$$|\tilde{h}(f)| = \sqrt{\frac{5}{24}} \frac{\mathcal{M}_z^{\frac{5}{6}}}{\pi^{\frac{2}{3}} D_L} \times \begin{cases} f^{-\frac{7}{6}} & f < f_{\rm merg} \\ f_{\rm merg}^{-\frac{1}{2}} f^{-\frac{2}{3}} & f_{\rm merg} \le f < f_{\rm ring} \\ f_{\rm merg}^{-\frac{1}{2}} f_{\rm ring}^{-\frac{2}{3}} \frac{\sigma^2}{4(f - f_{\rm ring})^2 + \sigma^2} & f_{\rm ring} \le f < f_{\rm cut}, \end{cases} \tag{10}$$

where the redshift-dependent chirp mass, $\mathcal{M}_z$, is given in terms of the total mass , M, and the symmetric mass ratio $\eta \equiv m_1 m_2 / M^2$ of the binary by $\mathcal{M}_z \equiv (1+z)\mathcal{M}_0 = (1+z) M \eta^{\frac{3}{5}}$, and D_L denotes the luminosity distance of the binary. This implies the canonical $\Omega_{\rm GW} \propto f^{\frac{2}{3}}$ scaling during the inspiral phase (Phinney 2001). The frequencies $f_{\rm merg}$, $f_{\rm ring}$, $f_{\rm cut}$, and σ are parameterised

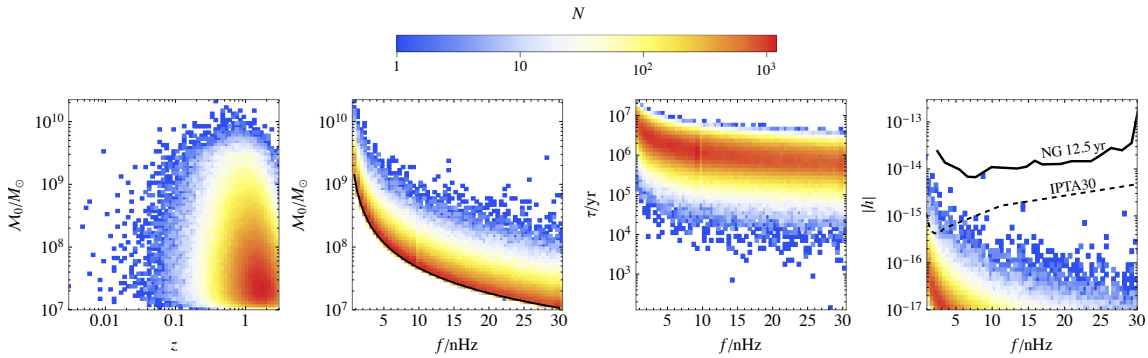


Fig. 3. Distributions of the binary parameters for a sampling of the BH binary population with $0 < z < 3$, $1\,\mathrm{nHz} < f < 30\mathrm{nHz}$ and $\mathcal{M}_0 > 10^9 (f/\mathrm{nHz})^{-\frac{4}{3}}$. The symmetric mass ratio is integrated with the lower bound $\eta > 0.001$. In the right panel, the black curve shows the NANOGrav 95% CL constraint on continuous GW sources (Arzoumanian et al. 2023) and the projected sensitivity of IPTA30 at SNR = 8.

as $f_j = \eta^{\frac{5}{3}}(a_j\eta^2 + b_j\eta + c_j)/(\pi\mathcal{M}_z)$ where a_j, b_j, and c_j are coefficients whose fitted values are given in Table I of Ajith et al. (2008).

We comment briefly on potential uncertainties in the GW energy density spectrum. If orbital decay is driven by other processes in addition to GW emission, such as viscous drag, then the coalescence time is shortened and the total GW spectrum is suppressed by a factor of $T_{\rm tot}/T_{\rm GW}$, where $T_{\rm tot}$, $T_{\rm GW}$ denote the characteristic hardening timescales (Kocsis & Sesana 2011; Kelley et al. 2017a). Here we omit this effect, assuming that the signal is dominated by binaries for which the hardening is driven mainly by GW emission. Additionally, we consider only circular binaries, which, in the inspiral phase, radiate monochromatically at twice the orbital frequency. We note, however, that the GW spectrum of eccentric binaries would contain higher harmonics, with the fundamental harmonic (at the orbital frequency) becoming dominant at eccentricities $e > 0.4$ (Taylor et al. 2016). In addition, the GW luminosity would be enhanced by a factor of $(1 + 73/24e^2 + 37/96e^4)/(1 - e^2)^{\frac{7}{2}}$ (Peters & Mathews 1963; Enoki & Nagashima 2007; Kelley et al. 2017b).

Figure 2 shows the GW energy density spectrum from BH binaries in different mass ranges. The black dashed curve corresponds to $p_{\rm BH} = 1$, namely, the assumption that each galactic merger would produce a BH merger with $m_2 > m_1 > 10^3\,M_\odot$. This naive assumption leads to a spectrum exceeding that observed in the PTA band.

We stress that the spectrum in Eq. (8), which is displayed in Fig. 2, does not necessarily correspond to a stochastic GW background as it does not distinguish between individual resolvable events and unresolvable events that would contribute to a GW background. The $\Omega_{\rm GW} \propto f^{\frac{2}{3}}$ tail at low frequencies arises from nearly monochromatic signals generated by inspiralling BH binaries. Because of this, $\Omega_{\rm GW}$ will fluctuate depending on the specific realization of the binary population, so the spectrum in Fig. 2 should be interpreted as the mean $\Omega_{\rm GW}$ obtained by averaging over many potential realizations of the binary population. We return to the probability distribution of $\Omega_{\rm GW}$ in the next subsection and in Appendix A.

The grey band in Fig. 2 was obtained by choosing a fixed value of $p_{\rm BH}$ so that the amplitude of the total GW spectrum from BH binaries matches the IPTA observations in the nHz range, which requires $p_{\rm BH} = 0.17^{+0.18}_{-0.08}$. The colored bands show the contributions to the total IPTA spectrum from different ranges of the BH masses. In comparing the dark blue, light blue, and green bands, we can see that the largest contribution to the GW spectrum in the PTA window comes from $M > 10^9\,M_\odot$ SMBH binaries, consistent with earlier studies (Wyithe & Loeb 2003; Sesana et al. 2004, 2008, 2009; Enoki et al. 2004; Kelley et al. 2017a; Izquierdo-Villalba et al. 2021; Bécsy et al. 2022).

Some characteristics of these binaries are shown in Fig. 3. The left panel shows that the total masses are $\lesssim 10^{10}$, and the redshifts are typically $O(1)$. The middle left panel shows that the typical masses increase with decreasing frequency and the middle right panel shows that the binary lifetimes are only weakly correlated with their emission frequencies, with lifetimes $\gtrsim 10^4$ yr being generally favored. The right panel shows the GW strain from these binaries, together with the upper limit on continuous GW sources from NANOGrav 12.5 years data (Arzoumanian et al. 2023) and the projected sensitivity of IPTA30 (Chen et al. 2017; Kaiser & McWilliams 2021) at a signal-to-noise ratio of SNR = 8, which is similar to that expected for SKA (Janssen et al. 2015).

Returning to Fig. 2, we see that the GW energy spectrum in the LISA frequency range is dominated by lighter BH binaries in the mass range of $M \in (10^6, 10^9)\,M_\odot$, shown in light blue, and the contribution in the AEDGE band is dominated by IMBHs with masses in the range of $M \in (10^3, 10^6)\,M_\odot$, shown in green. The low-frequency GW emissions from the early stages of the inspirals of these binaries also contribute in the lower-frequency bands, but these contributions are subdominant compared to the signals from higher-mass SMBH mergers[4].

We recall that the estimated spectrum in Fig. 2 assumes a constant $p_{\rm BH}$. Clearly, the relative contribution from binaries with lighter BH masses $<10^9\,M_\odot$ could be enhanced (reduced) by increasing (decreasing) $p_{\rm BH}$ for these binaries. In order for mergers of BHs with masses in the range $(10^6, 10^9)\,M_\odot$ to dominate the PTA signal $p_{\rm BH} \simeq 1$ would be required in this mass

[4] In addition to the experiments shown in Fig. 2, we have also considered the prospective sensitivities to GWs using the astrometric data from the *Gaia* and *Nancy Grace Roman* space telescopes (Wang et al. 2021). Their nominal sensitivities lie well above the black dashed line in Fig. 2 corresponding to $p_{\rm BH} = 1$, but a possible improvement of the Roman sensitivity might enable the detection of SMBH binary signals at frequencies $\in (10^{-7}, 10^{-6})$ Hz. Moreover, while the prospective power-law integrated sensitivity to GWs through binary resonance by laser ranging of the Moon (Blas & Jenkins 2022) reaches the gray band shown in Fig. 2, the most likely signal at $f \sim 10^{-6}$ Hz is well below that (see Sect. 3.2) and consists of a few nearly monochromatic sources.

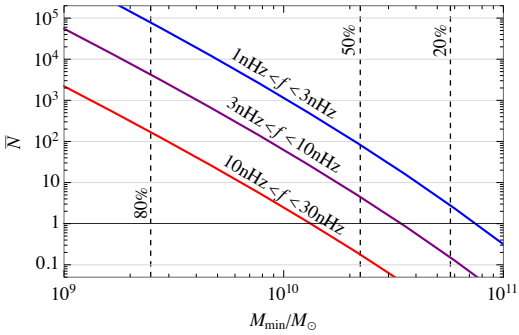


Fig. 4. Expected number of SMBH binaries heavier than $M_{\rm min}$ and with symmetric mass ratio $\eta > 0.01$ emitting GWs in the indicated frequency range for $p_{\rm BH} = 0.17$. The vertical dashed lines indicate the fractions of the total GW signal in each frequency band generated by binaries with larger masses.

range, but even this maximal enhancement would be insufficient for mergers with $M < 10^6\,M_\odot$ to contribute significantly to the PTA signal. That said, we emphasize that the GW spectrum in either the LISA or AEDGE frequency range could reach the black dashed line if $p_{\rm BH} \simeq 1$ in the relevant mass range, and could also be enhanced if Population III stars make a significant contribution to the BH spectrum. Conversely, the spectrum could be suppressed in these ranges if $p_{\rm BH}$ is smaller than the value that fits the PTA data.

3.2. Distribution of sources

The number of BH binaries at redshift z emitting in a given frequency band can be estimated from the time the binary spends in that frequency band, which gives (Sesana et al. 2008)

$$\frac{{\rm d}N}{{\rm d}\mathcal{M}_z {\rm d}\eta {\rm d}z {\rm d}\ln f} = \frac{{\rm d}\lambda}{{\rm d}\mathcal{M}_z {\rm d}\eta {\rm d}z}\left|\frac{{\rm d}\tau}{{\rm d}\ln f}\right|, \tag{11}$$

where, assuming circular orbits and orbital decay via GW emission (Peters 1964), the coalescence time of an inspiralling binary emitting GWs with frequency $(1+z)f$ is[5]

$$\tau(f) = \frac{5}{256\pi^{\frac{8}{3}}\mathcal{M}_z^{\frac{5}{3}}} f^{-\frac{8}{3}}. \tag{12}$$

We show in Fig. 4 the expected numbers of binaries with total masses of $M \geq M_{\rm min}$ and mass ratios $q \geq 0.01$ that emit GWs in several frequency bands in the range $1\,{\rm nHz} < f < 30\,{\rm nHz}$ (corresponding to the sensitivity ranges of PTAs). As seen in Fig. 2, SMBH binaries generate the dominant contribution to the GW signal in this frequency range. The fractions of the total GW signal in each frequency band that are generated by binaries with larger masses are indicated by the vertical dashed lines. For example, we see that almost 10^5 (10^2) binaries with $M > 3\times10^9\,M_\odot$ generate 80% of the calculated GW signal in the $1\,{\rm nHz} < f < 3\,{\rm nHz}$ ($10\,{\rm nHz} < f < 30\,{\rm nHz}$) frequency band, whereas 50% of the calculated GW signal in the $3\,{\rm nHz} < f < 10\,{\rm nHz}$ band may be generated by just one source.

[5] We remark that, in Eq. (11), the observed time spent in a given frequency band is redshifted when compared to the time in the binary rest frame. This redshift has been accounted for in Eq. (12).

These examples indicate that the expected signal, particularly in higher-frequency bins, may be comprised of a limited number of nearly monochromatic signals from heavy SMBH binaries ($>10^9\,M_\odot$), implying sizeable fluctuations around the smooth $f^{\frac{2}{3}}$ spectrum that would be obtained in the limit of a large population of heavy inspiralling binaries.

In order to study the prospective SMBH binary population in the PTA band via a Monte Carlo approach, we generated realizations of the population from the probability density function of binaries emitting at a given frequency. We divide the frequency range into bins (f_j, f_{j+1}) populated with $N(f_j)$ binaries. A realization of the GW background from each bin can then be obtained as (see Eq. (A.6)):

$$\Omega_{\rm GW}(f_j) = \frac{1}{\ln(f_{j+1}/f_j)} \sum_{k=1}^{N(f_j)} \Omega^{(1)}_{\rm GW}(\boldsymbol{\theta}_k), \tag{13}$$

where the contribution from an individual binary emitting in this frequency band is given by

$$\Omega^{(1)}_{\rm GW}(\boldsymbol{\theta}) = \left|\frac{{\rm d}\tau}{{\rm d}\ln f}\right|^{-1} \frac{2\pi}{5}\frac{f^3|\tilde{h}(f)|^2}{\rho_c}, \tag{14}$$

and $\boldsymbol{\theta} \equiv \{\mathcal{M}_z, z, \eta, f\}$ are parameters describing the binary. The number of binaries $N(f_j)$ in this frequency bin is drawn from a Poisson distribution with the expectation value $\bar{N}(f_j) \equiv \int_{f_j}^{f_{j+1}} {\rm d}N$ and the binary parameters are generated randomly according to the distribution

$$p(\mathcal{M}_z, \eta, z, f) \propto \frac{{\rm d}N}{{\rm d}\mathcal{M}_z {\rm d}\eta {\rm d}z {\rm d}\ln f}. \tag{15}$$

This is the expression that was used to calculate the distributions of binary parameters shown in Fig. 3 for a sampling with $p_{\rm BH} = 0.17$ of 4×10^5 SMBH binaries with redshifts $0 < z < 3$, mass ratios $\eta > 0.01$, and GW frequencies $1\,{\rm nHz} < f < 30\,{\rm nHz}$. We display only binaries with chirp mass $\mathcal{M}_0 > 10^9 (f/{\rm nHz})^{-\frac{4}{3}}$, so that $\bar{N}(f_j) > O(10^4)$ for each 1 nHz frequency bin.

In order to generate the statistical distributions of $\Omega_{\rm GW}$, it is sufficient to consider a smaller set of parameters, and the distribution Eq. (11) for the full population can be reduced to a simpler distribution $P(\Omega^{(1)}, f)$ for individual sources $\Omega^{(1)}_{\rm GW}$ emitting at a frequency f, which can be expressed via two one-parameter functions depending on the merger rate model, as discussed in Appendix A.

The upper left panel of Fig. 5 illustrates the frequency spectra found in 1600 Monte Carlo realizations of the SMBH binary population drawn from a statistical distribution similar to that shown in Fig. 3. The black solid curve shows the mean of the spectra, the grey bands show the 1σ and 2σ CL regions of the GW spectra, and the red line is the spectrum found in one representative realization. We note the importance of fluctuations, particularly at higher frequencies where fewer SMBH binaries contribute, with individual binaries becoming distinguishable at frequencies $f \gtrsim 10\,{\rm nHz}$. We find that the median spectrum, shown by the dashed line, has a somewhat lower slope than the analytic result shown as the green line. Although the numerically obtained mean spectrum (black), lies below the analytic expectation, it will approach it slowly if the number of Monte Carlo realizations is increased. The tendency of the mean spectrum to lie along the upper side of the 1σ CL range is due to the fact that the spectrum has a long tail at high values of $\Omega_{\rm GW}$ generated by occasional nearby binaries: the median spectrum always

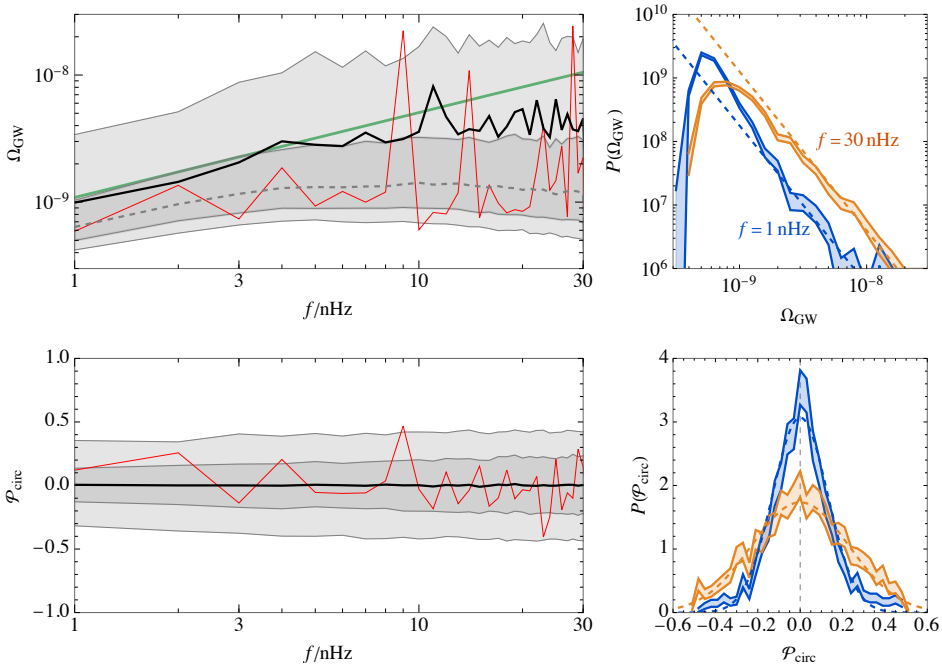


Fig. 5. GW energy spectrum $\Omega_{\rm GW}$ and the fractional circular polarization $\mathcal{P}_{\rm circ}$ calculated from 1600 Monte Carlo realizations of the SMBH binary population. In the left panels, the black solid curve shows the mean and the gray bands the 68% and 95% CL regions of the Monte Carlo realizations, and the red curve corresponds to one of the realizations shown in Fig. 3. In the upper left panel the green line shows the mean GW energy density (8). The distributions of $\Omega_{\rm GW}$ and $\mathcal{P}_{\rm GW}$ at 1 nHz (blue) and at 30 nHz (orange) calculated from the Monte Carlo realizations are shown in the right panels. In the upper right panel, the dashed lines show $P \propto \Omega^{-\frac{5}{2}}$, and in the lower right panel, the dashed curves show Gaussian fits to these distributions of width $\sigma = 0.13$ (blue) and $\sigma = 0.23$ (orange). In this figure, $p_{\rm BH} = 0.17$ is assumed.

lies within the 1σ CL range. All in all, consistently with earlier Monte Carlo studies of the GW signal (Sesana et al. 2008; Kocsis & Sesana 2011), we find that typical spectra tend to fall below the $\Omega_{\rm GW} \propto f^{\frac{2}{3}}$ expectation in higher frequency bins, while a few bins display sharp peaks.

As there are a finite number of sources, Eq. (8) is subject to statistical fluctuations, which arise mostly from the possibility of having a few strong sources nearby. To obtain an order-of-magnitude estimate, we focus on the closest binaries and ignore the redshift dependence. In this case, $\Omega^{(1)}_{\rm GW} \propto D_L^{-2}$, by Eqs. (8) and (10), and the probability of finding an event at D_L is $P^{(1)}(D_L) \propto D_L^2$. Thus $\langle\Omega^{(1)}_{\rm GW}\rangle \propto \int_{D_{L,\rm min}}^{D_{L,\rm max}} {\rm d}D_L P(D_L)\Omega^{(1)}_{\rm GW} \propto D_{L,\rm max}$, where, in this simplified approach, $D_{L,\rm max}$ is some large luminosity distance at which $\Omega^{(1)}_{\rm GW}$ gets suppressed, and $D_{L,\rm min}$ is the distance to the nearest possible source. On the other hand, $\langle(\Omega^{(1)}_{\rm GW})^2\rangle \propto 1/D_{L,\rm min}$. Therefore, we expect the mean of the GW signal to be determined by faraway sources, while the variance is set by a few close-by binaries. Moreover, we can estimate from $P^{(1)}(D_L)$ that $\Omega^{(1)}_{\rm GW}$ has a relatively flat power-law tail at large values,

$$P^{(1)}(\Omega) \propto \Omega^{-\frac{5}{2}}, \quad \text{when} \quad \Omega \to \infty. \tag{16}$$

Since the closest distance to massive BH binaries is constrained, this tail will be cut off at $\Omega^{(1)}_{\rm GW,max} \propto D^{-2}_{L,\rm min}$. Even with this cutoff, the mean and the variance are still not very useful characteristics of the uncertainties, and we find it more illuminating to estimate the confidence intervals around the median value.

The upper right panel of Fig. 5 displays the distributions of $\Omega_{\rm GW}$ at 1 nHz (blue) and at 30 nHz (orange) found in 1600 Monte Carlo realizations of the SMBH binary population. We note that the distributions at both frequencies have tails that approach the analytical result $P \propto \Omega^{-\frac{5}{2}}$ (Eq. (16)), as indicated by the dashed lines. These tails exhibit explicitly why the mean (black solid) line in Fig. 5 tends to lie above the 68% CL band, while the median (gray dashed) line lies within it at all frequencies. We note that the overall shapes of the distributions resemble the analytic calculations shown in the left panel of Fig. A.1. In summary, even if a bin has a high number of contributing events (i.e., $\bar{N} \to \infty$), the distribution of $\Omega_{\rm GW}$ does not converge to a Gaussian because it retains its $\Omega_{\rm GW}^{-5/2}$ tail. Since the variance is not well behaved, we define the width of the distribution $\Delta_{\Omega_{\rm GW}}$ as the width of the 68% confidence interval, as in Fig. 5. The width-to-mean ratio scales as:

$$\Delta_{\Omega_{\rm GW}}/\langle\Omega_{\rm GW}\rangle \propto \bar{N}^{-\frac{1}{3}}, \tag{17}$$

when $\bar{N} \gg 1$, as is shown in Appendix A. This scaling is slower than the typical $1/\sqrt{\bar{N}}$ scaling predicted by the central limit theorem.

3.3. GW polarization

Recent studies have argued that the circular polarization of the signal can be used to estimate whether the SGWB comes from a handful of sources or a relatively large population of binaries (Kato & Soda 2016; Conneely et al. 2019;

Hotinli et al. 2019; Belgacem & Kamionkowski 2020; Sato-Polito & Kamionkowski 2022; Valbusa Dall'Armi et al. 2023). We recall that the left and right circular GW polarization amplitudes from a binary with inclination angle θ are

$$|h_{L,R}(f)|^2 = \frac{1}{16}(1 \pm \cos\theta)^4 \, |h(f)|^2, \tag{18}$$

where

$$|h(f)| = \frac{4\mathcal{M}_z^{\frac{5}{3}}}{D_L}(\pi f)^{\frac{2}{3}} \tag{19}$$

is the maximal GW strain from an inspiralling binary, and the gravitational Stokes parameters are defined by

$$\begin{aligned} I(f) &= |h_L(f)|^2 + |h_R(f)|^2, \\ V(f) &= |h_L(f)|^2 - |h_R(f)|^2. \end{aligned} \tag{20}$$

The fractional amount of circular polarization of the SGWB can be characterized by the quantity

$$\mathcal{P}_{\rm circ}(f) = \frac{\sum_i V_i(f)}{\sum_i I_i(f)}, \tag{21}$$

where the sums are over all binaries in a fixed frequency range.

If the GW signal is dominated by a single source, then $\mathcal{P}_{\rm circ}$ depends only on the inclination angle and large circular polarizations are preferred, with $\mathcal{P}_{\rm circ} > 0.87(0.2)$ at the 68% CL (95% CL). On the other hand, if the GW signal is dominated by several ($N_{\rm dom} \gtrsim 10$) sources of comparable strengths, then the $\mathcal{P}_{\rm circ}(f)$ will be approximately Gaussian with a width determined by fluctuations in σ_V. Since $\sigma_V/\langle I\rangle_\theta = 1.17$ for a single source, we can estimate that (see Appendix A for details)

$$\sigma_{\mathcal{P}_{\rm circ}} \approx \sigma_V/\langle I\rangle_\theta \approx 1.17/\sqrt{N_{\rm dom}}. \tag{22}$$

The lower-left panel of Fig. 5 illustrates the distributions of the circular polarization $\mathcal{P}_{\rm circ}(f)$ found in the sample of 1600 Monte Carlo realizations of the SMBH binary population with $M > 10^9\,M_\odot$ and $\eta > 0.01$ whose frequency spectra were illustrated in Fig. 5. We see that the mean value of $\mathcal{P}_{\rm circ}(f) \approx 0$, as expected, but large statistical fluctuations are possible even at the 1σ level. This phenomenon is visible in the red curve, which shows results from one of the Monte Carlo realizations. The large fluctuations reflect the fact that the observable GW spectrum could be due to a very limited number of sources, particularly at higher frequencies.

The lower-right panel of Fig. 5 displays the distributions of $\mathcal{P}_{\rm circ}$ at 1 nHz (blue) and at 30 nHz (orange) found in 1600 Monte Carlo realizations of the SMBH binary population. We note that the statistical distribution is indeed broader at the higher frequency, as expected. The dashed curves are Gaussian fits to polarization distributions with widths $\sigma = 0.13$ (blue) and $\sigma = 0.23$ (orange). These fits are very accurate as also demonstrated in the right panel of Fig. A.1.

As the variance of the circular polarization of the signal depends on the number of binaries that dominate the signal, it would be suppressed if the merger rate of the heaviest binaries were suppressed, which could be the case if $p_{\rm BH}$ is not universal. By suppressing $p_{\rm BH}$, for instance, at $M > 10^9 M_\odot$ and enhancing $p_{\rm BH}$ for $10^9 M_\odot > M > 10^6 M_\odot$, we can accommodate the IPTA common-spectrum effect. In this case, the GW energy spectrum $\Omega_{\rm GW}$ becomes smoother and approaches the naive $\Omega_{\rm GW} \propto f^{\frac{2}{3}}$ behavior while suppressing the circular polarization $\mathcal{P}_{\rm circ}$. Moreover, in this case, $p_{\rm BH}$ for the lighter binaries would be enhanced, which would increase the number of signals in the sensitivity range of LISA and possibly AEDGE.

3.4. Prospects for future GW observatories

As discussed earlier, BH binaries can generate two qualitatively distinct classes of signal: long, nearly monochromatic signals from the slowly-evolving inspiralling phase and relatively short signals from the merger and ringdown. The detections of these two types of sources need to be considered separately.

The number of detectable nearly monochromatic GW signals that arise from inspiralling BH binaries is

$$N_{\rm insp} = \int_{\mathcal{T}+\tau_{\rm min}}^{\infty} {\rm d}\tau \int {\rm d}\lambda \, p_{\rm det}\left[\frac{{\rm SNR}_c}{{\rm SNR}(\tau, z)}\right], \tag{23}$$

where we account only for binaries whose coalescence time τ is longer than $\tau_{\rm min} = 1$ day and we estimate the SNR for a detector characterized by the noise power spectrum $S_n(f)$ as

$${\rm SNR}(\tau, z) = \sqrt{\int_0^{\mathcal{T}} {\rm d}t \, \frac{2|h(\tau - t, z)|^2}{S_n(f(\tau - t, z))}}. \tag{24}$$

The optimal time-dependent inspiral strain $|h(t)| = |h(f(t))|$ is given by Eq. (19) and the GW frequency as a function of time is given by Eq. (12). Analogously, the expected number of BH binary merger events is

$$N_{\rm merg} = \mathcal{T} \int {\rm d}\lambda \, p_{\rm det}\left[\frac{{\rm SNR}_c}{{\rm SNR}(z)}\right], \tag{25}$$

where

$${\rm SNR}(z) = \sqrt{\int_{f(\tau_{\rm min})}^{\infty} {\rm d}f \, \frac{4|\tilde{h}(f)|^2}{S_n(f)}}. \tag{26}$$

The cut at $f(\tau_{\rm min})$ here implies that we only consider the SNR only from the last day of the signal. In both cases, the detection probability $p_{\rm det}$ accounts for the detector's antenna patterns and includes the average over the binary inclination, sky location, and polarization (Finn & Chernoff 1993; Gerosa et al. 2019), the noises S_n include the foregrounds from stellar mass BH binaries and white dwarf binaries (see e.g., Lewicki & Vaskonen 2021). Furthermore, we used ${\rm SNR}_c = 8$ for the detection threshold and $\mathcal{T} = 1$ yr for the observation time.

In the monochromatic limit, Eq. (24) is simplified to ${\rm SNR} = |h(f)|\sqrt{2\mathcal{T}/S_n(f)}$. Using this, we can convert the prospected noise $S_n(f)$ of IPTA30 to a lower bound on the strain $|h(f)|$ for which the SNR exceeds the detection threshold. As indicated by the very long coalescence times in the middle right panel of Fig. 3, the signals in the IPTA30 band can be adequately approximated as monochromatic. In the right panel of Fig. 3 the dashed black curve shows the GW strain $|h(f)|$ that for 20 year observation time with IPTA30 gives SNR = 8. We see that IPTA30 can potentially resolve several, $O(10)$, monochromatic GW signals from SMBH binaries.

At higher frequencies, $f > 10^{-6}$ Hz, we find that LISA is the only detector capable of measuring the near-monochromatic sources arising in the early stages of the inspiral phase, namely, more than one day from the beginning of the merger phase. It is clear that AEDGE is not able to observe these near-monochromatic sources because the coalescence time of binaries heavier than $10^3\,M_\odot$ is less than a day when they enter the AEDGE sensitivity window[6]. Figure 6 illustrates the prospects

[6] In this work we have cut the AEDGE sensitivity at 3 mHz. This cut is motivated by the potential Newtonian gravity backgrounds (Hogan et al. 2011; Graham et al. 2017). We note, however, that dedicated studies of these backgrounds for AEDGE have not been performed.

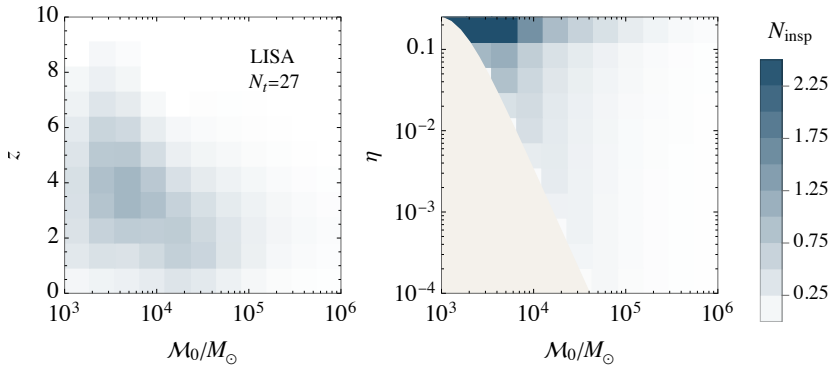


Fig. 6. Expected numbers of near-monochromatic GW signals generated more than a day before the merger that would be detectable in a year of LISA observation, as functions of the chirp mass $\mathcal{M}_0$ and redshift z (left panel) and of the chirp mass and symmetric mass ratio η (right panel). The merger rates assume the abundance of IMBHs described by Eq. (6) for BHs heavier than $10^3\,M_\odot$ and a universal merger efficiency factor of $p_{\rm BH} = 0.17$ inferred from IPTA data, as discussed in the text.

to detect the inspiral signals with LISA. The shading in the panels correspond to the number of events whose parameters fall within each rectangular bin. The distribution peaks below $\mathcal{M}_0 = 10^4 M_\odot$ reflecting the binary population, that increases towards lower masses, and the cut-off at $10^3\,M_\odot$. Similarly, because the GW signal (Eq. (10)) from inspirals does not explicitly depend on η, the η distribution in the right panel of Fig. 6 arises solely from the binary population. The binary merger rate peaks at $4 < z < 6$ (see Fig. 1) but LISA cannot spot the majority of the binaries beyond $z \sim 4$.

We find that, at $z < 10$, there are a total of $O(10^7)$ near-monochromatic sources in the frequency range of $10^{-5}\,{\rm Hz} < f < 0.1\,{\rm Hz}$[7]. As only a small fraction of these can be resolved by LISA, we expect that the rest constitute a significant stochastic GW background. It should be noted that the potential redshift dependence of $p_{\rm BH}$ introduces additional uncertainties in the merger rate at higher redshifts ($z \gtrsim 3$), which is not well probed by the PTA measurements. We leave a detailed study of this background and its detectability with LISA for future work.

Figure 7 illustrates AEDGE and LISA prospects for observing GW events from BHs less than a day after the merger. The expected total number of detectable events is slightly larger for LISA than for AEDGE. Due to the different frequency ranges, LISA can spot heavier binaries, $\mathcal{M}_0 > 10^4\,M_\odot$, while AEDGE can see lighter binaries, $\mathcal{M}_0 < 2\times10^3\,M_\odot$. LISA could (in principle) detect even heavier mergers, namely, of $\mathcal{M}_0 \lesssim 10^8\,M_\odot$, but since such mergers are so rare it is unlikely that LISA will see any mergers above $10^6\,M_\odot$. The expected number of detectable mergers peaks at $4 < z < 6$ for both of the experiments. For AEDGE, the z distribution of the detectable mergers reflects the binary merger rate (see Fig. 1) while for LISA the z distribution peaks at slightly lower z because the SNR of $\mathcal{M}_0 \sim 10^4$ binaries for LISA is not as high as it is for AEDGE. The lower panels show that, for all detectors, the BH mass ratio peaks at the highest values. For AION-km and ET we find that the total expected number of events per year is less than three.

We note that most of the IMBH events detectable by AEDGE during the last day prior to the merger (see the left panels in Fig. 7) will also have been detected by LISA during the previous infall stage, as seen in Fig. 6. This opens up prospects for using LISA data to predict when and in what direction AEDGE will observe IMBH mergers, sharpening tests of general relativity and giving advance warnings for searches for possible multi-messenger signals.

We caution that there are observational uncertainties in the low mass cut-off in Eq. (6), due to the difficulty of measuring very faint active galactic nuclei (AGNs) or inactive BHs in dwarf galaxies. As mentioned above, surveys such as eRASS and AMUSE (Miller et al. 2015) are already constraining this low-mass region and are compatible with the assumed cut-off, although a heavier mass cut-off may be favored (Chadayammuri et al. 2023). This could easily be achieved in models with modified initial fluctuation spectra (Hütsi et al. 2023).

The cut-off mass is tightly related to the SMBH formation mechanism since a BH in the halo centre cannot be lighter than the seed that originated the growth. For a full review, we refer to Volonteri et al. (2021). In this sense, the cut-off we have considered assumes the existence of some light seeds, which augment the possible GW signal. Independently of the growth through halo merging, IMBHs can also be born inside dense stellar media like nuclear and globular star clusters. The growth from stellar masses happens because of repeated encounters inside these dense environments. Estimating the GW emission from such encounters is an active field of research (Fragione & Kocsis 2018; Fragione et al. 2019; Fragione & Loeb 2023) and their signals may contribute significantly to the expected number of IMBH mergers.

4. Conclusions

Here, we describe the results from a model for the PTA nHz common-process signal based on a simulation of massive BH mergers. The magnitude of the signal depends on the merger probability, $p_{\rm BH}$, which is a product of the probabilities that a pair of halos contain massive BHs and the probability that they will merge. We sidestepped the considerable uncertainties in modeling these probabilities by fitting a mass-independent value of $p_{\rm BH}$ to the PTA signal, finding $p_{\rm BH} = 0.17$, with a factor of two uncertainty. With this assumption, the dominant contribution to the PTA signal is made by mergers with total masses $>10^9\,M_\odot$ and about 10% from masses $<10^9\,M_\odot$. The PTA mergers would have redshifts $O(1)$ and have mass asymmetries $\gtrsim$10.

[7] We emphasize, however, that there could be lower-mass BHs that are remnants of Population III stars, which are not included in our analysis.

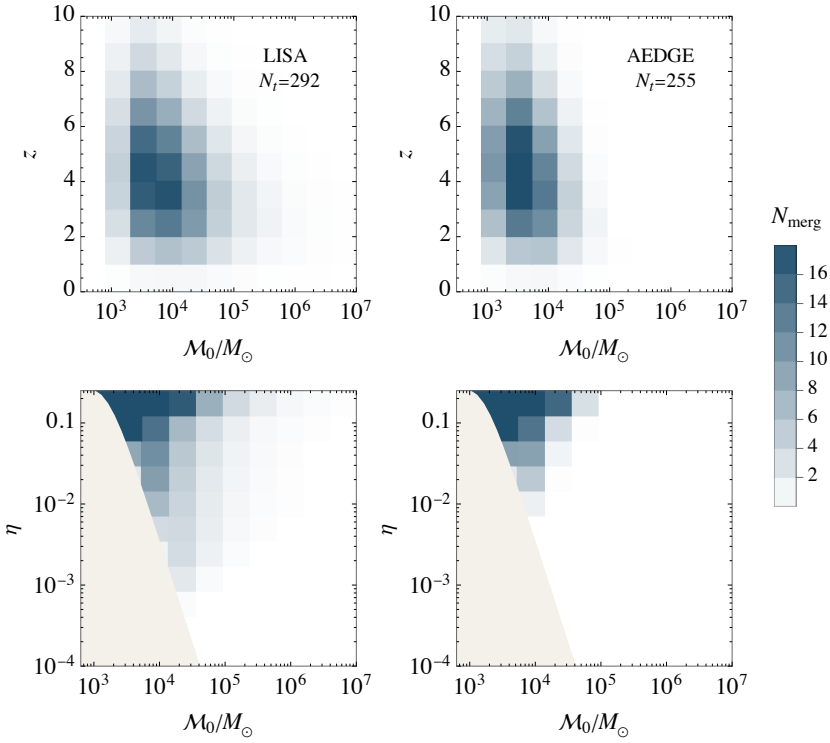


Fig. 7. Expected numbers of detectable GW events generated during the last day before the merger in a year of observation by LISA (left panels) and AEDGE (right panels) as functions of the chirp mass $\mathcal{M}_0$ and redshift z (upper panels) and of chirp mass and symmetric mass ratio η (lower panels). The merger rates assume the abundance of IMBHs described by Eq. (6) for BHs heavier than $10^3\,M_\odot$ and a universal merger efficiency factor of $p_{\rm BH} = 0.17$ inferred from the IPTA data.

The number of mergers contributing most of the PTA signal at frequencies $O(10)$ nHz is limited. Consequently, the frequency spectrum becomes quite irregular, the spectral index may deviate from the analytic value of 2/3, individual mergers may be distinguished, and there may be detectable circular polarization. These will be interesting targets for future experiments in the nHz range, including PTAs and SKA.

Assuming the same mass-independent value of $p_{\rm BH}$ as for the PTA signal, there would be observable signals from mergers with total masses $\in(10^3, 10^6)\,M_\odot$ in the LISA experiment and from mergers with total masses $\in(10^3, 10^5)\,M_\odot$ in the AEDGE experiment. Data from these experiments will be able to check the accuracy of the constancy of $p_{\rm BH}$. In principle, their signals could be even larger than our estimates if $p_{\rm BH}$ is closer to unity for masses below the PTA range, or if there is a significant GW contribution seeded by Population III stars. However, a lower value for $p_{\rm BH}$ and, hence, a lower event rate cannot be excluded.

Acknowledgements. This work was supported by European Regional Development Fund through the CoE program grant TK133 and by the Estonian Research Council grants PRG803 and PSG869. The work of J.E. was supported by the United Kingdom STFC Grant ST/T000759/1. The work of V.V. was supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101065736.

References

Ajith, P., Babak, S., Chen, Y., et al. 2008, Phys. Rev. D, 77, 104017
Antoniadis, J., Arzoumanian, Z., Babak, S., et al. 2022, MNRAS, 510, 4873
Arzoumanian, Z., Baker, P. T., Blumer, H., et al. 2020, ApJ, 905, L34
Arzoumanian, Z., Baker, P. T., Blecha, L., et al. 2023, ApJL, submitted [arXiv:2301.03608]
Ashoorioon, A., Rezazadeh, K., & Rostami, A. 2022, Phys. Lett. B, 835, 137542
Badurina, L., Bentine, E., Blas, D., et al. 2020, JCAP, 05, 011
Badurina, L., Buchmueller, O., Ellis, J., et al. 2021, Phil. Trans. A. Math. Phys. Eng. Sci., 380, 20210060
Barkana, R., & Loeb, A. 2001, Phys. Rept., 349, 125
Bécsy, B., Cornish, N. J., & Kelley, L. Z. 2022, ApJ, 941, 119
Begelman, M. C., Blandford, R. D., & Rees, M. J. 1980, Nature, 287, 307
Belgacem, E., & Kamionkowski, M. 2020, Phys. Rev. D, 102, 023004
Benson, A. J., Kamionkowski, M., & Hassani, S. H. 2005, MNRAS, 357, 847
Berti, E., Cardoso, V., & Will, C. M. 2006, Phys. Rev. D, 73, 064030
Blanco-Pillado, J. J., Olum, K. D., & Wachter, J. M. 2021, Phys. Rev. D, 103, 103512
Blas, D., & Jenkins, A. C. 2022, Phys. Rev. Lett., 128, 101103
Blasi, S., Brdar, V., & Schmitz, K. 2021, Phys. Rev. Lett., 126, 041305
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Boroson, T. A., & Lauer, T. R. 2009, Nature, 458, 53
Buchmuller, W., Domcke, V., & Schmitz, K. 2020, Phys. Lett. B, 811, 135914
Chadayammuri, U., Bogdan, A., Ricarte, A., & Natarajan, P. 2023, ApJ, 946, 51
Chen, S., Middleton, H., Sesana, A., Del Pozzo, W., & Vecchio, A. 2017, MNRAS, 468, 404
Chen, S., Caballero, R. N., Guo, Y. J., et al. 2021, MNRAS, 508, 4970
Conneely, C., Jaffe, A. H., & Mingarelli, C. M. F. 2019, MNRAS, 487, 562
De Luca, V., Franciolini, G., & Riotto, A. 2021, Phys. Rev. Lett., 126, 041303
De Paolis, F., Ingrosso, G., & Nucita, A. A. 2002, A&A, 388, 470
De Paolis, F., Ingrosso, G., & Nucita, A. A. 2004, A&A, 426, 379
Dodelson, S. 2003, Modern Cosmology (Amsterdam: Academic Press)
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
El-Neaj, Y. A., Alpigiani, C., Amairi-Pyka, S., et al. 2020, EPJ Quant. Technol., 7, 6
Ellis, J., & Lewicki, M. 2021, Phys. Rev. Lett., 126, 041304
Enoki, M., & Nagashima, M. 2007, Prog. Theor. Phys., 117, 241

Enoki, M., Inoue, K. T., Nagashima, M., & Sugiyama, N. 2004, ApJ, 615, 19
Erickcek, A. L., Kamionkowski, M., & Benson, A. J. 2006, MNRAS, 371, 1992
Event Horizon Telescope Collaboration (Akiyama, K., et al.) 2019, ApJ, 875, L1
Event Horizon Telescope Collaboration (Akiyama, K., et al.) 2022, ApJ, 930, L12
Finn, L. S., & Chernoff, D. F. 1993, Phys. Rev. D, 47, 2198
Fragione, G., & Kocsis, B. 2018, Phys. Rev. Lett., 121, 161103
Fragione, G., & Loeb, A. 2023, ApJ, 944, 81
Fragione, G., Grishin, E., Leigh, N. W. C., Perets, H. B., & Perna, R. 2019, MNRAS, 488, 47
Gerosa, D., Ma, S., Wong, K. W. K., et al. 2019, Phys. Rev. D, 99, 103004
Goncharov, B., Shannon, R. M., Reardon, D. J., et al. 2021, ApJ, 917, L19
Graham, M. J., Djorgovski, S. G., Stern, D., et al. 2015, Nature, 518, 74
Graham, P. W., Hogan, J. M., Kasevich, M. A., Rajendran, S., & Romani, R. W. 2017, ArXiv e-prints [arXiv:1711.02225]
Greene, J. E., Strader, J., & Ho, L. C. 2020, ARA&A, 58, 257
Hellings, R. W., & Downs, G. S. 1983, ApJ, 265, L39
Hogan, J. M., Johnson, D. M. S., Dickerson, S., et al. 2011, Gen. Rel. Grav., 43, 1953
Holgado, A. M., Sesana, A., Sandrinelli, A., et al. 2018, MNRAS, 481, L74
Hotinli, S. C., Kamionkowski, M., & Jaffe, A. H. 2019, Open J. Astrophys., 2, 8
Hütsi, G., Raidal, M., Urrutia, J., Vaskonen, V., & Veermäe, H. 2023, Phys. Rev. D, 107, 043502
Iguchi, S., Okuda, T., & Sudou, H. 2010, ApJ, 724, L166
Izquierdo-Villalba, D., Sesana, A., Bonoli, S., & Colpi, M. 2021, MNRAS, 509, 3488
Janssen, G., Hobbs, G., McLaughlin, M., et al. 2015, PoS, AASKA14, 037
Kaiser, A. R., & McWilliams, S. T. 2021, CQG, 38, 055009
Kato, R., & Soda, J. 2016, Phys. Rev. D, 93, 062003
Kelley, L. Z., Blecha, L., & Hernquist, L. 2017a, MNRAS, 464, 3131
Kelley, L. Z., Blecha, L., Hernquist, L., Sesana, A., & Taylor, S. R. 2017b, MNRAS, 471, 4508
King, O. G., Hovatta, T., Max-Moerbeck, W., et al. 2013, MNRAS, 436, 114
Kocsis, B., & Sesana, A. 2011, MNRAS, 411, 1467
Kohri, K., & Terada, T. 2021, Phys. Lett. B, 813, 136040
Kovačević, A. B., Songsheng, Y.-Y., Wang, J.-M., & Popovic, L. C. 2022, A&A, 663, A99
Lacey, C. G., & Cole, S. 1993, MNRAS, 262, 627
Lewicki, M., & Vaskonen, V. 2021, Eur. Phys. J. C, 83, 168
LIGO Scientific & Virgo Collaborations (Abbott, B. P., et al.) 2016, Phys. Rev. Lett., 116, 061102
LIGO Scientific, Virgo & KAGRA Collaborations (Abbott, R., et al.) 2021, ArXiv e-prints [arXiv:2111.03634]
Lippai, Z., Frei, Z., & Haiman, Z. 2009, ApJ, 701, 360
LISA Collaboration (Amaro-Seoane, P., et al.) 2017, ArXiv e-prints [arXiv:1702.00786]
Merritt, D. 2013, CQG, 30, 244005
Miller, B. P., Gallo, E., Greene, J. E., et al. 2015, ApJ, 799, 98
O'Neill, S., Kiehlmann, S., Readhead, A. C. S., et al. 2022, ApJ, 926, L35
Peters, P. C. 1964, Phys. Rev., 136, B1224
Peters, P. C., & Mathews, J. 1963, Phys. Rev., 131, 435
Phinney, E. S. 2001, ArXiv e-prints [arXiv:astro-ph/0108028]
Planck Collaboration VI. 2020, A&A, 641, A6
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Rieger, F. M., & Mannheim, K. 2000, A&A, 359, 948
Sathyaprakash, B., Abernathy, M., Acernese, F., et al. 2012, CQG, 29, 124013
Sato-Polito, G., & Kamionkowski, M. 2022, Phys. Rev. D, 106, 023004
Sesana, A., Haardt, F., Madau, P., & Volonteri, M. 2004, ApJ, 611, 623
Sesana, A., Vecchio, A., & Colacino, C. N. 2008, MNRAS, 390, 192
Sesana, A., Vecchio, A., & Volonteri, M. 2009, MNRAS, 394, 2255
Taylor, S. R., Huerta, E. A., Gair, J. R., & McWilliams, S. T. 2016, ApJ, 817, 70
Valbusa Dall'Armi, L., Nishizawa, A., Ricciardone, A., & Matarrese, S. 2023, Phys. Rev. Lett., 131, 041401
Valtonen, M. J., et al. 2008, Nature, 452, 851
Vaskonen, V., & Veermäe, H. 2021, Phys. Rev. Lett., 126, 051303
Vogelsberger, M., Genel, S., Springel, V., et al. 2014, Nature, 509, 177
Volonteri, M., Habouzit, M., & Colpi, M. 2021, Nat. Rev. Phys., 3, 732
Wang, Y., Pardo, K., Chang, T.-C., & Doré, O. 2021, Phys. Rev. D, 103, 084007
Wyithe, J. S. B., & Loeb, A. 2003, ApJ, 590, 691

Appendix A: Statistics of the GW energy spectrum and circular polarization

A.1. Distribution of $\Omega_{\rm GW}(f)$

The GW energy spectrum arising from a population of BH binaries can be expressed as the sum of the contributions from individual binaries. We first consider the contribution to $\Omega_{\rm GW}$ coming from a single inspiralling binary:

$$\Omega^{(1)}_{\rm GW}(\boldsymbol{\theta}) = \left|\frac{{\rm d}\tau}{{\rm d}\ln f}\right|^{-1} \frac{2\pi}{5}\frac{f^3|\tilde{h}(f)|^2}{\rho_c}\,, \tag{A.1}$$

where $\boldsymbol{\theta} \equiv \{\mathcal{M}_z, z, f\}$ are parameters describing the binary. The GW energy spectrum arising from the inspiralling binary population (or any of its subpopulations) arises from the sum of nearly monochromatic components

$$\Omega_{\rm GW}(f) \approx \sum_i \Omega^{(1)}_{\rm GW}(\boldsymbol{\theta}_i)\delta(\ln f/f_i)\,. \tag{A.2}$$

This approximation is valid as long as the change in frequency over the observation period is smaller than the spectral resolution. The parameters $\boldsymbol{\theta}_i$ as well as $\Omega^{(1)}_{\rm GW}$ are independent and identically distributed. Thus, the statistical properties of $\Omega_{\rm GW}(f)$ can be inferred from the distribution of $\Omega^{(1)}_{\rm GW}(f)$, which is given by

$$\begin{aligned} P^{(1)}(\Omega|f) &= \frac{1}{n(f)}\int {\rm d}\lambda \left|\frac{{\rm d}\tau}{{\rm d}\ln f}\right| \delta\left(\Omega - \Omega^{(1)}_{\rm GW}\right) \\ &= \frac{\mathcal{F}(\Omega f^{-\frac{10}{3}})}{f\Omega^{\frac{3}{2}} n(f)}\,, \end{aligned} \tag{A.3}$$

where we define the following function:

$$\begin{aligned} \mathcal{F}(x) \equiv \int {\rm d}z{\rm d}\eta \Bigg[&\frac{\partial\lambda(\mathcal{M}_z, z)}{\partial\mathcal{M}_z\partial\eta\partial z} \\ &\times \frac{\mathcal{M}_z}{16\sqrt{10\rho_c}D_L}\Bigg]_{\mathcal{M}_z=\frac{1}{\pi}\left(\frac{5\pi}{8}D_L^3\rho_c x\right)^{\frac{3}{10}}} \end{aligned} \tag{A.4}$$

and the spectral source density:

$$n(f) \equiv \int {\rm d}\mathcal{M}_z{\rm d}\eta{\rm d}z \frac{{\rm d}N}{{\rm d}\mathcal{M}_z{\rm d}\eta{\rm d}z{\rm d}\ln f}\,. \tag{A.5}$$

We see that, for circular inspiralling binaries, the four-parameter distribution (11) can be reduced to two functions of a single parameter. Further reductions are unlikely, as these functions depend on the model of the binary merger rate, which we assume to stem from the merger rate of galaxies.

The average energy spectrum in a frequency bin (f_j, f_{j+1}) can be obtained from

$$\begin{aligned} \Omega_{\rm GW}(f_j) &\equiv \frac{1}{\ln(f_{j+1}/f_j)}\int_{f_j}^{f_{j+1}} \Omega_{\rm GW}(\boldsymbol{\theta}; f){\rm d}\ln f \\ &= \frac{1}{\ln(f_{j+1}/f_j)}\sum_{i=1}^{N(f_j)} \Omega^{(1)}_{{\rm GW},i}\,, \end{aligned} \tag{A.6}$$

where the number of binaries $N(f_j)$ is drawn from a Poisson distribution with the expected value,

$$\bar{N}(f_j) = \int_{f_j}^{f_{j+1}} {\rm d}\ln f\, n(f)\,, \tag{A.7}$$

and the $\Omega^{(1)}_{{\rm GW},i}$ are drawn from a random distribution $P^{(1)}(\Omega, f) = P^{(1)}(\Omega|f)n(f)/\bar{N}(f_j)$ with $f \in (f_j, f_{j+1})$. In the following, we suppress f_j.

The moment generating function of $\Omega_{\rm GW}$ is given by:

$$\begin{aligned} M_{\Omega_{\rm GW}}(s) &\equiv \langle \exp\left(s\Omega_{\rm GW}\right)\rangle \\ &= \sum_{N\geq 0} p_N \prod_{i=1}^{N}\left\langle \exp\left(s\Omega^{(1)}_{\rm GW}\right)\right\rangle \\ &= \sum_{N\geq 0}\frac{\bar{N}^N}{N!}e^{-\bar{N}}\left(M_{\Omega^{(1)}_{\rm GW}}(s)\right)^{\bar{N}} \\ &= \exp\left[\bar{N}\left(M_{\Omega^{(1)}_{\rm GW}}(k) - 1\right)\right]\,. \end{aligned} \tag{A.8}$$

Thus, the cumulant generating function of $\Omega_{\rm GW}(f_j)$, that is,

$$\begin{aligned} K_{\Omega_{\rm GW}}(s) &\equiv \ln M_{\Omega_{\rm GW}}(s) \\ &= \int_{f_j}^{f_{j+1}} {\rm d}N\left[\exp(s\,\Omega^{(1)}_{\rm GW}) - 1\right]\,. \end{aligned} \tag{A.9}$$

is proportional to the generating function of $M_{\Omega^{(1)}_{\rm GW}}(s)$, implying that the nth cumulant of $\Omega_{\rm GW}$ is

$$\kappa_n[\Omega_{\rm GW}] = \bar{N}\langle(\Omega^{(1)}_{\rm GW})^n\rangle = \int_{f_j}^{f_{j+1}} {\rm d}N(\Omega^{(1)}_{\rm GW})^n\,. \tag{A.10}$$

Therefore, the mean $\kappa_1(\Omega_{\rm GW}) \equiv \langle\Omega_{\rm GW}\rangle$ matches the expectation value given in Eq. (8), while the variance,

$$\kappa_2(\Omega_{\rm GW}) \equiv \langle(\delta\Omega_{\rm GW})^2\rangle = \bar{N}\langle(\Omega^{(1)}_{\rm GW})^2\rangle \tag{A.11}$$

is divergent due to the long $P^{(1)}(\Omega) \propto \Omega^{-\frac{5}{2}}$ tail (16), unless a minimal distance to the closest BH binary is imposed. In particular, as we will demonstrate briefly, this long tail is preserved in the distribution of $\Omega_{\rm GW}$, i.e., the higher cumulants are not diminished when summing several instances of $P^{(1)}(\Omega) \propto \Omega^{-\frac{5}{2}}$.

The moment-generating function $M_{\Omega_{\rm GW}}(-s)$ is a Laplace transform of the probability distribution. Thus as an alternative to the Monte Carlo approach adopted in the main text, the probability distribution of $\Omega_{\rm GW}(f_j)$ can be obtained by an inverse Laplace transform of the moment generating function given in Eq. (A.8). It can be expressed as

$$\begin{aligned} P(\Omega_{\rm GW}) = \frac{1}{2\pi\bar{N}}\int_{-\infty}^{\infty} {\rm d}s\, e^{is\Omega_{\rm GW}/\bar{N}}\times \\ \times \exp\left(\bar{N}\left\langle\exp\left(-is\Omega^{(1)}_{\rm GW}/\bar{N}\right) - 1\right\rangle\right)\,, \end{aligned} \tag{A.12}$$

where the average is taken over the single event distribution $P^{(1)}$.

The large N limit. – The central limit theorem (and Eq. (A.11)) dictates that when $N \to \infty$, then the distribution (A.6) should approach a Gaussian with its relative width scaling as $\sigma_{\Omega_{\rm GW}}/\langle\Omega_{\rm GW}\rangle \propto 1/\sqrt{N}$. However, this is not the case when the second cumulant diverges. Considering a generic distribution with a tail (droping the "GW" subindex for the sake of brevity):

$$P^{(1)}(\Omega) \overset{\Omega\to\infty}{\sim} C\Omega^{-\frac{5}{2}}, \tag{A.13}$$

where C is a constant. The second moment $\langle\Omega^2\rangle$ diverges. We assume that the small Ω behavior is such that the first moment $\bar{\Omega}^{(1)} \equiv \langle\Omega^{(1)}\rangle$ is finite. The $\bar{N} \to \infty$ asymptotic of the average in Eq. (A.12) can then be computed by separating it into the mean

and a term for which the expectation value can be determined by approximating the distribution by its tail (A.13):

$$\begin{aligned}&\bar{N}\left\langle\exp\left(-is\frac{\Omega^{(1)}}{\bar{N}}\right)-1\right\rangle\\&=-is\bar{\Omega}^{(1)}+\bar{N}\left\langle\exp\left(-is\frac{\Omega^{(1)}}{\bar{N}}\right)-1+is\frac{\Omega^{(1)}}{\bar{N}}\right\rangle\\&\overset{N\to\infty}{\sim}-is\bar{\Omega}^{(1)}+e^{i\pi\frac{3}{4}}\frac{\sqrt{\pi}4C}{3\sqrt{\bar{N}}}s^{\frac{3}{2}}\,.\end{aligned}\tag{A.14}$$

In summary, the $N\to\infty$ distribution asymptotes to

$$P(\Omega)\approx(C\bar{N})^{-\frac{2}{3}}\mathcal{P}\left((C\bar{N})^{-\frac{2}{3}}(\Omega-\bar{\Omega})\right),\tag{A.15}$$

where $\bar{\Omega}\equiv\bar{N}\bar{\Omega}^{(1)}$ denotes the expectation value of Ω and

$$\mathcal{P}(x)\equiv\mathrm{Re}\frac{e^{i\pi/6}}{\pi}\int_0^\infty \mathrm{d}s\,\exp\left(ie^{i\pi/6}xs-\frac{4\sqrt{\pi}}{3}s^{\frac{3}{2}}\right)\tag{A.16}$$

is a universal function that is independent of the single event distribution – all information about the latter enters through $\bar{\Omega}^{(1)}$ and C. This limiting case reveals a few crucial features:

1. Since $\mathcal{P}(x)\sim x^{-5/2}$ when $x\to\infty$, the power-law behavior of the tail is preserved in the large $\bar{N}$ limit:
$$P(\Omega)\overset{\Omega\gg\bar{\Omega}}{\approx}\frac{C\bar{N}}{(\Omega-\bar{\Omega})^{\frac{5}{2}}}\,.\tag{A.17}$$
2. When the number of contributing events N is increased by a factor of A to AN, the distribution evolves as
$$P^{(AN)}(\Omega)=A^{-\frac{2}{3}}P^{(N)}\left(A^{-\frac{2}{3}}(\Omega-\bar{\Omega}^{(AN)})+\bar{\Omega}^{(N)}\right).\tag{A.18}$$
3. Although the variance diverges, we can define the width of the distribution $\Delta_\Omega\equiv\Omega_+-\Omega_-$ from the confidence interval (Ω_-,Ω_+) centred around, e.g., the median. Eq. (A.18) then implies the scaling
$$\Delta_\Omega\propto\bar{N}^{2/3}\,.\tag{A.19}$$
So, the relative width $\Delta_\Omega/\bar{\Omega}$ will approach zero, but at a slower pace than when the central limit theorem applies.
4. Using the method of steepest decent, we find that
$$\mathcal{P}(x)\overset{\Omega\ll\bar{\Omega}}{\approx}\frac{\sqrt{\bar{\Omega}-\Omega}}{2\pi C\bar{N}}e^{-\frac{(\bar{\Omega}-\Omega)^3}{12\pi(C\bar{N})^2}}\,,\tag{A.20}$$
implying values smaller than the mean, that is $\Omega<\bar{\Omega}$, are much less likely than for Gaussian distributions, as they are suppressed by an exponent of a cube.

As an illustration, we consider a population of sources for which individual signals follow the simple long-tailed distribution

$$P^{(1)}(\Omega)\propto\Omega^{-\frac{5}{2}}\theta(\Omega-\Omega_{\min}),\tag{A.21}$$

which mimics the statistics of $\Omega_{\rm GW}$ from a realistic BH population found in section 3.2. In Fig. A.1, we show the resulting distributions of $\Omega/\bar{\Omega}$ if the signal consists of 10, 100, and 10^5 sources. The mean is given by $\bar{\Omega}\equiv 3N\Omega_{\min}$. The distribution is computed by generating random realizations of the source populations (shown by points) and by using (A.12) (shown by solid lines). These two approaches are in excellent agreement. As expected, one can observe that the total signal will inherit the long $\Omega^{-5/2}$ tail from the distribution of individual sources (shown by the dashed line). At large Ω, the tail approaches Eq. (A.17) in all cases, while the asymptotic (A.15) works well for $N=10^5$, but predicts a slightly too wide peak for $N\leq 10^5$. As predicted, the peak gets narrower and moves closer to the expectation value, when N is increased. We note, however, that although the shape of the simplified distribution in Fig. A.1 resembles the distribution in Fig. 5 obtained from a model of the heavy BH binary population, the shape around the peak in the latter depends also on the z dependence and the binary mass distribution.

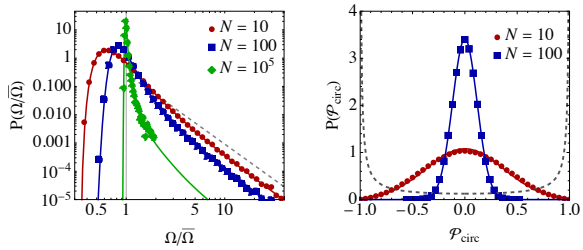


Fig. A.1. Distributions of the total signal strengths from 1, 10, 100, and 10^5 sources whose independent signal strengths have identical $\Omega^{-\frac{5}{2}}$ tails (left). Distribution of $\mathcal{P}_{\rm circ}$ for 1, 10 and 100 identical sources (right). The dashed line shows the distribution from a single source, the solid lines show the analytic estimates, and the distributions obtained from an explicit Monte Carlo analysis are shown by the points.

A.2. Distribution of $\mathcal{P}_{\rm circ}$

Considering now the distribution of the circular polarization

$$\mathcal{P}_{\rm circ}\equiv\frac{\sum_{i=1}^N V_i}{\sum_{i=1}^N I_i}\,,\tag{A.22}$$

which depends on the distribution and number N of sources. We denote the contributions from individual sources by V_i, I_i given by Eqs. (18-20). As above, we are working in a fixed frequency bin and, thus, suppressing the frequency dependence in the arguments of the functions.

The polarization will vanish on average, that is, at $\langle\mathcal{P}_{\rm circ}\rangle=0$, but polarization measurements can provide information from the size of the deviation from that expectation. For a single dominant source, the contribution from the amplitude drops out and we find

$$\mathcal{P}_{\rm circ}=\frac{4c_\theta(1+c_\theta^2)}{1+6c_\theta^2+c_\theta^4}\,,\tag{A.23}$$

where $c_\theta\equiv\cos(\theta)$. The corresponding probability distribution is shown as the dashed curve in Fig. A.1 and can be seen to be dominated by the region $|\mathcal{P}_{\rm circ}|\approx 1$.

If the signal is dominated by several sources, the analytic treatment of $\mathcal{P}_{\rm circ}$ is complicated due to correlations between V and I. The mean and variance of I and V are given by

$$\begin{aligned}\langle I\rangle&=\frac{4}{5}\langle|h|^2\rangle\,,&\langle V\rangle&=0\,,\\\langle\delta I^2\rangle&=\frac{284}{315}\langle(\delta|h|^2)^2\rangle\,,&\langle\delta V^2\rangle&=\frac{92}{105}\langle(\delta|h|^2)^2\rangle\,,\end{aligned}\tag{A.24}$$

where we used the fact that the inclination is independent of other parameters of the binary and the fluctuations in $|h|^2$ are computed as in (A.11). The latter is expected to be large. To

estimate the fluctuations in $\mathcal{P}_{\rm circ}$ in the limit when the number of sources is large, we first assume that I can be replaced by its average over inclinations $\langle I\rangle_\theta \equiv (4/5)\sum_i |h_i|^2$, but not over signal strengths, so that

$$\mathcal{P}_{\rm circ} \approx V/\langle I\rangle_\theta \equiv \sum_{i=1}^{N} p_i f_V(c_{\theta,i})\,, \tag{A.25}$$

where $p_i \equiv (4/5)|h_i|^2/\langle I\rangle_\theta \in [0,1]$ is the fractional contribution to the signal from source i and $f_V(c_\theta) = (5/4)c_\theta(1+c_\theta^2)$ contains the inclination dependence of V. In this way, it is possible to contain the effect of the large fluctuations in $\langle I\rangle_\theta$ in the random fractions p_i, which vary only in the range $[0,1]$. We find:

$$\langle \delta\mathcal{P}^2_{\rm circ}\rangle = \frac{115}{84} N\langle |p_i|^2\rangle\,. \tag{A.26}$$

We note that $N\langle|p_i|^2\rangle$ takes values in the range $[1/N, 1]$ and could be interpreted as a measure of the effective number of dominant sources $1/N_{\rm dom}$. For instance, in the limiting case where the signal is sourced by N binaries that contribute equally, that is, $p_i = 1/N$, we would obtain

$$\sigma_{\mathcal{P}_{\rm circ}} = 1.17/\sqrt{N}\,, \tag{A.27}$$

as also stated in Eq. (22). In the right panel of Fig. A.1, approximate Gaussian distributions with a width given by (A.27) and shown as solid lines are compared to distributions obtained from explicitly generated random populations of $\mathcal{P}_{\rm circ}$ arising from exactly 10 and 100 sources of equal strength (shown by points). We see an excellent agreement between the naive Gaussian approximation and the Monte Carlo estimate already for $N = 10$.

Finally, we remark that the approximation (A.25) above can be improved by expanding $\mathcal{P}_{\rm circ}$ in $\delta I \equiv I - \langle I\rangle_\theta$, that is, $\mathcal{P}_{\rm circ} = V/\langle I\rangle_\theta(1-\delta I/\langle I\rangle_\theta + \ldots)$. This allows us to account for correlations between V and δI, for instance, when computing the variance (A.26),

$$\langle \delta\mathcal{P}^2_{\rm circ}\rangle = \frac{115}{84} N\langle |p_i|^2\rangle - \frac{2855}{1386} N\langle |p_i|^3\rangle + \ldots \tag{A.28}$$

The corrections arising from higher powers of $\delta I/\langle I\rangle_\theta$ are suppressed by increasing powers of $1/N$.

Appendix 7

VII
J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data. *Phys. Rev. D*, 109(2):L021302, 2024

Letter

Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data

John Ellis,[1,2,3,*] Malcolm Fairbairn,[2,†] Gert Hütsi,[1,‡] Juhan Raidal,[1,§] Juan Urrutia,[1,4,‖] Ville Vaskonen,[1,5,6,¶] and Hardi Veermäe[1,**]

[1]*Keemilise ja Bioloogilise Füüsika Instituut, Rävala puiestee 10, 10143 Tallinn, Estonia*
[2]*Physics Department, King's College London, Strand, London, WC2R 2LS, United Kingdom*
[3]*Theoretical Physics Department, CERN, CH 1211 Geneva, Switzerland*
[4]*Department of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia*
[5]*Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy*
[6]*Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 12 July 2023; accepted 30 November 2023; published 10 January 2024)

The NANOGrav and other pulsar timing arrays (PTAs) have recently announced evidence for nHz gravitational waves (GWs) that may originate from supermassive black hole (SMBH) binaries. The spectral index of the GW signal differs from that predicted for binary evolution by GW emission alone, and we show that environmental effects such as dynamical friction with gas, stars, and dark matter improve the consistency of the SMBH binary model with the PTA data. We comment on the possible implications of environmental effects for PTA observations of fluctuations in the GW frequency spectrum and measurements of GWs at higher frequencies.

DOI: 10.1103/PhysRevD.109.L021302

Introduction. Several ongoing pulsar timing array (PTA) projects—the North American Nanohertz Observatory for Gravitational Waves (NANOGrav), the European PTA, the Parkes PTA, and the Chinese PTA—have recently released their latest data [1–4]. Most importantly, the NANOGrav Collaboration has found evidence [5] in NANOGrav 15-year (NG15) data for the Hellings-Downs quadrupolar correlation [6] (see also [4,7,8])—a key characteristic of gravitational waves (GWs)—in the common-spectrum process observed previously by them [9] and other PTAs [10–12]. This is a landmark in GW astronomy, of significance comparable to the first indirect detection of GWs emitted by binary pulsars [13] and the first direct observation of GW emissions from stellar-mass black hole (BH) binaries [14]. This discovery of a nHz stochastic GW background opens a new window on astrophysical processes that were previously unobserved. We assume here that the origin of the PTA signal is a population of supermassive black hole (SMBH) binaries (see also [15,16]) and explore the astrophysical implications of this possibility.[1]

Accretion around SMBHs drives active galactic nuclei (AGNs) and is thought to play a role in the formation of SMBH binaries [59,60]. At an early stage, the binary evolution is driven by dynamical friction with gas, stars, and dark matter [61]. However, a prerequisite for SMBH mergers is the formation of tightly bound binaries that can radiate GWs efficiently. It is not yet understood how this stage of SMBH binary evolution is reached, an issue commonly known as the "final parsec problem" [61]. If the binaries can overcome it, an important source of evolution at subparsec scales is energy loss via GW emission until it is the main driver of the evolution close to merging. SMBH binaries driven only by GW emission are naively expected to produce an almost flat background with a spectral index $\gamma = 13/3$ [62]. The recent NG15 data [5] on GWs, although compatible with this value at the 99% CL, prefer a lower value of γ. This hints that the GWs may be emitted by SMBH binaries that are experiencing additional energy loss via environmental effects, which

*john.ellis@cern.ch
†malcolm.fairbairn@kcl.ac.uk
‡gert.hutsi@kbfi.ee
§juhan.raidal@student.manchester.ac.uk
‖juan.urrutia@kbfi.ee
¶ville.vaskonen@pd.infn.it
**hardi.veermae@cern.ch

[1]Numerous interpretations of the NANOGrav 12.5-year data [9] based on cosmological sources have been put forward, including cosmic strings and domain walls [17–27], first-order phase transitions [28–37], and primordial fluctuations [38–57]. Many of these models with a spectral slope $\gamma > 13/3$ are disfavored by the current data [58], for instance, cosmic strings or primordial black hole seeds for SMBHs.

 © 2024 American Physical Society

might be related to whatever dynamics overcomes the final parsec problem.

In this Letter, we follow [63] in using the extended Press-Schechter (EPS) formalism to estimate the SMBH binary formation and merger rates. We go beyond [63] by including scatter in the SMBH-halo mass relation and by modifying the evolution of the SMBH binaries to include energy loss via environmental effects as well as GWs. Our analysis improves existing SMBH interpretations [15] of the NG15 data by accurately modeling the long-tailed statistical fluctuations in the GW spectrum without relying on Gaussian approximations. We find that the best fit to the NG15 data [5] is given by a phenomenological model of SMBH binary evolution that includes environmental energy loss, with a log-likelihood difference of 10 compared to purely GW-driven evolution, and hence is favored by more than 2σ. Within this framework, we find that the efficiency for the production of SMBH mergers $p_{\rm BH}$ is probably quite high and close to the maximal allowed value. These results provide important constraints on astrophysical scenarios for the interlinked dynamics of SMBHs and their host galaxies, and suggest promising rates for observing mergers of lower-mass BHs in detectors sensitive to GWs with higher frequencies.

SMBH background with environmental effects. The mean of the GW energy density spectrum generated by a population of SMBH binaries can be estimated as [62]

$$\Omega_{\rm GW}(f) \equiv \frac{1}{\rho_c}\frac{{\rm d}\rho_{\rm GW}}{{\rm d}\,\ln f} = \frac{1}{\rho_c}\int {\rm d}\lambda \frac{1+z}{4\pi D_L^2}\frac{{\rm d}E_{\rm GW}}{{\rm d}\,\ln f_r}, \quad (1)$$

where D_L denotes the source luminosity distance, ${\rm d}E_{\rm GW}/{\rm d}\,\ln f_r$ gives the energy emitted by the source per logarithmic frequency interval, and $f_r \equiv (1+z)f$ denotes the frequency in the source frame, i.e., at the time of emission. The differential BH merger rate in the observer frame is

$${\rm d}\lambda = {\rm d}\mathcal{M}{\rm d}\eta{\rm d}z\frac{1}{1+z}\frac{{\rm d}V_c}{{\rm d}z}\frac{{\rm d}R_{\rm BH}}{{\rm d}\mathcal{M}{\rm d}\eta}, \quad (2)$$

where $\mathcal{M}$ denotes the binary chirp mass, η its symmetric mass ratio, V_c the comoving volume, and $R_{\rm BH}$ the comoving BH merger rate density. It is given by

$$\frac{{\rm d}R_{\rm BH}}{{\rm d}m_1{\rm d}m_2} = p_{\rm BH}(m_1,m_2,z)\int {\rm d}M_1{\rm d}M_2\frac{{\rm d}R_h}{{\rm d}M_1{\rm d}M_2} \times p_{\rm occ}(m_1|M_1,z)p_{\rm occ}(m_2|M_2,z), \quad (3)$$

where $m_{1,2}$ are the masses of the merging BHs, $M_{1,2}$ are the masses of their host halos, R_h is the halo merger rate that we estimate with the EPS formalism [64–66], and $p_{\rm BH} \leq 1$ combines the SMBH occupation fraction in galaxies with the efficiency for the BHs to merge following the merging of their host halos. For simplicity, we assume a constant value for $p_{\rm BH}$, which we treat as a free parameter to be determined by fitting the NG15 data. As the GW background arises from a relatively narrow range of SMBH masses and redshifts, we expect that extending the parametrization by letting $p_{\rm BH}$ vary would have a minor effect on our conclusions. The halo mass-SMBH mass relation is encoded in $p_{\rm occ}$ that is the probability distribution of the SMBH mass.[2] We model the spread in the mass relation with a log-normal distribution:

$$p_{\rm occ}(m|M,z) = \frac{1}{\sqrt{2\pi}m\sigma}\exp\left[-\frac{\ln(m/\bar{m})^2}{2\sigma^2}\right] \quad (4)$$

with $\log_{10}(\bar{m}/M_\odot) = 8.95 + 1.4\log_{10}(M_*/10^{11}M_\odot)$ and $\sigma = 1.1$ following the fit to observations of inactive galaxies in [67,68], and relate the stellar mass M_* to the halo mass $M_{\rm halo}$ using the fit provided in [69]. The effect of using the fit to AGN and not including the scatter is discussed in the Supplementary Material [70].

The GW spectrum from an inspiraling binary is determined by its orbital evolution, and following NANOGrav [5] we assume circular binaries. The eccentricity of the orbits (considered, e.g., in [78]) would affect the GW spectrum by introducing higher harmonics, increasing the total power emitted in GWs, and modifying the frequency spectral index [79]. However, big eccentricities $e > 0.9$ would lead to an attenuation of the background due to the acceleration of the binary inspiral [80].

The binary may lose energy through a combination of GW emission and dissipative environmental effects:

$$\dot{E} = -\dot{E}_{\rm GW} - \dot{E}_{\rm env}, \quad (5)$$

where the dot denotes differentiation with respect to time. This energy loss causes orbital decay and growth in the orbital frequency. SMBH binaries are thought to go through several stages as they harden, starting with dynamical friction, followed by hardening due to close stellar encounters with stars populating the so-called "loss-cone" orbits [81], and finally, dissipation caused by viscous drag due to circumbinary gas disks [61]. The crossing of the final parsec is determined mostly by these last two effects, with the latter being the dissipation mechanism that is commonly thought to remain active after the GW emission has become significant (see, e.g., [82–84]), thus leading to faster binary hardening compared to the GW-only case.

The impact of the gas is not fully established; e.g., [85,86] argue that accretion from circumbinary disks might have quite the opposite effect; instead of hardening the binary, it might lead to the expansion of the orbit. More recent work predicts, however, inspiraling for more realistic cooler, thinner disks [87], nonzero eccentricities [88], and unequal masses [89]. Any slowing of the infall of binaries contributing in the low-frequency bins would worsen the

[2]Compared to Eq. (5) of Ref. [63], the normalization of $p_{\rm occ}$ is absorbed into $p_{\rm BH}$.

SMBH fit, whereas extra dissipation improves it, as we show below.

The characteristic timescales for the above processes are

$$t_{\rm GW} \equiv |E|/\dot{E}_{\rm GW} = 4\tau, \qquad t_{\rm env} \equiv |E|/\dot{E}_{\rm env}, \tag{6}$$

where $\tau = (5/256)(\pi f_r)^{-8/3}\mathcal{M}^{-5/3}$ denotes the coalescence time of the binary assuming GW emission alone. The effective timescale for the binary evolution is then $t_{\rm eff}^{-1} = t_{\rm GW}^{-1} + t_{\rm env}^{-1}$. Since the binding energy of the binary is $E = -(\pi f_r)^{2/3}\mathcal{M}^{5/3}/2$, where f_r is the frequency of the emitted GW, it follows that

$$\frac{{\rm d}\ln f_r}{{\rm d}t} = \frac{3}{2}t_{\rm GW}^{-1}\left(1 + \frac{t_{\rm GW}}{t_{\rm env}}\right), \tag{7}$$

and thus,

$$\frac{{\rm d}E_{\rm GW}}{{\rm d}\ln f_r} = \frac{{\rm d}E_{\rm GW}}{{\rm d}t}\frac{{\rm d}t}{{\rm d}\ln f_r} = \frac{1}{3}\frac{(\pi f_r)^{\frac{2}{3}}\mathcal{M}^{\frac{5}{3}}}{1 + t_{\rm GW}/t_{\rm env}}. \tag{8}$$

When $t_{\rm GW} \ll t_{\rm env}$, we obtain the usual spectrum for GW-driven coalescence. As the PTAs observe only a narrow range of frequencies, environmentally driven decay can be approximated by

$$\frac{t_{\rm env}}{t_{\rm GW}} = \left(\frac{f_r}{f_{\rm GW}}\right)^{\alpha_1+\alpha_2\ln f_r/f_{\rm GW}+\ldots}, \tag{9}$$

where $f_{\rm GW}$ denotes a reference frequency above which GW emission becomes dominant, and α_n are coefficients series expansion of $\ln t_{\rm env}/t_{\rm GW}$ in $\ln f_r$. At leading order, we can drop $\alpha_{n>1}$ terms and define $\alpha_1 \equiv \alpha$, so that the environmental effects are parametrized by a power-law dependence on the orbital frequency (cf. [90–93]) and choose[3]

$$f_{\rm GW}(\mathcal{M}, \eta, z) = f_{\rm ref}\left(\frac{\mathcal{M}}{10^9 M_\odot}\right)^{-\beta}, \tag{10}$$

where $f_{\rm ref}, \alpha > 0$ are phenomenological parameters that we constrain with the NG15 data (we note that $t_{\rm env}$ increases with f_r if $\alpha > 8/3$). Depending on the environmental mechanism for orbital decay, the range $\beta = 0.2$–0.8 has been considered in the literature [90,91,93]. Following Ref. [93], we fix $\beta = 0.4$. As the signal is generated by binaries in a relatively narrow BH mass range, we expect our results to be only weakly dependent on β.

Analysis. As reported by NANOGrav, a power-law fit of the form $\Omega_{\rm GW} = A(f/f_{\rm yr})^{5-\gamma}$ excludes the $\gamma = 13/3$ scaling naively expected for a background of SMBH binaries whose evolution is driven by GW emission at the 95% CL.[4] However, as shown in Fig. 1, due to the stochastic nature of the signal, the power law fails to capture many of its properties, and it could differ significantly from any one particular realization of the background. Moreover, due to the environmental effects, even the mean GW background is not a simple power law. Therefore, we use the full information provided in [5] for the probability distribution functions at each frequency bin $P_{\rm data}(\Omega, f_i)$ represented by the orange "violins" in Fig. 6 of Supplementary Material [70].

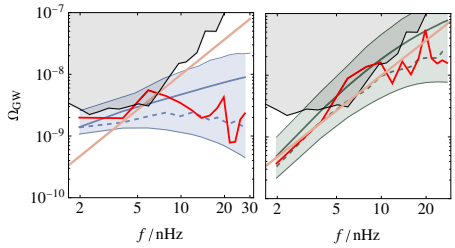


FIG. 1. The best-fit GW energy density for GW-driven binaries (left) and GW and environmentally driven effects (right), compared to the best power-law fit to the NG15 data (orange). The mean energy densities (1) are shown as full lines, and the most probable values as dashed lines, and in each case one random realization is shown in red and the 95% CL ranges are shaded. The gray curve shows the NG15 upper limit on signals from individually resolvable binaries [94].

Following [63], we generate Monte Carlo realizations of the stochastic GW background using the same frequency binning as the NANOGrav Collaboration, f_i. A realization of the SMBH background in a specific frequency bin is given by

$$\Omega_{\rm GW}(f_j) = \frac{1}{\ln(f_{j+1}/f_j)}\sum_{k=1}^{N(f_j)}\Omega_{\rm GW}^{(1)}(\vec{\theta}_b^k), \tag{11}$$

where $\vec{\theta}_b \equiv \{\mathcal{M}, z, \eta, f\}$ denotes the parameters of a binary, and $N(f_j)$ is drawn from a Poisson distribution determined by the expected number of binaries $\bar{N}(f_j) = \int_{f\in(f_j,f_{j+1})}{\rm d}\lambda{\rm d}\tau$ contributing to each bin. The contribution from an individual binary emitting at some frequency is

$$\Omega_{\rm GW}^{(1)}(f, \vec{\theta}_b) = \frac{1+z}{4\pi D_L^2\rho_c}\frac{{\rm d}E_{\rm GW}}{{\rm d}t}. \tag{12}$$

To decrease the computation time, rather than generating realizations of the binary population, we instead generate values of $\Omega_{\rm GW}(f_j)$ directly. Assuming that there are no correlations between the frequency bins, i.e., no binaries cross between bins during the period of observation, the

[3]This parametrization of the environmental effects does not account for the ratio between the masses of the components of the binary. However, the majority of the binaries are expected to have mass ratios close to unity [63]. An analogous power-law model was used in the NANOGrav analysis [5].

[4]As a consistency check, we repeated our analysis for a power-law ansatz and reproduced the posteriors reported by NANOGrav [5].

statistical properties of $\Omega_{\rm GW}(f)$ can be inferred from the distribution of $\Omega^{(1)}_{\rm GW}(f,\vec{\theta}_b)$, which is given by[5]

$$P^{(1)}(\Omega|f,\vec{\theta}_f) \propto \int {\rm d}\lambda \left|\frac{{\rm d}t}{{\rm d}\ln f_r}\right| \delta(\Omega - \Omega^{(1)}_{\rm GW}), \quad (13)$$

and depends only on the parameters of the merger rate and binary evolution $\vec{\theta}_f = (p_{\rm BH}, \alpha, \beta, f_{\rm ref})$.

Finally, the distribution $P(\Omega|f_j,\vec{\theta}_f)$ of the total GW energy density (11) in each bin is estimated by dividing the signal into two pieces: the signal from the strongest sources and the rest (see the Supplementary Material [70] for details). The first component is modeled using a Monte Carlo approach simulating 4×10^5 realizations of the signal from these strong sources using single event distributions (13). The rest of the signal follows a narrow Gaussian distribution and can be simply modeled by its average. The probability distribution of $\Omega_{\rm GW}(f_j)$ possesses a long tail which it inherits from $P^{(1)}$.[6] Such long tails are expensive to resolve with the Monte Carlo approach, so we use $P^{(1)}$ to construct the large Ω tail analytically. This improves the statistical modeling of $\Omega_{\rm GW}(f_j)$ at large amplitudes for which a single source is expected to dominate the signal. We represent the probability distributions of Ω with and without the environmental effects for the best-fit parameter values in Fig. 6 of Supplementary Material [70] by the green and blue violins.

Our approach provides an accurate and fast[7] way of resolving the distribution $P(\Omega|f_j,\vec{\theta}_f)$ of spectral fluctuations in each frequency bin. This makes it feasible to perform accurate scans over a wide range of model parameters $\vec{\theta}_f$, and is thus a step forward from earlier analyses that relied on Gaussian process interpolation for SMBH population synthesis [15] or a relatively small number of realizations of the full SMBH population [78].

The likelihood of a given model $\vec{\theta}_f$ is

$$l(f_i|\vec{\theta}_f) \propto \prod_i \int {\rm d}\Omega\, P_{\rm data}(f_i|\Omega) P(\Omega|f_i,\vec{\theta}_f), \quad (14)$$

where the effective number of variables of $\vec{\theta}_f$ depends on the hypothesis made for the population: If the energy loss due to GW emission dominates—corresponding to very slow environmental energy loss ($t_{\rm env}\to\infty$)—the only parameter is $p_{\rm BH}$, whereas environmental energy loss is characterized by two additional parameters, $f_{\rm ref}$ and α. Finally, to compare the fits, we define the likelihood ratio

$$\mathcal{C}_{\rm max} = \frac{\max_{\vec{\theta}_f} l(\vec{\theta}_f|H_1)}{\max_{\vec{\theta}_f} l(\vec{\theta}_f|H_2)}. \quad (15)$$

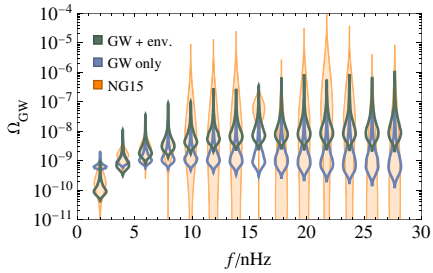


FIG. 2. The NG15 data on the energy density of GWs $\Omega_{\rm GW}$ (shown in orange) compared with the best fit assuming energy loss by GWs only (shown in blue), and the best fit including also energy loss by the environmental effects (shown in green). The posteriors for the parameters of the fits are shown in Fig. 3.

Results. We see in Fig. 2 that the NG15 data (shown in orange) are better fitted by the model including environmental effects (shown in green) than by the model where binaries evolve purely by emitting GWs (shown in blue). In particular, the fit including environmental energy loss captures better the dip in the lowest frequency bin at $f = 2$ nHz, whereas the GW-driven model has to balance the tilt by lowering the overall merging efficiency. We note that suppression of the signal at low frequencies is expected also when the binaries are eccentric [95].

We find that the best-fit value of the merging efficiency in the case of purely GW-driven binaries is $p_{\rm BH} = 0.07^{+0.05}_{-0.07}$, where the uncertainties are reported at the 68% CL. For the model including environmental effects, the fit prefers largervalues of $p_{\rm BH}$. In this case, we show in Fig. 3 the two-dimensional marginalized projections of the three-dimensional parameter space featuring the 68%, 95%, and 99% CL regions with different shades of green and the best fit as a white dot. The marginalized distributions of the three parameters are also shown. The best-fit values are $p_{\rm BH} = 0.84$, $\alpha = 2.0$, and $f_{\rm ref} \simeq 34$ nHz. We note, however, that there are significant correlations between the parameters, and the marginalized probability distributions are broad and asymmetric. The preferred value of α is consistent with previous model estimates and the value $\alpha = 8/3$ for which $t_{\rm env}$ would be frequency independent. The central value of $f_{\rm ref}$ corresponds to the apparent break in the spectral index seen in Fig. 6 of Supplementary Material [70].

The model that includes environmental effects gives a larger best-fit likelihood but also has more parameters, namely, $(p_{\rm BH}, \alpha, f_{\rm ref})$. The likelihood ratio with respect to the best fit for the purely GW-driven binaries (i.e., the limit

[5]The binary inclination can be exactly accounted for by including the inclination-angle-dependent prefactor in $\Omega^{(1)}_{\rm GW}$ and integrating over the inclination angle in $P^{(1)}$. We have checked that including the inclination angle in $P^{(1)}$ has a negligible effect, and thus we use the inclination angle averaged $\Omega^{(1)}_{\rm GW}$.

[6]It asymptotes to $P(\Omega|f,\vec{\theta}_f) \sim \Omega^{-\frac{5}{2}}$ as $\Omega\to\infty$, thus implying a divergent variance [63].

[7]It takes $\mathcal{O}(1)$ seconds per bin to compute $P(\Omega|f_j,\vec{\theta}_f)$ on an Apple M1 Pro 8-core processor.

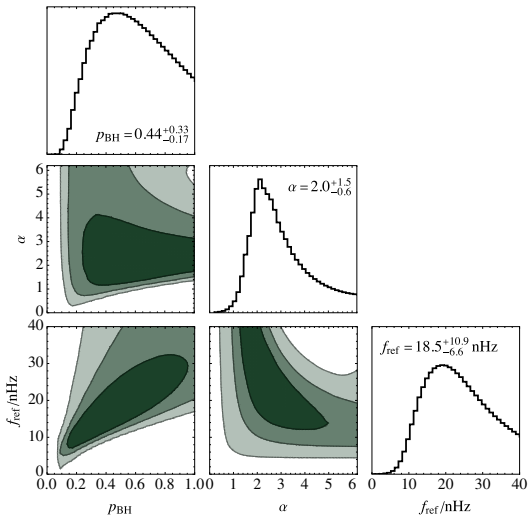


FIG. 3. Fit of the SMBH binary model with environmental effects parametrized by $f_{\rm ref}$, α, and $p_{\rm BH}$ to the NG15 data [5], for $\beta = 2/5$.

$f_{\rm ref} \to 0$) corresponds to $\Delta\chi^2 \sim -2\log\ell_{\rm max} = 11$. In a Gaussian approximation, since the difference in the number of degrees of freedom between the hypotheses is 2, the model including environmental effects is favored over the model with pure GW evolution at the 2σ CL.

We note that the lower bound on $\Omega_{\rm GW}$ in the eighth bin at $f = 16$ nHz is relatively high. Large upward fluctuations are a feature of the long-tailed distribution of the GW spectra from SMBHs [63] and are not penalized heavily in a likelihood analysis. Relatively nearby SMBH binaries can generate such peaks. Specifically, the best fits for the GW-only and GW + environment models shown in Fig. 1 predict that 0.44 and 0.85 SMBH binaries are expected to be resolved with the current NANOGrav sensitivity, respectively. Intriguingly, we find that in both cases the individually resolvable binaries are most likely to be found in the 4–7 nHz frequency range, which coincides with a potential monochromatic GW signal seen by NANOGrav [94] and EPTA [96].

Discussion. The PTA results open an exciting new way to probe the environments of SMBH binaries. We show in Fig. 4 the effective timescales for the evolution of the binaries that best fit the NG15 data. We see that at the smallest radii, the binaries evolve by emitting GWs, whereas at larger distances the binary timescales are close to being flat. This suggests that at these distances the main mechanism driving binary evolution is viscous drag. Simulations show that in this regime the combination with other effects, e.g., loss-cone scattering as well as GW emission, leads to a distribution that is close to flat,

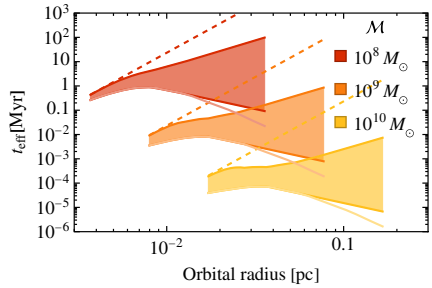


FIG. 4. Effective timescales $t_{\rm eff}$ for fits to the NG15 data shown as functions of the separations at which binaries are emitting GWs in the NANOGrav band (1–30 nHz), and for chirp masses $\mathcal{M} \in (10^8, 10^{10})M_\odot$, which is the range found in [63] to dominate the stochastic GW background. The dashed lines are for energy loss by GWs alone, and the shaded regions are those favored at the 68% CL for a combination of energy loss by environmental effects and GWs. The two lower lines are computed with $\alpha < 4$ and $\alpha < 6$ as priors.

although the uncertainties are large. In general, we find that the phenomenological fit is consistent with the expected evolution of binaries in galactic environments.

As discussed in [63], although there are large uncertainties, extrapolation of the model predicts that GWs from mergers of lower-mass BHs should be observable in the higher frequency ranges where the LISA mission [97] and proposed atom interferometers such as AION [98] and AEDGE [99] are most sensitive, as well as other projects targeting the dHz frequency band [100,101]. LISA would be most sensitive to mergers of BHs with masses $(10^3, 10^7)M_\odot$ and AEDGE to the $(10^2, 10^5)M_\odot$ mass range. Observations of binaries of such intermediate-mass BHs could cast light on the mechanism for the assembly of SMBHs as well as probe strong gravity in a range beyond the reach of terrestrial laser interferometers. To the extent that the results of this Letter suggest a high merger efficiency $p_{\rm BH} \sim 0.5$, the merger rates might be greater than was estimated in [63], improving the prospects for seeing inspiraling intermediate-mass BH binaries by LISA and AEDGE to $\mathcal{O}(10^3)$ and by km-scale terrestrial atom interferometers to $\mathcal{O}(10)$.

Acknowledgments. The work of G. H., V. V., and H. V. was supported by European Regional Development Fund through the Center of Excellence program Grant No. TK133 and by the Estonian Research Council Grants No. PSG869 and No. PRG803. The work of J. E. and M. F. was supported by the United Kingdom STFC Grants No. ST/T000759/1 and No. ST/T00679X/1. The work of V. V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101065736.

[1] G. Agazie *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **951**, L9 (2023).
[2] J. Antoniadis *et al.* (EPTA Collaboration), Astron. Astrophys. **678**, A48 (2023).
[3] A. Zic *et al.* (Parkes Pulsar Timing Array Collaboration), arXiv:2306.16230.
[4] H. Xu *et al.* (Chinese Pulsar Timing Array Collaboration), Res. Astron. Astrophys. **23**, 075024 (2023).
[5] G. Agazie *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **951**, L8 (2023).
[6] R. W. Hellings and G. S. Downs, Astrophys. J. Lett. **265**, L39 (1983).
[7] J. Antoniadis *et al.* (European Pulsar Timing Array Collaboration), Astron. Astrophys. **678**, A50 (2023).
[8] D. J. Reardon *et al.*, Astrophys. J. Lett. **951**, L6 (2023).
[9] Z. Arzoumanian *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **905**, L34 (2020).
[10] B. Goncharov *et al.* (Parkes Pulsar Timing Array Collaboration), Astrophys. J. Lett. **917**, L19 (2021).
[11] S. Chen *et al.* (European Pulsar Timing Array Collaboration), Mon. Not. R. Astron. Soc. **508**, 4970 (2021).
[12] J. Antoniadis *et al.* (International Pulsar Timing Array Collaboration), Mon. Not. R. Astron. Soc. **510**, 4873 (2022).
[13] R. A. Hulse and J. H. Taylor, Astrophys. J. Lett. **195**, L51 (1975).
[14] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[15] G. Agazie *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **952**, L37 (2023).
[16] J. Antoniadis *et al.* (European Pulsar Timing Array Collaboration), arXiv:2306.16227.
[17] J. Ellis and M. Lewicki, Phys. Rev. Lett. **126**, 041304 (2021).
[18] S. Datta, A. Ghosal, and R. Samanta, J. Cosmol. Astropart. Phys. 08 (2021) 021.
[19] R. Samanta and S. Datta, J. High Energy Phys. 05 (2021) 211.
[20] W. Buchmuller, V. Domcke, and K. Schmitz, Phys. Lett. B **811**, 135914 (2020).
[21] S. Blasi, V. Brdar, and K. Schmitz, Phys. Rev. Lett. **126**, 041305 (2021).
[22] W. Buchmuller, V. Domcke, and K. Schmitz, J. Cosmol. Astropart. Phys. 12 (2021) 006.
[23] J. J. Blanco-Pillado, K. D. Olum, and J. M. Wachter, Phys. Rev. D **103**, 103512 (2021).
[24] R. Z. Ferreira, A. Notari, O. Pujolas, and F. Rompineve, J. Cosmol. Astropart. Phys. 02 (**2023**) 001.
[25] H. An and C. Yang, arXiv:2304.02361.
[26] Z.-Y. Qiu and Z.-H. Yu, Chin. Phys. C **47**, 085104 (2023).
[27] Z.-M. Zeng, J. Liu, and Z.-K. Guo, Phys. Rev. D **108**, 063005 (2023).
[28] Z. Arzoumanian *et al.* (NANOGrav Collaboration), Phys. Rev. Lett. **127**, 251302 (2021).
[29] X. Xue *et al.*, Phys. Rev. Lett. **127**, 251303 (2021).
[30] Y. Nakai, M. Suzuki, F. Takahashi, and M. Yamada, Phys. Lett. B **816**, 136238 (2021).
[31] P. Di Bari, D. Marfatia, and Y.-L. Zhou, J. High Energy Phys. 10 (2021) 193.
[32] A. S. Sakharov, Y. N. Eroshenko, and S. G. Rubin, Phys. Rev. D **104**, 043005 (2021).
[33] S.-L. Li, L. Shao, P. Wu, and H. Yu, Phys. Rev. D **104**, 043510 (2021).
[34] A. Ashoorioon, K. Rezazadeh, and A. Rostami, Phys. Lett. B **835**, 137542 (2022).
[35] M. Benetti, L. L. Graef, and S. Vagnozzi, Phys. Rev. D **105**, 043520 (2022).
[36] J. Barir, M. Geller, C. Sun, and T. Volansky, arXiv:2203.00693.
[37] M. Hindmarsh and J. Kume, J. Cosmol. Astropart. Phys. 04 (2023) 045.
[38] V. Vaskonen and H. Veermäe, Phys. Rev. Lett. **126**, 051303 (2021).
[39] V. De Luca, G. Franciolini, and A. Riotto, Phys. Rev. Lett. **126**, 041303 (2021).
[40] N. Bhaumik and R. K. Jain, Phys. Rev. D **104**, 023531 (2021).
[41] K. Inomata, M. Kawasaki, K. Mukaida, and T. T. Yanagida, Phys. Rev. Lett. **126**, 131301 (2021).
[42] K. Kohri and T. Terada, Phys. Lett. B **813**, 136040 (2021).
[43] G. Domènech and S. Pi, Sci. China Phys. Mech. Astron. **65**, 230411 (2022).
[44] S. Vagnozzi, Mon. Not. R. Astron. Soc. **502**, L11 (2021).
[45] R. Namba and M. Suzuki, Phys. Rev. D **102**, 123527 (2020).
[46] S. Sugiyama, V. Takhistov, E. Vitagliano, A. Kusenko, M. Sasaki, and M. Takada, Phys. Lett. B **814**, 136097 (2021).
[47] Z. Zhou, J. Jiang, Y.-F. Cai, M. Sasaki, and S. Pi, Phys. Rev. D **102**, 103527 (2020).
[48] J. Lin, S. Gao, Y. Gong, Y. Lu, Z. Wang, and F. Zhang, Phys. Rev. D **107**, 043517 (2023).
[49] K. Rezazadeh, Z. Teimoori, S. Karimi, and K. Karami, Eur. Phys. J. C **82**, 758 (2022).
[50] M. Kawasaki and H. Nakatsuka, J. Cosmol. Astropart. Phys. 05 (2021) 023.
[51] W. Ahmed, M. Junaid, and U. Zubair, Nucl. Phys. **B984**, 115968 (2022).
[52] Z. Yi and Q. Fei, Eur. Phys. J. C **83**, 82 (2023).
[53] Z. Yi, J. Cosmol. Astropart. Phys. 03 (2023) 048.
[54] V. Dandoy, V. Domcke, and F. Rompineve, SciPost Phys. Core **6**, 060 (2023).
[55] J.-X. Zhao, X.-H. Liu, and N. Li, Phys. Rev. D **107**, 043515 (2023).
[56] G. Ferrante, G. Franciolini, A. Iovino Junior, and A. Urbano, arXiv:2305.13382.
[57] Y. Cai, M. Zhu, and Y.-S. Piao, arXiv:2305.10933.
[58] A. Afzal *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **951**, L11 (2023).
[59] L. Mayer, Classical Quantum Gravity **30**, 244008 (2013).
[60] G. Foreman, M. Volonteri, and M. Dotti, Astrophys. J. **693**, 1554 (2009).
[61] M. C. Begelman, R. D. Blandford, and M. J. Rees, Nature (London) **287**, 307 (1980).
[62] E. S. Phinney, arXiv:astro-ph/0108028.
[63] J. Ellis, M. Fairbairn, G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe, Astron. Astrophys. **676**, A38 (2023).
[64] W. H. Press and P. Schechter, Astrophys. J. **187**, 425 (1974).

[65] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophys. J. **379**, 440 (1991).
[66] C. G. Lacey and S. Cole, Mon. Not. R. Astron. Soc. **262**, 627 (1993).
[67] J. Kormendy and L. C. Ho, Annu. Rev. Astron. Astrophys. **51**, 511 (2013).
[68] A. E. Reines and M. Volonteri, Astrophys. J. **813**, 82 (2015).
[69] G. Girelli, L. Pozzetti, M. Bolzonella, C. Giocoli, F. Marulli, and M. Baldi, Astron. Astrophys. **634**, A135 (2020).
[70] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevD.109.L021302 for further discussion on the implications of different scaling relations as well as a more developed mathematical formalism of the computation of the background, which includes Refs. [71–77].
[71] Y. Harikane *et al.*, arXiv:2303.11946.
[72] D. D. Kocevski *et al.*, Astrophys. J. Lett. **954**, L4 (2023).
[73] R. Maiolino *et al.*, arXiv:2308.01230.
[74] F. Pacucci, B. Nguyen, S. Carniani, R. Maiolino, and X. Fan, Astrophys. J. Lett. **957**, L3 (2023).
[75] R. Aversa, A. Lapi, G. de Zotti, F. Shankar, and L. Danese, Astrophys. J. **810**, 74 (2015).
[76] I. Delvecchio *et al.*, Astrophys. J. **892**, 17 (2020).
[77] T. Di Matteo, D. Angles-Alcazar, and F. Shankar, arXiv:2304.11541.
[78] J. Antoniadis *et al.* (EPTA Collaboration), arXiv:2306.16227.
[79] M. Enoki and M. Nagashima, Prog. Theor. Phys. **117**, 241 (2007).
[80] L. Z. Kelley, L. Blecha, L. Hernquist, A. Sesana, and S. R. Taylor, Mon. Not. R. Astron. Soc. **471**, 4508 (2017).
[81] D. Merritt, Classical Quantum Gravity **30**, 244005 (2013).
[82] P. J. Armitage and P. Natarajan, Astrophys. J. Lett. **567**, L9 (2002).
[83] A. I. Macfadyen and M. Milosavljevic, Astrophys. J. **672**, 83 (2008).
[84] Y. Tang, A. MacFadyen, and Z. Haiman, Mon. Not. R. Astron. Soc. **469**, 4258 (2017).
[85] D. J. Muñoz, R. Miranda, and D. Lai, Astrophys. J. **871**, 84 (2019).
[86] M. S. L. Moody, J.-M. Shi, and J. M. Stone, Astrophys. J. **875**, 66 (2019).
[87] C. Tiede, J. Zrake, A. MacFadyen, and Z. Haiman, Astrophys. J. **900**, 43 (2020).
[88] D. J. D'Orazio and P. C. Duffell, Astrophys. J. Lett. **914**, L21 (2021).
[89] P. C. Duffell, D. D'Orazio, A. Derdzinski, Z. Haiman, A. MacFadyen, A. L. Rosen, and J. Zrake, Astrophys. J. **901**, 25 (2020).
[90] Z. Haiman, Z. Haiman, B. Kocsis, B. Kocsis, K. Menou, and K. Menou, Astrophys. J. **700**, 1952 (2009); **937**, 129 (E) (2022).
[91] A. Sesana, Classical Quantum Gravity **30**, 224014 (2013).
[92] L. Z. Kelley, L. Blecha, and L. Hernquist, Mon. Not. R. Astron. Soc. **464**, 3131 (2017).
[93] M. M. Kozhikkal, S. Chen, G. Theureau, M. Habouzit, and A. Sesana, arXiv:2305.18293.
[94] G. Agazie *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **951**, L50 (2023).
[95] E. A. Huerta, S. T. McWilliams, J. R. Gair, and S. R. Taylor, Phys. Rev. D **92**, 063010 (2015).
[96] J. Antoniadis *et al.* (EPTA Collaboration), arXiv:2306.16226.
[97] P. Amaro-Seoane *et al.* (LISA Collaboration), arXiv:1702.00786.
[98] L. Badurina *et al.* (AION Collaboration), J. Cosmol. Astropart. Phys. 05 (2020) 011.
[99] Y. A. El-Neaj *et al.* (AEDGE Collaboration), Eur. Phys. J. Quantum Technol. **7**, 6 (2020).
[100] V. Corbin and N. J. Cornish, Classical Quantum Gravity **23**, 2435 (2006).
[101] T. Nakamura *et al.*, Prog. Theor. Exp. Phys. **2016**, 093E01 (2016).

Appendix 8

VIII
J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Probing Supermassive Black Hole Seed Scenarios with Gravitational-wave Measurements. *Astrophys. J.*, 964(1):11, 2024

The Astrophysical Journal, 964:11 (7pp), 2024 March 20 https://doi.org/10.3847/1538-4357/ad27d5
© 2024. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

Probing Supermassive Black Hole Seed Scenarios with Gravitational-wave Measurements

John Ellis[1,2,3], Malcolm Fairbairn[1], Juan Urrutia[3,4], and Ville Vaskonen[3,5,6]
[1] King's College London, Strand, London WC2R 2LS, UK; malcolm.fairbairn@kcl.ac.uk, john.ellis@cern.ch
[2] Theoretical Physics Department, CERN, Geneva, Switzerland
[3] Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia; juan.urrutia@kbfi.ee, ville.vaskonen@pd.infn.it
[4] Department of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia
[5] Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy
[6] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy
Received 2024 January 17; revised 2024 February 8; accepted 2024 February 9; published 2024 March 11

Abstract

The process whereby the supermassive black holes (SMBHs) populating the centers of galaxies have been assembled remains to be established, with the relative importance of seeds provided by collapsed Population III stars, black holes formed in nuclear star clusters via repeated mergers, or direct collapses of protogalactic disks yet to be determined. In this paper we study the prospects for casting light on this issue by future measurements of gravitational waves emitted during the inspirals and mergers of pairs of intermediate-mass black holes (IMBHs), discussing in particular the roles of prospective measurements by LISA and the proposed atom interferometers AION and AEDGE. We find that the expected number of detectable IMBH binaries is $\mathcal{O}(100)$ for LISA and AEDGE and $\mathcal{O}(10)$ for AION in low-mass seeds scenarios and goes down to $\mathcal{O}(10)$ for LISA and below one for AEDGE and AION in high-mass seed scenarios. This allows all of these observatories to probe the parameters of the seed model, in particular, if at least a fraction of the SMBHs arises from a low-mass seed population. We also show that the measurement accuracy of the binary parameters is, in general, best for AEDGE, which sees very precisely the merger of the binary.

Unified Astronomy Thesaurus concepts: Intermediate-mass black holes (816); Gravitational waves (678); Cosmology (343); Supermassive black holes (1663)

1. Introduction

Most galaxies contain supermassive black holes (SMBHs) heavier than $10^6\,M_\odot$ (Kormendy & Ho 2013), and the existence of black holes (BHs) with masses between a few and $\sim 80\,M_\odot$ has been established by observations of X-ray binaries (Remillard & McClintock 2006) and by the measurements of gravitational waves (GWs) with frequencies ~ 100 Hz emitted during their mergers (Abbott et al. 2019, 2021a, 2021b). Various other observations point to the existence of intermediate-mass black holes (IMBHs) with masses in the range 10^4–$10^6\,M_\odot$, but their mass function and redshift distribution is known only very poorly (Greene et al. 2020). This lack of information about IMBHs impedes our understanding of how SMBHs have been assembled (Reines 2022).

The main possibilities for seeding SMBH assembly include collapsed Population III stars (Madau & Rees 2001; Bromm & Loeb 2003), BHs formed in nuclear star clusters via repeated mergers (Atakan Gurkan et al. 2004; Portegies Zwart et al. 2004; Natarajan 2021), or direct collapses of protogalactic disks in which fragmentation is suppressed (Sesana et al. 2004; Begelman et al. 2006; Mayer et al. 2007; Volonteri et al. 2008; Tanaka & Li 2014; Inayoshi et al. 2015; Izquierdo-Villalba et al. 2023). All of these mechanisms are capable of reproducing the properties of the observed SMBH population for suitable values of assembly parameters, such as accretion rates, but can differ significantly in their predictions for the spectrum of IMBH mergers at different redshifts; see Volonteri et al. (2021) for an overview. Signatures of these mechanisms may include either light seeds, with masses $\sim 10^2$–$10^3\,M_\odot$, or heavy seeds, with masses $\sim 10^4$–$10^5\,M_\odot$, at $z \lesssim 10$. These are currently unconstrained by data, but can in principle be probed by future GW and other measurements (Sesana et al. 2011; Hartwig et al. 2016; Krolik et al. 2019; Mangiagli et al. 2020; Volonteri et al. 2020; Haidar et al. 2022; Dong-Páez et al. 2023).

The purpose of this paper is to investigate what progress can be made in distinguishing between the SMBH assembly scenarios with planned future GW experiments at frequencies larger than 0.1 mHz in light of recent data on nHz GWs from pulsar timing array (PTA) experiments, highlighting the potential capabilities of atom interferometers. Whereas PTAs are sensitive to the early inspiralling phase of the SMBHs heavier than $10^9 M_\odot$, the future GW experiments will probe directly the infalls and mergers of binaries lighter than $10^7 M_\odot$.

NANOGrav (Agazie et al. 2023a) and other PTA experiments (Antoniadis et al. 2023a; Xu et al. 2023; Zic et al. 2023) have recently reported the observation of a stochastic background of GWs at frequencies in the nHz range, for which the most conservative astrophysical interpretation is that binary systems of SMBHs are emitting them with masses $\sim 10^9\,M_\odot$ (Agazie et al. 2023b, 2023c; Antoniadis et al. 2023b; Ellis et al. 2023). Naive extrapolation of binary merger models to lower BH masses suggests that GWs from IMBH binaries may be observable at higher frequencies between 0.1 mHz and 1 Hz, for example by the LISA spaceborne laser interferometer experiment or atom interferometer experiments (Ellis et al. 2024). In principle, there are two regimes where the formation

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

channels of the SMBHs may be distinguished. Either at $z \gtrsim 7$ when the seeds are assembling and scaling relations are not that strong (Treister et al. 2013; Ricarte & Natarajan 2018a), or by observing the low-mass occupation fraction of dwarf galaxies and dark matter halos at more recent times (Ricarte & Natarajan 2018b; Volonteri et al. 2021). The extrapolation to higher frequencies of a model that can fit the NANOGrav background (Ellis et al. 2023) predicts that the majority of detectable binaries will be at $z < 7$, so the focus of this study is to constrain the latter. Our work complements analogous studies that have been performed with electromagnetic observations of active galactic nuclei (AGNs; Kelly & Shen 2013; Miller et al. 2015; Gallo & Sesana 2019; Chadayammuri et al. 2023).

To extend these observations to the GW spectrum and establish a multimessenger signal, we estimate how well the spaceborne laser interferometer experiment LISA (Amaro-Seoane et al. 2017) and the atom interferometers AION-1 km (Badurina et al. 2020, 2021) and AEDGE (El-Neaj et al. 2020; Badurina et al. 2021) could recover the occupation fraction and the weight of the different channels contributing to the formation of the SMBHs.[7] Our approach is to generate populations of the IMBH binaries, estimate how well the binary parameters could be recovered, and then construct posteriors for the parameters of the merger rate. To extend the deduced merger rate from the NANOGrav observations, we assume that the scaling relation extends to lower masses in a light-seed scenario. Although there is evidence for this assumption (Baldassare et al. 2020), there is no consensus among simulations and semianalytical models (Fontanot et al. 2015). However, under this assumption, we predict that laser and atom interferometers will be to place new and competitive constraints after their first years of observation, shedding light on the origin of SMBHs.

2. Model

We use the extended Press–Schechter formalism (Press & Schechter 1974; Bond et al. 1991; Lacey & Cole 1993) to estimate the rate R_h of coalescences of galactic halos of masses $M_{1,2}$. Assuming conservatively that each of the halos includes at most one BH and that the BHs merge with the halos with probability $p_{\rm BH}$ (Ellis et al. 2023), the merger rate of BHs can be written as

$$\frac{dR_{\rm BH}(p_{\rm BH}, \boldsymbol{\theta})}{dm_1 dm_2} \approx p_{\rm BH} \int dM_1 dM_2 \times \frac{dR_h}{dM_1 dM_2} \prod_{j=1,2} p_{\rm occ}(m_j|M_j, z, \boldsymbol{\theta}), \quad (1)$$

where $p_{\rm occ}(m_{\rm BH}|M_v, z, \boldsymbol{\theta})$ is the occupation fraction of BHs of mass $m_{\rm BH}$ in halos of mass M_v at redshift z, and the mechanisms for SMBH formation are characterized by a set of parameters $\boldsymbol{\theta}$. We fix the value of the merging efficiency $p_{\rm BH}$ to 0.84, which corresponds to the best fit of the SGWB measured by the NANOGrav collaboration (Ellis et al. 2024). This fit was obtained by including environmental effects on the binary evolution. Such effects are preferred by the fit but will not affect the binaries at the frequencies considered in this study.

[7] See also the related discussion of the capability of the proposed TianQin experiment in Wang et al. (2019).

We model the occupation fraction $p_{\rm occ}(m_{\rm BH}|M_v, z, \boldsymbol{\theta})$ by a sum over different SMBH seed channels:

$$p_{\rm occ}(m_{\rm BH}|M_v, z, \boldsymbol{\theta}) = \sum_j \frac{p_j(m_{\rm BH}, \boldsymbol{\theta}_j)}{\sqrt{2\pi}\, m_{\rm BH}\sigma} \exp\left[-\frac{\ln(m_{\rm BH}/\bar{m})^2}{2\sigma^2}\right], \quad (2)$$

where the lognormal distribution parameterizes the halo mass–BH mass relation and p_j parameterizes the low-mass cut of the massive BH population. We compute the latter by using the observed halo mass–stellar mass relation $M_*(M_v)$ (Girelli et al. 2020), the mean stellar mass–BH mass relation obtained from observations of inactive galaxies[8] (Kormendy & Ho 2013; Reines & Volonteri 2015), $\log_{10}(\bar{m}/M_\odot) = 8.95 + 1.4\log_{10}(M_*/10^{11}M_\odot)$, and the observed scatter of $\sigma = 1.1$ in the latter.[9]

At low redshifts, $z \lesssim 7$, different SMBH seed scenarios are reflected in differences in the low-mass end of the occupation fraction. This is accounted for by the functions $p_j(m_{\rm BH}, \boldsymbol{\theta}_j)$ for which we consider the following parametric form:

$$p_j(m_{\rm BH}, \boldsymbol{\theta}_j) = \frac{f_j}{2}\left(1 + \mathrm{erf}\left[\frac{\log_{10}(m_{\rm BH}/m_{{\rm cut},j})}{w_j}\right]\right) \quad (3)$$

with the parameters $\boldsymbol{\theta}_j = (f_j, m_{{\rm cut},j}, w_j)$, where f_j is the fraction of SMBHs produced by a given mechanism j, $m_{{\rm cut},j}$ characterizes a cut on the minimal SMBH mass, and w_j is the spread with which the minimal-mass cut is applied.[10] We fix $\sum_j f_j = 1$. The function p_j is shown in the left panel of Figure 1 for different values of $m_{\rm cut}$ and w. In the following analysis, we consider two cases: (1) a one-component model where the SMBH population arises from one seed population and (2) a two-component model that includes a population arising from light seeds, e.g., BH remnants of Population III stars, and a population arising from heavy seeds, e.g., BH nuclei of protogalaxies. We show an example of the occupation fraction $p_{\rm occ}$ in the latter case in the right panel of Figure 1.[11]

We emphasize that the above estimate of the BH merger rate is subject to various uncertainties. In particular, we neglect the additional delays related to the binary evolution following a halo merger (Hao et al. 2024) as well as possible difficulties for the SMBH to shrink to the galactic cores in dwarf galaxies (Dayal et al. 2020; Dunn et al. 2020), and we extrapolate the halo mass–BH mass relation to much lower masses than where it is currently measured. However, since our goal is mainly to illustrate and compare the capabilities of LISA, AION, and AEDGE, a detailed discussion of overall uncertainties is beyond the scope of this work.

[8] Another measurement of the stellar mass–BH mass relation comes from AGN observations. As discussed in Ellis et al. (2024), this relation might not reflect the full low-z massive BH population but only its low-mass tail.

[9] Notice that we use the lognormal distribution in Equation (2), so this scatter corresponds to 0.47 dex as in Reines & Volonteri (2015).

[10] A qualitatively similar parameterization of the occupation fraction was adopted in Miller et al. (2015) and in Chadayammuri et al. (2023) for studies of SMBH seed models with electromagnetic observations of AGNs.

[11] There may be other BH binary populations that can complicate the searches of the population related to the assembly of SMBHs. In particular, IMBH binaries may form in dense stellar clusters (Rasskazov et al. 2020; Fragione et al. 2022). The merger rate of this binary population, however, depends differently on redshift than the merger rate we consider and may become relevant only at $z \lesssim 1$.

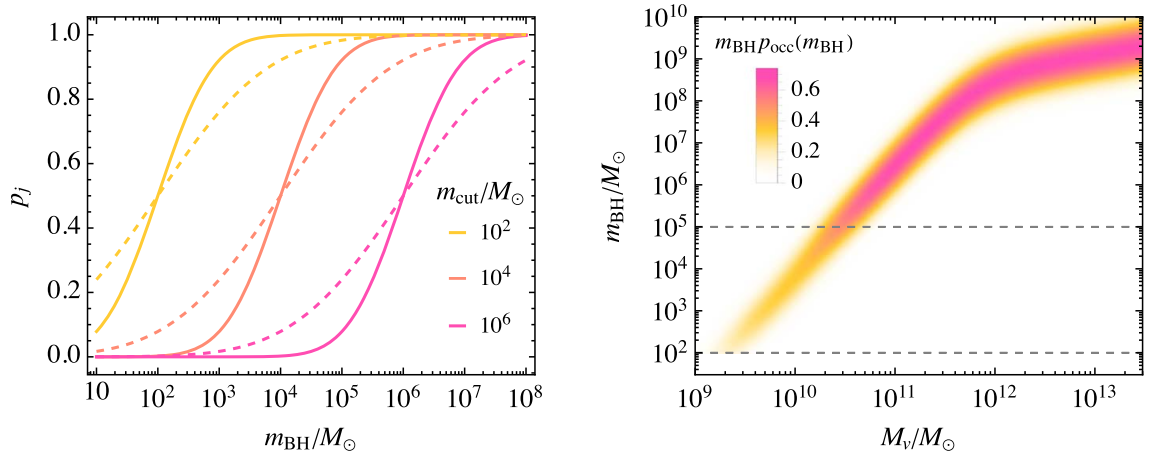


Figure 1. Left panel: the low-mass cut on the massive BH population for different values of $m_{\rm cut}$, for $w = 1$ (solid) and $w = 2$ (dashed). Right panel: the halo mass–BH mass relation. The color coding shows the BH occupation fraction $p_{\rm occ}$ as a function of the halo mass M_v and the BH mass $m_{\rm BH}$ for a scenario with $f_1 = f_2 = 0.5$, $w_1 = w_2 = 1$, $m_{\rm cut,1} = 100\,M_\odot$, and $m_{\rm cut,2} = 10^5\,M_\odot$.

3. GW Data Analysis

For each specific SMBH formation model, characterized by the parameters $\boldsymbol{\theta}$, we generate n Monte Carlo (MC) realizations of the expected binary populations using Equation (1). The number of binaries N_j in any given MC realization follows a Poisson distribution whose mean is the expected number $\bar{N}(\boldsymbol{\theta})$ of detectable binaries during one year of observation. The latter is given by

$$\bar{N}(\boldsymbol{\theta}) = \int_0^\infty d\tau \int d\lambda(\boldsymbol{\theta}, \boldsymbol{x})\, p_{\rm det}\left[(S/N_c)/(S/N(\boldsymbol{x}, \tau))\right], \quad (4)$$

where S/N($\boldsymbol{x}$) is the signal-to-noise ratio of the GW signal from a binary with parameters $\boldsymbol{x}$ in an optimal source-detector system for a given GW detector, $p_{\rm det}$ is the detection probability that accounts for the source location and orientation (Finn & Chernoff 1993; Gerosa et al. 2019), and $S/N_c = 8$.

To characterize how well the parameters of the formation mechanism can be measured, we compute the likelihood for each of the MC realizations (Mandel et al. 2019; Hütsi et al. 2021):

$$\ell_j(\boldsymbol{\theta}) \propto e^{-\bar{N}(\boldsymbol{\theta})} \prod_{i=1}^{N_j} \int d\lambda(\boldsymbol{\theta}, \boldsymbol{x}) P_i(\boldsymbol{x}_i|\boldsymbol{x}), \quad (5)$$

where $\boldsymbol{x}_i = (m_{1,i}, m_{2,i}, z_i)$ are the parameters of the ith binary in the generated population, $\boldsymbol{x} = (m_1, m_2, z)$ are the integration variables and

$$d\lambda(\boldsymbol{\theta}, \boldsymbol{x}) \equiv \frac{1}{1+z}\frac{dV_c}{dz}\frac{dR_{\rm BH}(p_{\rm BH}, \boldsymbol{\theta})}{dm_1 dm_2} dm_1 dm_2 dz. \quad (6)$$

The distributions $p_i(\boldsymbol{x}_i|\boldsymbol{x})$ account for the accuracy with which the parameters of the binary i can be measured. We assume that $p_i(\boldsymbol{x}_i|\boldsymbol{x})$ follows a multivariate Gaussian distribution centered around the true values $\boldsymbol{x}_i$, and estimate the covariance matrix from the Fisher information, $\Sigma_{kl}(\boldsymbol{x}_i) = \Gamma^{-1}_{kl}(\boldsymbol{x}_i)$. Finally, we compute the expected likelihood as

$$\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} \ell_j(\boldsymbol{\theta}) \quad (7)$$

and use that to estimate the expected accuracy at which the model parameters $\boldsymbol{\theta}$ can be measured.

We express the optimal S/N and the Fisher matrix $\Gamma_{kl}(\boldsymbol{x})$ of a signal as (Poisson & Will 1995)

$$\begin{aligned} S/N(\boldsymbol{x}) &= \sqrt{\langle \tilde{h}(\boldsymbol{x})|\tilde{h}(\boldsymbol{x})\rangle}\,, \\ \Gamma_{kl}(\boldsymbol{x}) &= \langle \partial_k \tilde{h}(\boldsymbol{x})|\partial_l \tilde{h}(\boldsymbol{x})\rangle, \end{aligned} \quad (8)$$

by defining the inner product of two complex functions $a(f)$ and $b(f)$ as

$$\langle a(f)|b(f)\rangle = 2\int_{f(\tau)}^{f(\tau-\mathcal{T})} df\, \frac{a^*(f)b(f) + b^*(f)a(f)}{S_n(f)}, \quad (9)$$

where $\mathcal{T}$ denotes the observation time and $S_n(f)$ is the noise power spectral density of the considered GW experiment. The partial derivatives in the Fisher matrix are taken with respect to the parameters of the template. We take into account the binary component masses m_1 and m_2, its redshift z, the phase of the GW signals ϕ_c, and its coalescence time τ, but not the binary sky location or inclination, over which we simply average. To estimate the Fourier transform of the GW strain, $\tilde{h}(\boldsymbol{x})$, we use the inspiral-merger-ringdown template (Ajith et al. 2008).

4. Results

4.1. Binary Population

We perform the analysis for LISA, AEDGE, and AION-1 km. As can be seen in the left panel of Figure 2, LISA is expected to observe more binaries than AEDGE. For low-mass seed scenarios, this difference is small but it grows with $m_{\rm cut}$ implying that LISA will perform much better than AEDGE for heavy-seed scenarios. In particular, the expected number of binaries observable with AEDGE is less than one for heavy-

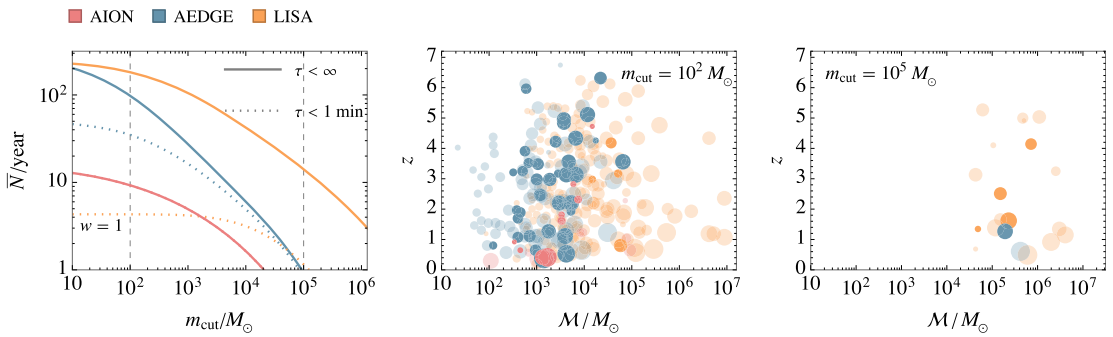


Figure 2. Left panel: the expected numbers of binaries detectable by AION-1 km, AEDGE, and LISA during 1 yr of observation, as functions of $m_{\rm cut}$. The solid curves show all detectable binaries whereas the dotted curves show only those for which the last 1 minute of the merger is seen. Middle and right panels: explicit examples of the detectable binaries for a light-seed and a heavy-seed scenario. The sizes of the dots are $\propto \ln[{\rm S/N}^{-1}]$ with the minimum size corresponding to ${\rm S/N} = 10^4$ and the maximal to ${\rm S/N} = 10$. The darker dots correspond to binaries for which the last 1 minute of the merger is seen.

seed scenarios with $m_{\rm cut} > 10^5\,M_\odot$.[12] For Figure 2 we have fixed the width of the mass cut to $w = 1$, but we have checked that small changes to w do not alter significantly these results. Focusing on binaries that can be observed within 1 minute of the merger, the number of events for LISA is reduced to only approximately five events for the light-seed case. In the heavy-seed case, LISA will not see the last 1 minute of the binaries if $m_{\rm cut} > 10^5\,M_\odot$. On the other hand, AEDGE would detect most of the binaries until within 1 minute of the merger, except for those with $\mathcal{M}_c \lesssim 100\,M_\odot$. AION-1 km has a similar range to AEDGE but, because it has less sensitivity, the number of events is decreased to around 10 if $m_{\rm cut} \sim 10^2 M_\odot$ and further reduced to one event on average for $2 \times 10^4\ M_\odot$.

An explicit realization of the binary population in a light-seed scenario with ($m_{\rm cut} = 10^2\,M_\odot$) is shown in the middle panel of Figure 2, and in a heavy-seed scenario with ($m_{\rm cut} = 10^5\,M_\odot$) in the right panel of Figure 2. We see that AEDGE and, to a lesser extent, AION-1 km will probe IMBH binaries over a range of masses that are too light for LISA, covering the mass range covered by terrestrial laser interferometers from $\mathcal{M}_c < 100 M_\odot$ to $\sim 10^4\,M_\odot$. LISA, as expected, will probe heavier binaries up to $\mathcal{M}_c \sim 10^7\,M_\odot$. The types of events each experiment can detect are quite different. Considering the final stages within 1 minute of a merger, AEDGE binaries can explore the range $\mathcal{M} \in (10^2$–$10^4)\,M_\odot$ out to redshifts $z \sim 7$. LISA, on the other hand, observes only the last minutes of heavier binaries with $\mathcal{M} \in (10^4$–$10^6)\,M_\odot$ and $z \lesssim 4$. Finally, AION-1 km will detect a handful of events for $m_{\rm cut} < 10^4\ M_\odot$, at low redshifts, $z < 4$, and mainly in the $(10^3$–$10^4)\,M_\odot$ mass range, mostly in the last moments before the merger.

4.2. Binary Parameters

In Figure 3 we display the accuracies with which LISA, AEDGE, and AION-1 km can measure the binary parameters for two representative choices of the binary component masses, a lighter binary with $(m_1, m_2) = (500, 300)\,M_\odot$ (upper panels) and a heavier binary with $(m_1, m_2) = (10^4, 5 \times 10^4)\,M_\odot$ (lower panels), as functions of the binary redshift z (left panels) and the time to merge τ (right panels). In general, the measurement accuracy of the binary parameters m_1, m_2, and z are similar, and the measurement accuracy is best with AEDGE. The lighter binary enters the AION/AEDGE sensitivity band about a day before the merger. Consequently, the binary is not observable for AION/AEDGE if its coalescence time at the beginning of the experiment is longer than the observation time ($T = 1$ yr). LISA instead sees only the inspiralling part until less than 1 day from the merger and can see binaries whose coalescence time is $\tau \lesssim 10$ yr. The heavier binary is seen with AION/AEDGE for an even shorter time, but since the signal is strong the measurement accuracy is very good. In this case, also LISA sees the merger but the measurement with AEDGE is still more accurate unless the redshift of the binary is high, $z \gtrsim 4$, or its coalescence time is longer than the observation time.

For many binaries, the measurement accuracy with AEDGE or LISA is excellent. The probability distributions of the binary parameters can then be approximated by delta functions in the computation of the likelihood Equation (5) so that the integrals over the binary parameters become trivial because the merger rate is essentially constant over such small parameter ranges. This reduces significantly the computational time for the likelihood. We adopt the delta-function approximation when all binary parameters are measured with better than 10% accuracy.

4.3. Seed Scenarios

To estimate how accurately $m_{\rm cut}$ could be measured with 1 yr data samples, we have generated $n = 20$ realizations of the detectable binary populations for a given value of $m_{\rm cut}$ considering a single SMBH seed population. The left panel of Figure 4 shows on the vertical axis the 95% confidence level (CL) ranges of $m_{\rm cut}$ estimates for $m_{\rm cut} \in [10^2,\ 10^6]\,M_\odot$ and taking a flat prior for $\log m_{\rm cut}$ over the range $m_{\rm cut} \in [10^2,\ 10^6]\ M_\odot$. The solid and dashed curves correspond to $w = 1$ and $w = 2$ and show that the measurement accuracy of $m_{\rm cut}$ in these two cases is very similar. We see that the seed mass is best recovered with LISA with a small uncertainty of a factor ~ 2. For $m_{\rm cut} \lesssim \times\, 10^3\,M_\odot$ the accuracy with AEDGE is similar to LISA, but the 95% CL range for AEDGE grows for heavier masses. When $m_{\rm cut} \gtrsim 10^5\,M_\odot$, only a lower bound is obtained with AEDGE because the expected number of detectable

[12] We have cut the AEDGE sensitivity at 0.01 Hz. Depending on the orbit of the satellites, the Newtonian gravity backgrounds may affect the sensitivity at frequencies below that (Hogan et al. 2011). If these backgrounds do not affect the sensitivity, the expected number of binaries for AEDGE will increase significantly, especially in the heavy-seed scenarios. For AION-1 km we have considered the low-noise model (Badurina et al. 2023).

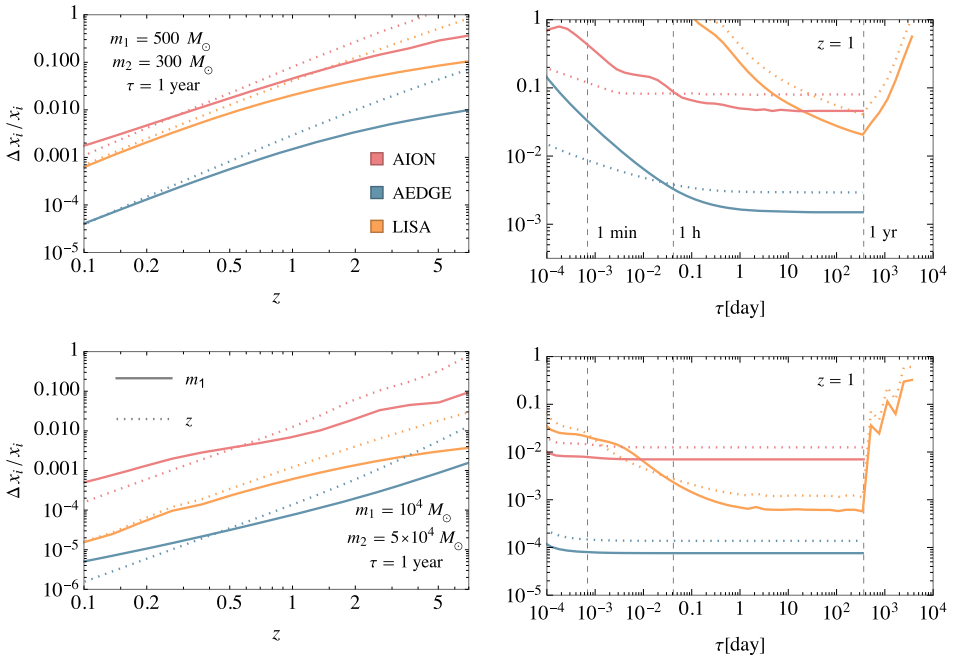


Figure 3. Measurement accuracies for the binary parameters: the upper and lower panels correspond to two binaries whose component masses are fixed. In the left panels, the binary is observed for the last 1 yr and the errors are shown as a function of its redshift. In the right panels, $z = 1$ and the errors are shown as a function of the binary coalescence time at the beginning of a 1 yr observation. The errors for the masses of both BHs are almost the same.

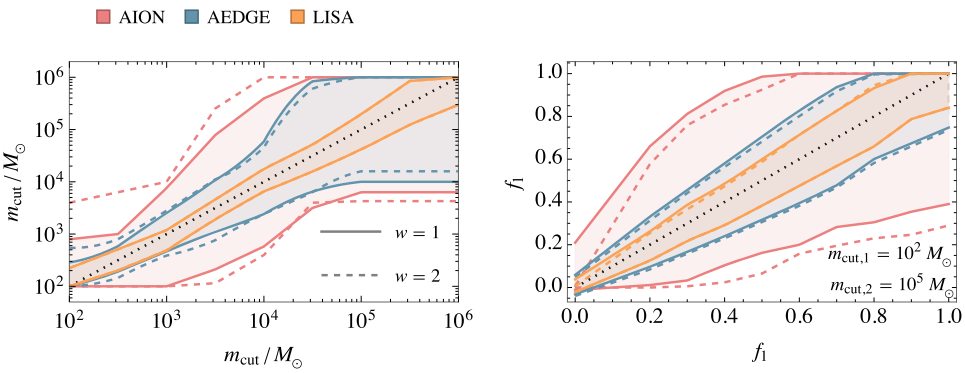


Figure 4. Left panel: the 95% CL accuracy with which LISA, AEDGE, and AION-1 km could measure $m_{\rm cut}$ over the range $[10^2, 10^6]\ M_\odot$. Right panel: the 95% CL accuracy with which LISA, AEDGE, and AION-1 km could measure the fraction of light seeds, f_1, assuming an input mixture of seeds with masses 10^2 and $10^5\ M_\odot$ and $f_2 = 1 - f_1$. The solid and dashed curves in both panels correspond, respectively, to $w = 1$ and $w = 2$.

events is less than one in the heavy-seed scenarios. On the other hand, already AION-1 km could measure $m_{\rm cut} \lesssim 3 \times 10^4\ M_\odot$ within 1 order of magnitude at 95% CL and could place a 95% CL lower bound of $m_{\rm cut} \gtrsim 4 \times 10^3\ M_\odot$ if $m_{\rm cut} \gtrsim 3 \times 10^4\ M_\odot$.

In the right panel of Figure 4, we have considered the possibility that there are two different SMBH seed populations, a light one with $m_{\rm cut,1} = 100\ M_\odot$, and a heavy one with $m_{\rm cut,2} = 10^5\ M_\odot$. We have again generated $n = 20$ realizations of the detectable binary populations for a given value of f_1 with $f_2 = 1 - f_1$. In the right panel of Figure 4, we show how accurately f_1 can be recovered at the 95% CL. We see that LISA could measure f_1 with an accuracy $\sim$10%, whereas AEDGE could measure $f_{\rm light}$ with an accuracy $\sim$10%–20%. AION-1 km has a much bigger uncertainty but it would be precise enough that it can differentiate at the 95% CL a pure light-seed from a pure heavy-seed scenario.

For the right panel of Figure 4, we have assumed delta-function priors on $m_{\rm cut,1}$ and $m_{\rm cut,2}$ that might be theoretically motivated. Instead, in Figure 5, we consider a benchmark case with $m_1 = 10^2\ M_\odot$, $m_2 = 10^5\ M_\odot$, and $f_1 = f_2 = 0.5$, and show the one- and two-dimensional posteriors for their measurements with AION-1 km (top panel), AEDGE (middle panel), and LISA (bottom panel) assuming the log-uniform priors

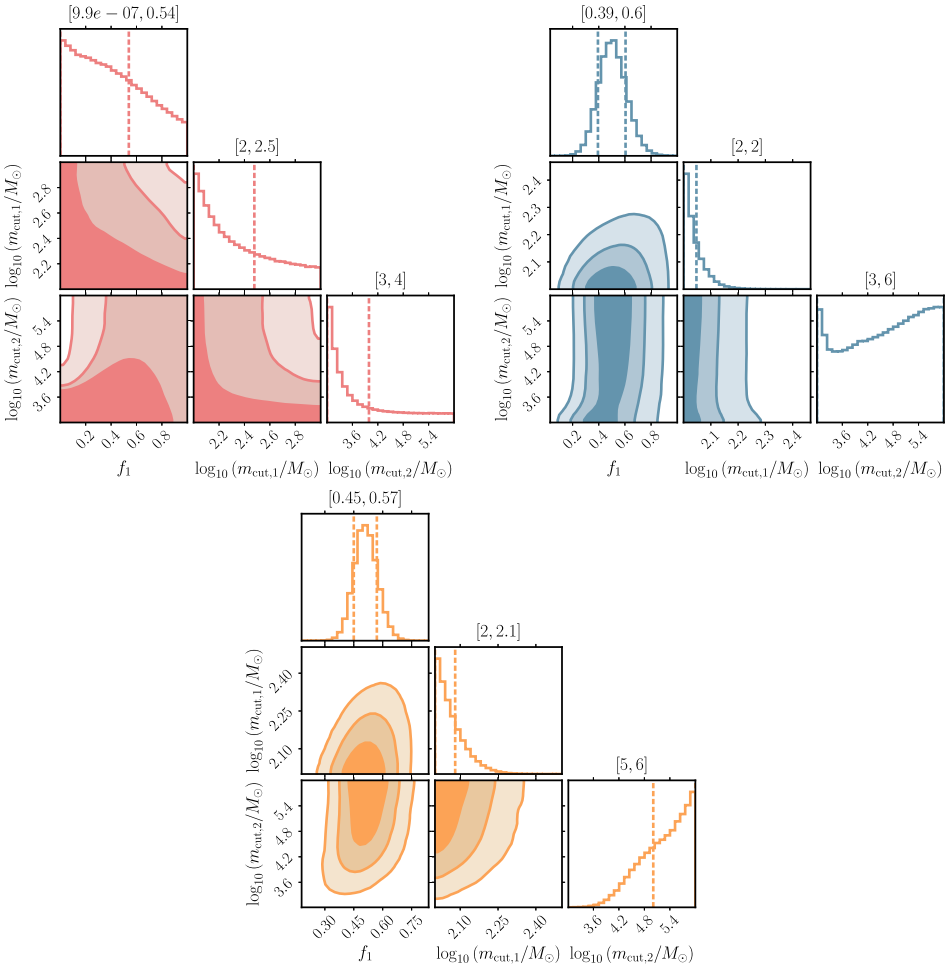


Figure 5. The one- and two-dimensional posteriors of the merger rate parameters (f_1, $m_{\rm cut,1}$, $m_{\rm cut,2}$) for a two-component seed model with $m_{\rm cut,1} = 10^2\,M_\odot$ $m_{\rm cut,2} = 10^5\,M_\odot$ and $f_1 = 0.5$ for AION-1 km (top panel), AEDGE (middle panel), and LISA (bottom panel). The contours enclose the 68%, 95%, and 99% CL regions and the dashed vertical lines show the 68% CL ranges in the marginalized posteriors.

$2 < \log_{10}(m_{\rm cut,1}/M_\odot) < 3$ and $3 < \log_{10}(m_{\rm cut,2}/M_\odot) < 6$ for the cutoff masses. For simplicity, we keep $w_1 = w_2 = 1$ fixed. We see that AION-1 km provides a good estimate of $m_{\rm cut,1}$, but provides only limited information on $m_{\rm cut,2}$ and f_1. If we would take the prior $m_{\rm cut,2} \gtrsim 2 \times 10^4\,M_\odot$, then we would reproduce the lower bound on f_1 with AION-1 km seen in the right panel of Figure 4. On the other hand, AEDGE provides good estimates of both $m_{\rm cut,1}$ and f_1 but also does not constrain $m_{\rm cut,2}$ significantly while LISA not only provides good estimates of both $m_{\rm cut,1}$ and f_1 but also constrains $m_{\rm cut,2}$ significantly. The marginalized posteriors of f_1 for AEDGE and LISA roughly match to the results shown in the right panel of Figure 4.

5. Conclusions

We have described in this paper the capabilities of the planned spaceborne laser interferometer LISA and the proposed atom interferometers AEDGE and AION-1 km to observe mergers of intermediate-mass BHs, measure their parameters, and discriminate between different seed scenarios for the assembly of SMBHs. We have considered the extended Press–Schechter to model the coalescences of galactic halos and estimate a rate for mergers of SMBHs that is compatible with the PTA signals for GWs in the nHz range. We have extrapolated this model to different SMBH seed scenarios by parameterizing the low-mass cutoff of the massive BH population. Using this parameterization, we have estimated the possible rates for IMBH mergers and assessed their detectability and measurability.

We have found that, although LISA has a high rate for observing the early infall stages of IMBH binaries for all the masses studied, this detector loses many binaries as the merger time approaches. Both AEDGE and AION-1 km have higher rates than LISA for detections within 1 minute of the merger. We have shown that AEDGE has the best perspectives for

detecting mergers of IMBHs weighing $\lesssim 10^4\,M_\odot$ whereas LISA has better perspectives for IMBHs weighing $\gtrsim 10^5\,M_\odot$. The better detection rates translate into smaller uncertainties in the measurements by AEDGE of binary parameters for IMBHs weighing $\lesssim 10^5\,M_\odot$.

We have estimated the accuracy with which a lower cutoff on the BH seed mass, $m_{\rm cut}$ could be extracted from the prospective GW data. We find that both LISA and AEDGE could determine $m_{\rm cut}$ with precision $\lesssim$20% if $m_{\rm cut} \lesssim 10^4\,M_\odot$, whereas LISA could determine $m_{\rm cut}$ with better precision than AEDGE if $m_{\rm cut} \gtrsim 10^4\,M_\odot$. We also find that both LISA and AEDGE have interesting capabilities for distinguishing between scenarios with different mixtures of seeds with 10^2 and $10^5\,M_\odot$. AION-1 km could also provide some information, particularly in scenarios with a population of low-mass seeds.

Our results indicate that the spaceborne laser interferometer LISA and atom interferometers AEDGE and AION-1 km have interesting and complementary capabilities for measuring IMBH mergers and distinguishing between different seed scenarios for the assembly of SMBHs. We should emphasize that our study has been exploratory and should be complemented by an improved modeling of the SMBH seed scenarios and more detailed studies of the instrumental capabilities of GW interferometers. It would also be interesting to extend the analysis to assess the prospects for multimessenger observations and study the prospects for measuring higher-order multipoles of the GW signals that would allow, for example, for new probes of strong gravity.

Acknowledgments

The work of J.E. was supported by the United Kingdom STFC grants ST/X000753/1 and ST/T00679X/1, and that of M.F. was also supported by the United Kingdom STFC grant ST/X000753/1. The work of J.U. and V.V. was supported by European Regional Development Fund through the CoE program grant TK202 and by the Estonian Research Council grant PRG803. The work of V.V. was also partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736.

ORCID iDs

John Ellis https://orcid.org/0000-0002-7399-0813
Malcolm Fairbairn https://orcid.org/0000-0002-0566-4127
Juan Urrutia https://orcid.org/0000-0002-6035-6610
Ville Vaskonen https://orcid.org/0000-0003-0003-2259

References

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019, PhRvX, 9, 031040
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021a, PhRvX, 11, 021053
Abbott, R., Abbott, T. D., Acernese, F., et al. 2021b, PhRvX, 13, 011048
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023a, ApJL, 951, L8
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023b, ApJL, 952, L37
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023c, ApJL, 951, L50
Ajith, P., Babak, S., Chen, Y., et al. 2008, PhRvD, 77, 104017
Amaro-Seoane, P., Audley, H., Babak, S., et al. 2017, arXiv:1702.00786
Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2023b, arXiv:2306.16227
Antoniadis, J., Babak, S., Bak Nielsen, A.-S., et al. 2023a, A&A, 678, A48
Atakan Gurkan, M., Freitag, M., & Rasio, F. A. 2004, ApJ, 604, 632
Badurina, L., Bentine, E., Blas, D., et al. 2020, JCAP, 05, 011
Badurina, L., Buchmueller, O., Ellis, J., et al. 2021, RSPTA, 380, 20210060
Badurina, L., Gibson, V., McCabe, C., & Mitchell, J. 2023, PhRvD, 107, 055002
Baldassare, V. F., Dickey, C., Geha, M., & Reines, A. E. 2020, ApJ, 898, L3
Begelman, M. C., Volonteri, M., & Rees, M. J. 2006, MNRAS, 370, 289
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Bromm, V., & Loeb, A. 2003, ApJ, 596, 34
Chadayammuri, U., Bogdan, A., Ricarte, A., & Natarajan, P. 2023, ApJ, 946, 51
Dayal, P., Volonteri, M., Choudhury, T. R., et al. 2020, MNRAS, 495, 3065
Dong-Páez, C. A., Volonteri, M., Beckmann, R. S., et al. 2023, A&A, 676, A2
Dunn, G., Holley-Bockelmann, K., & Bellovary, J. 2020, ApJ, 896, 72
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2023, A&A, 676, A38
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2024, PhRvD, 109, L021302
El-Neaj, Y. A., Alpigiani, C., Amairi-Pyka, S., et al. 2020, EPJQT, 7, 6
Finn, L. S., & Chernoff, D. F. 1993, PhRvD, 47, 2198
Fontanot, F., Monaco, P., & Shankar, F. 2015, MNRAS, 453, 4113
Fragione, G., Loeb, A., Kocsis, B., & Rasio, F. A. 2022, ApJ, 933, 170
Gallo, E., & Sesana, A. 2019, ApJL, 883, L18
Gerosa, D., Ma, S., Wong, K. W. K., et al. 2019, PhRvD, 99, 103004
Girelli, G., Pozzetti, L., Bolzonella, M., et al. 2020, A&A, 634, A135
Greene, J. E., Strader, J., & Ho, L. C. 2020, ARA&A, 58, 257
Haidar, H., Habouzit, M., Volonteri, M., et al. 2022, MNRAS, 514, 4912
Hao, W., Kouwenhoven, M. B. N., Spurzem, R., et al. 2024, MNRAS, 527, 10705
Hartwig, T., Volonteri, M., Bromm, V., et al. 2016, MNRAS, 460, L74
Hogan, J. M., Johnson, D. M. S., Dickerson, S., et al. 2011, GReGr, 43, 1953
Hütsi, G., Raidal, M., Vaskonen, V., & Veermäe, H. 2021, JCAP, 03, 068
Inayoshi, K., Visbal, E., & Kashiyama, K. 2015, MNRAS, 453, 1692
Izquierdo-Villalba, D., Colpi, M., Volonteri, M., et al. 2023, A&A, 677, A123
Kelly, B. C., & Shen, Y. 2013, ApJ, 764, 45
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Krolik, J. H., Volonteri, M., Dubois, Y., & Devriendt, J. 2019, ApJ, 879, 110
Lacey, C. G., & Cole, S. 1993, MNRAS, 262, 627
Madau, P., & Rees, M. J. 2001, ApJL, 551, L27
Mandel, I., Farr, W. M., & Gair, J. R. 2019, MNRAS, 486, 1086
Mangiagli, A., Klein, A., Bonetti, M., et al. 2020, PhRvD, 102, 084056
Mayer, L., Kazantzidis, S., Madau, P., et al. 2007, Sci, 316, 1874
Miller, B. P., Gallo, E., Greene, J. E., et al. 2015, ApJ, 799, 98
Natarajan, P. 2021, MNRAS, 501, 1413
Poisson, E., & Will, C. M. 1995, PhRvD, 52, 848
Portegies Zwart, S. F., Baumgardt, H., Hut, P., Makino, J., & McMillan, S. L. W. 2004, Natur, 428, 724
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Rasskazov, A., Fragione, G., & Kocsis, B. 2020, ApJ, 899, 149
Reines, A. E. 2022, NatAs, 6, 26
Reines, A. E., & Volonteri, M. 2015, ApJ, 813, 82
Remillard, R. A., & McClintock, J. E. 2006, ARA&A, 44, 49
Ricarte, A., & Natarajan, P. 2018a, MNRAS, 474, 1995
Ricarte, A., & Natarajan, P. 2018b, MNRAS, 481, 3278
Sesana, A., Gair, J., Berti, E., & Volonteri, M. 2011, PhRvD, 83, 044036
Sesana, A., Haardt, F., Madau, P., & Volonteri, M. 2004, ApJ, 611, 623
Tanaka, T. L., & Li, M. 2014, MNRAS, 439, 1092
Treister, E., Schawinski, K., Volonteri, M., & Natarajan, P. 2013, ApJ, 778, 130
Volonteri, M., Habouzit, M., & Colpi, M. 2021, NatRP, 3, 732
Volonteri, M., Lodato, G., & Natarajan, P. 2008, MNRAS, 383, 1079
Volonteri, M., Pfister, H., Beckmann, R. S., et al. 2020, MNRAS, 498, 2219
Wang, H.-T., Jiang, Z., Sesana, A., et al. 2019, PhRvD, 100, 043003
Xu, H., Chen, S., Guo, Y., et al. 2023, RAA, 23, 075024
Zic, A., Reardon, D. J., Kapur, A., et al. 2023, PASA, 40, e049

Appendix 9

IX
J. Ellis, M. Fairbairn, G. Hütsi, J. Urrutia, V. Vaskonen, and H. Veermäe. Consistency of JWST black hole observations with NANOGrav gravitational wave measurements. *Astron. Astrophys.*, 691:A270, 2024

A&A, 691, A270 (2024)
https://doi.org/10.1051/0004-6361/202450846
© The Authors 2024

Astronomy & Astrophysics

Consistency of JWST black hole observations with NANOGrav gravitational wave measurements

John Ellis[1,2,3], Malcolm Fairbairn[2], Gert Hütsi[3], Juan Urrutia[3,4,⋆], Ville Vaskonen[3,5,6], and Hardi Veermäe[3]

[1] Theoretical Physics Department, CERN, Geneva, Switzerland
[2] King's College London, Strand, London, WC2R 2LS, UK
[3] Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia
[4] Departament of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia
[5] Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy
[6] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy

Received 23 May 2024 / Accepted 30 September 2024

ABSTRACT

JWST observations have opened a new chapter in supermassive black hole (SMBH) studies, stimulating discussion of two puzzles: the abundance of high-z SMBHs and the fraction of dual active galactic nuclei (AGNs). We argue that the answers to these puzzles may be linked to an interpretation of the data on the nanohertz gravitational waves (GWs) discovered by NANOGrav and other pulsar timing arrays as SMBH binaries whose evolution is driven by interactions with their environments down to $O(0.1\,\mathrm{pc})$ separations. We show that the stellar mass–black hole mass correlations found in JWST data and in low-z inactive galaxies are similar, and present a global fit to these data, excluding low-z AGNs. Matching the NANOGrav and dual-AGN data requires that binary evolution due to environmental effects at separations below $O(1\,\mathrm{kpc})$ be rapid on cosmological timescales. According to this interpretation, the SMBHs in low-z AGNs are the tip of the iceberg of a local SMBH population in mainly inactive galaxies. This interpretation is consistent with the 'little red dots' observed with JWST being AGNs, and would favour the observability of GW signals from black hole binaries in LISA and decihertz GW detectors.

Key words. quasars: supermassive black holes – cosmology: theory – early Universe

1. Introduction

There has been some surprise at the number of supermassive black holes (SMBHs) discovered with JWST at high redshifts, z (see Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024a; Yue et al. 2023). Prior expectations for the number of SMBHs in the early Universe were largely based on empirical relationships between the SMBH mass–halo mass ratio and the SMBH mass–stellar mass ratio, such as that found by Reines & Volonteri (2015, hereafter RV15) in the local (low-z) population of active galactic nuclei (AGNs). From a statistical analysis, Pacucci et al. (Pacucci et al. 2023, hereafter P23) conclude that there was a $>3\sigma$ discrepancy between the SMBH mass/stellar mass ratios, $M_{\rm BH}/M_*$, of a subset of the high-z JWST data and this fit to local AGN data. We note, however, that RV15 also analysed a sample of local inactive galaxies (IGs), finding higher values of $M_{\rm BH}/M_*$. Moreover, larger values of $M_{\rm BH}/M_*$ had also been found by Izumi et al. (2021, hereafter I21) in a compilation of high-z SMBHs observed previously by JWST.

Another surprise in the JWST data was the discovery of several close pairs in a sample of AGNs (Perna et al. 2023), whereas cosmological simulations generally predicted at least an order of magnitude fewer dual AGNs. This observation was used in Padmanabhan & Loeb (2024) to study SMBH binary evolution. JWST observations have also revealed a population of 'little red dots' (LRDs) in the redshift range $z \sim 2{-}11$ that are interpreted as previously hidden AGNs that trace black hole (BH) growth in the early Universe (Labbé et al. 2023b; Akins et al. 2023; Barro et al. 2024; Labbé et al. 2023a; Matthee et al. 2024; Kocevski et al. 2024).

In parallel with the emergence of JWST data, pulsar timing arrays (PTAs) have detected the existence of a stochastic gravitational wave background (SGWB) in the nanohertz range (Agazie et al. 2023c; Antoniadis et al. 2024; Reardon et al. 2023; Xu et al. 2023), whose most plausible astrophysical interpretation is gravitational wave (GW) emission by SMBH binaries. Analysing their data under this assumption and guided by astrophysical priors, the NANOGrav Collaboration (Agazie et al. 2023b) found that their data were best fit by a $M_{\rm BH}/M_*$ ratio that was significantly higher than that found in the RV15 analysis of local AGNs. In parallel, in Ellis et al. (2024a) we found – using the extended Press-Schechter formalism (Press & Schechter 1974; Bond et al. 1991) to estimate the halo merger rate and assuming a fixed probability, $p_{\rm BH}$, for a halo merger to yield a BH merger – that a good fit to the NANOGrav data was possible with the $M_{\rm BH}/M_*$ ratio found by RV15 in their analysis of local IGs. It was also found in Ellis et al. (2024a) that using the RV15 fit to local AGNs significantly degraded the fit to the NANOGrav data: $\Delta\chi^2 = 43$ compared to the fit based on local IGs (see Fig. 9 of Ellis et al. 2024a and the discussion in the accompanying text, as well as the analysis below).

It has been suggested recently that the $M_{\rm BH}/M_*$ relationship may evolve with z (Pacucci & Loeb 2024), or that selection biases and measurement uncertainties (Li et al. 2024) may account for much of the tension between the JWST high-z SMBH data and the data on local AGNs considered in RV15.

⋆ Corresponding author; juan.urrutia@kbfi.ee

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

However, we took the prima facie evidence at face value to be able to compare the information on SMBHs provided by the high-z SMBH data and NANOGrav measurements. The puzzle of the masses of JWST BHs may be linked to the somewhat unexpected strength of the GW signal reported by NANOGrav and other PTAs (Agazie et al. 2023a), and can be reframed as a question as to why both datasets correspond better to the data on local IGs than to local AGNs.

In this paper we quantify in more detail the consistency between (i) the NANOGrav nanohertz GW data and (ii) the JWST data and previous data on high-z SMBHs, discussing both the SMBH mass/stellar mass ratio and the fraction of dual AGNs in the JWST data. The paper is organised as follows: In Sect. 2 we analyse the BH mass-stellar mass relations found in recent high-z data from JWST and other observations, showing that they are quite consistent with data on low-z IGs. In Sect. 3 we show that the high-z SMBH data are also consistent with our analysis of the NANOGrav data on the nanohertz GW background (Ellis et al. 2024a). In Sect. 4 we show that our interpretation of the NANOGrav and JWST data can also accommodate the unexpectedly high fraction of dual AGNs observed with JWST (Perna et al. 2023), if the hardening timescale due to halo-halo interactions is short on cosmological timescales. We also show that our interpretation of the JWST and NANOGrav data is consistent with the population of LRDs reported recently (Matthee et al. 2024; Kocevski et al. 2024), assuming that they are AGNs triggered during mergers. We present our conclusions in Sect. 5.

2. Analysis of the BH mass–stellar mass relation

We considered the 33 high-z AGNs observed by JWST compiled in the appendix. In addition, we also considered a sample of 21 high-z AGNs observed previously, extracted from the observations compiled in I21. Both samples are shown in Fig. 1. The errors in the BH masses were reported in I21, but not the uncertainties in the stellar masses: we estimated them to be approximately 0.3 dex. The figure also shows the local IG and AGN measurements compiled in RV15. For clarity, we do not show the error bars for these points, but we take the uncertainties reported in RV15 into account in the following analysis.

The new JWST SMBH observations have been interpreted by P23 as inconsistent with the local mass relationship, prompting, for example, the proposals that the SMBH-stellar mass relationship evolves with redshift (Pacucci & Loeb 2024) or that selection biases and measurement uncertainties in the high-z population may be largely responsible for the apparent tension with the local relationship (Li et al. 2024). As mentioned earlier, rather than pursuing these hypotheses, we recall the tension between different local active and inactive SMBH populations RV15 and emphasise the consistency of the JWST measurements with the local IG measurements as well as the other high-z measurements.

As commonly done in the literature, we assumed a power-law form for the relation between the SMBH mass and the stellar mass of the host galaxy, with a log-normal intrinsic scatter. The probability distribution of the SMBH mass, m, for a given stellar mass, M_*, is then parametrised by the magnitude, a, and logarithmic slope, b, of the mean and the width, σ, of the distribution:

$$\frac{\mathrm{d}P(m|M_*,\boldsymbol{\theta})}{\mathrm{d}\log_{10}m} = \mathcal{N}\Big(\log_{10}\frac{m}{M_\odot}\Big|a+b\log_{10}\frac{M_*}{10^{11}M_\odot},\sigma\Big), \tag{1}$$

where $\boldsymbol{\theta} = (a, b, \sigma)$ and $\mathcal{N}(x|\bar{x},\sigma)$ denotes the probability density of a Gaussian distribution with mean $\bar{x}$ and variance σ^2.

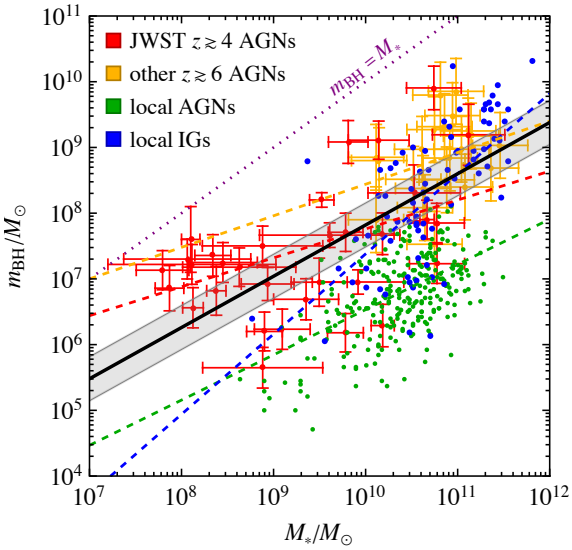


Fig. 1. Measurements and fits of the BH mass–stellar mass relation. The data points with errors are from JWST measurements (red; see the appendix) and other high-z measurements (yellow) from I21. The local AGN measurements (green) and the local IG measurements (blue), both from RV15, are shown without errors, but the errors are taken into account in the analysis. The dashed lines show the best fits to each of the datasets separately, and the solid black line shows the best fit combining the high-z data and the local IG data. The grey band shows the scatter in the latter fit.

We fitted the parameters a, b, and σ to the measured data. The likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_j \int \mathrm{d}\log_{10}m\,\mathrm{d}\log_{10}M\,\frac{\mathrm{d}P(m|M,\boldsymbol{\theta})}{\mathrm{d}\log_{10}m} \times \mathcal{N}(\log_{10}m|\log_{10}m_j)\mathcal{N}(\log_{10}M|\log_{10}M_j)\,, \tag{2}$$

where the product is over the data points, m_j and M_j denote the mean measured BH and stellar masses, and we assumed that the posteriors of the BH and stellar mass measurements are log-normal and uncorrelated. We used flat priors for a, b, and σ with $b > 0$.

As seen in Fig. 1, the high-z AGN measurements are largely consistent with each other and with the fit to the local IGs, though the latter prefers a slightly stronger tilt. The posteriors in a and b found in a global fit that includes all observations except the local AGNs, marginalised over σ, are shown in the upper panel of Fig. 2, and the posteriors of the fits to individual datasets are shown in the appendix. The lower panel of Fig. 2 displays uncertainty ellipses for the individual datasets as well as that for the global fit (shown in grey) that omits the local AGN data (shown in green). The global best fit is

$$a = 8.6, \quad b = 0.8, \quad \sigma = 0.8\,, \tag{3}$$

and the 68% CL parameter ranges are indicated by the vertical dashed lines in the upper panel of Fig. 2. The fit to the local AGN population prefers, instead, much lower a values but has a slope b similar to that in the global fit. We note that the linear relation ($b = 1$) is compatible with our global fit at the 68% CL.

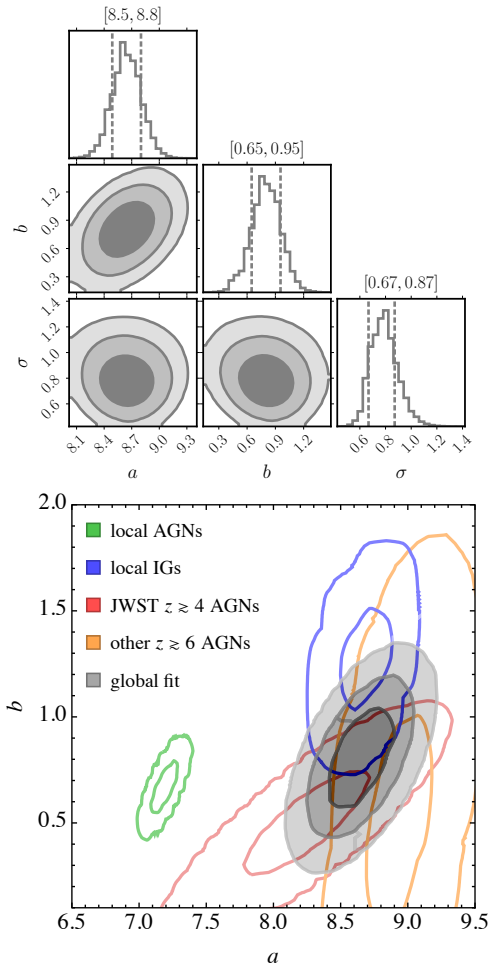


Fig. 2. *Upper panel*: corner plot showing the posteriors for the global fit to JWST (Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024a; Yue et al. 2023) and previous high-z SMBH data from I21 as well as low-z RV15 data on SMBHs in IGs. *Lower panel*: 68, 95, and 99.8% CL contours (grey) of the posteriors for a and b defined in Eq. (1), obtained in a global fit that includes all observations except local AGNs by marginalising over σ. For comparison, the 68 and 99.8% CL of individual datasets are shown: JWST high-z measurements (red; Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024a; Yue et al. 2023), other high-z measurements from I21 (yellow), local IGs (blue), and local AGNs from RV15 (green). The fits were performed using flat priors and including the uncertainties in the observations. Corner plots showing the posterior distributions for these fits are displayed in the appendix.

We quantified the overlaps between the posteriors for the fits to the different datasets i and j using the overlap coefficient

$$O_{i,j} = \int \mathrm{d}\boldsymbol{\theta} \min\left[\mathcal{L}_i(\boldsymbol{\theta}), \mathcal{L}_j(\boldsymbol{\theta})\right]. \tag{4}$$

For example, for two Gaussian distributions of similar width σ with a mean separation of 4σ the overlap coefficient is $O \approx 0.05$.

Table 1. Overlaps Eq. (4) between the posteriors obtained in the fits to different datasets.

	JWST $z \gtrsim 4$ AGNs	Other $z \gtrsim 6$ AGNs	Local AGNs	Local IGs	Global fit
JWST $z \gtrsim 4$ AGNs	1	0.07	10^{-6}	0.007	0.3
Other $z \gtrsim 6$ AGNs		1	10^{-9}	0.02	0.15
Local AGNs			1	10^{-9}	10^{-9}
Local IGs				1	0.08
Global fit					1

Notes. The cells of the table are colour-coded for clarity.

The overlap coefficients are tabulated in Table 1. We see numerically that the fit to the JWST data has better overlaps with the fits to the other high-z SMBHs and local IGs, as well as the global fit, than with the fit to local AGNs. The local AGN data overlap only slightly with the other high-z SMBHs, the local IGs, and the global fit.

3. Analysis of NANOGrav observations

We performed the analysis of the NANOGrav GW signal in the same way as in Ellis et al. (2023, 2024a,b). We summarise here the main features of the model: We related the SMBH merger rate to the halo merger rate, R_h, estimated from the Press-Schechter formalism (Press & Schechter 1974; Bond et al. 1991; Lacey & Cole 1993), as

$$\frac{\mathrm{d}R_{\rm BH}}{\mathrm{d}m_1\mathrm{d}m_2} \approx p_{\rm BH} \int \mathrm{d}M_{\rm v,1}\mathrm{d}M_{\rm v,2} \frac{\mathrm{d}R_h}{\mathrm{d}M_{\rm v,1}\mathrm{d}M_{\rm v,2}} \times \prod_{j=1,2} \frac{\mathrm{d}P(m_j|M_*(M_{{\rm v},j}, z))}{\mathrm{d}m_j}, \tag{5}$$

where $p_{\rm BH}$ characterises the BH merger efficiency and $M_*(M_{{\rm v},j}, z)$ is the halo mass-stellar mass relation that we took from Girelli et al. (2020). We included environmental effects parametrised by a spectral index, α, and a reference frequency,

$$t_{\rm int} = t_{\rm GW}\left[\frac{f_r}{f_{\rm GW}}\right]^{\alpha}, \tag{6}$$

where f_r is the frequency of the emitted GW. The environmental effects drive the binary evolution at frequencies above $f_{\rm GW}$, which we parametrised as $f_{\rm GW} = f_{\rm ref}(\mathcal{M}/M_\odot)^{-\beta}$. We fixed $\beta = 5/8$, corresponding to binary evolution by gas infall (see the discussion below Eq. (12)).

In Ellis et al. (2024a) we found that the local IG fit also gives a much better fit to the NANOGrav GW observations than the local AGN fit, and we now develop this point further. Figure 3 displays in a corner plot the posterior distributions for the parameters of this fit: α, $f_{\rm ref}$ and $p_{\rm BH}$. The posterior for α is compatible with models of environmental effects and the fit prefers large values of $f_{\rm ref} \gtrsim 10\,$nHz. The preferred range of $p_{\rm BH}$ is [0.16, 0.38]. The best fit is achieved at $p_{\rm BH} = 0.37$, $f_{\rm ref} = 27\,$nHz and $\alpha = 3.3$. In particular, the value $\alpha = 8/3$ expected for gas infall agrees well with the data. By fixing $\alpha = 8/3$, we find that the best fit is given by $p_{\rm BH} = 0.37$, $f_{\rm ref} = 30\,$nHz.

In the upper panel of Fig. 4 we compare our global fit Eq. (3) with the NANOGrav SGWB data (Agazie et al. 2023c) and the best fit SMBH binary GW signals obtained with the BH mass-stellar mass relations derived in P23 and Li et al. (2024). The

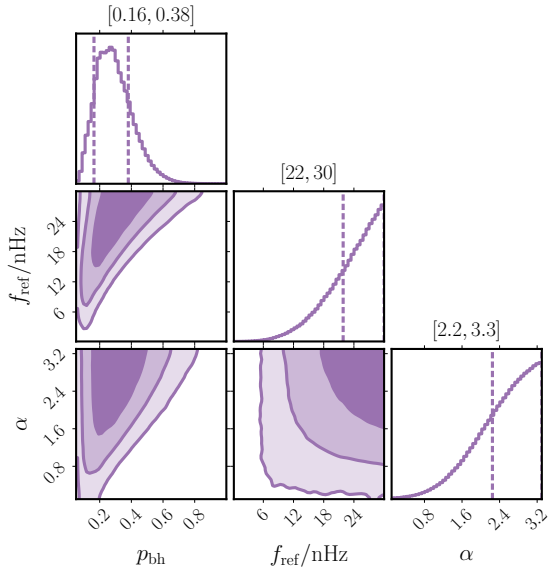


Fig. 3. Posterior distributions of the parameters for a fit to NANOGrav GW data (Agazie et al. 2023c) including energy loss due to interactions with the environment.

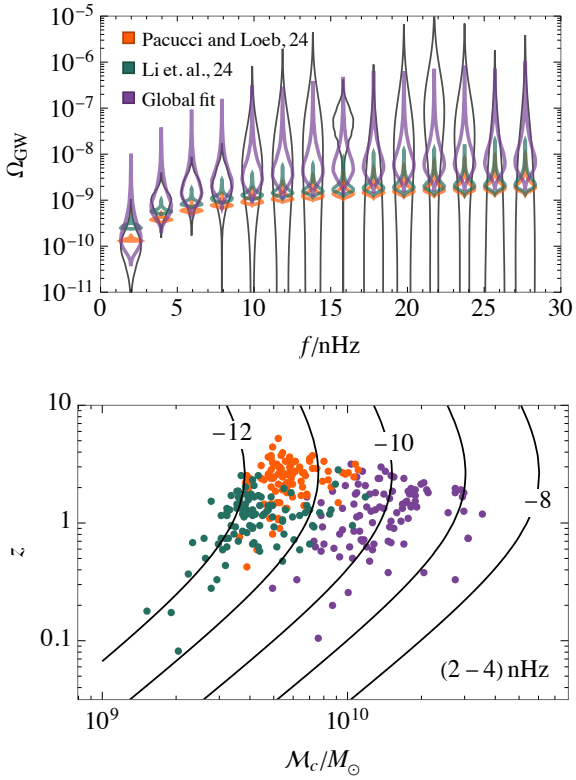


Fig. 4. *Upper panel*: best-fit nanohertz GW signals found using the BH mass–stellar mass relations proposed in Pacucci & Loeb (2024), Li et al. (2024), and our global fit (3) compared to the signal measured by NANOGrav (grey). We imposed $\alpha = 8/3$, as expected from gas infall (see the discussion below Eq. (12)). *Lower panel*: chirp masses and redshifts of 100 realisations of the strongest binary contributing to the first bin of the upper panel for the scenarios proposed in Pacucci & Loeb (2024), and Li et al. (2024) compared with our analysis (Ellis et al. 2024a). Contours of $\log_{10}\Omega_{\rm GW}$ are shown for comparison.

data and the best-fit simulations are represented by 'violins' whose widths represent the posteriors as functions of the GW density: following Ellis et al. (2024a), we calculated the overlaps between the violins to assess the qualities of the fits. For both P23 and Li et al. (2024), the best fits saturate at $p_{\rm BH} = 1$ (i.e. every galactic merger leads to a GW signal), which is an indication that the generated SGWB is insufficient. Comparing the best-fit likelihoods, we find that the PTA GW background favours the global fit mass relation at higher than the 3σ CL compared to the scaling relations proposed in Pacucci et al. (2023) and Li et al. (2024).

We show in the lower panel of Fig. 4 100 random realisations of the strongest binary contributing to the first NANOGrav bin, $f_{\rm GW} \in (2{-}4)$ nHz, for each of the three scenarios. In the case of Li et al. (2024), who advocated for a strong bias in the scaling relation, we can see that the SMBHs are too light, whereas in the case of Pacucci & Loeb (2024), who advocate for a steep z dependence, the heavy SMBH are merging mainly at $z > 1$, which also decreases the amplitude significantly.

Furthermore, in the lower panel of Fig. 4 we also see that BHs with $m_{\rm BH} < 10^9\,M_\odot$ are very unlikely to be among the strongest sources contributing to the SGWB. Given that the mass and z range probed by NANOGrav coincides with the dynamically measured SMBH masses, the high merger efficiency necessary to explain the observed GW background implies that the local IG population is not significantly biased. The local AGNs lie in a lower mass range, and it is an open question whether that population is biased. However, given the consistency between the high-z AGN measurements and the local IG measurements and the fact that at high z the AGN fraction is close to unity but drops quickly at $z \lesssim 1$ (Aird et al. 2018; Georgakakis et al. 2017), it seems that the local AGN population reflects the intrinsic low-mass tail of the full BH mass-stellar mass relationship.

Comparing the local AGN population to the global fit in Fig. 1 we see that most of the local AGNs lie more than 2σ below the mean of the global fit. Hence, according to this analysis, the AGN fraction in the local SMBH population can only be $\lesssim$5%, in agreement with X-ray observations (Aird et al. 2018; Georgakakis et al. 2017). To conclude, in contrast to Li et al. (2024), our analysis indicates that the local AGN population represents the tip of the iceberg of the local SMBH population. As discussed in Ellis et al. (2024a,c), our interpretation of the NANOGrav and JWST data would favour the observability of GW signals from BH binaries in space-borne laser interferometers, for example LISA (Amaro-Seoane et al. 2017), TianQin (Wang et al. 2019), and Taiji (Ruan et al. 2020), and decihertz GW detectors, for example AION (Badurina et al. 2020), AEDGE (El-Neaj et al. 2020), and DECIGO (Kawamura et al. 2021).

4. Dual AGN fraction

We found in Ellis et al. (2024a,c) that the quality of the SMBH fit to the NANOGrav data is improved by allowing for the

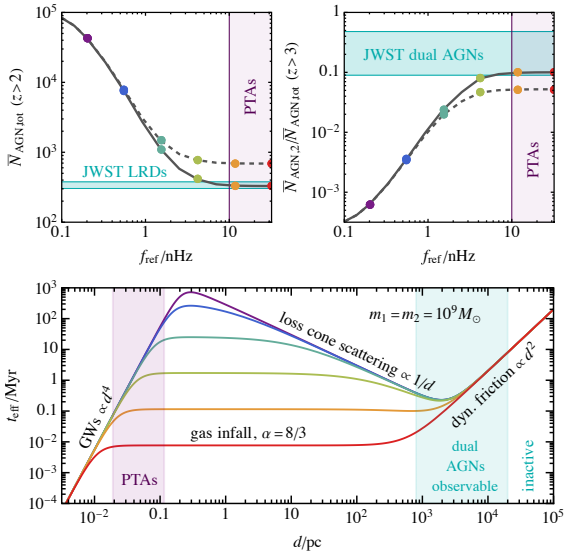


Fig. 5. *Upper panels*: expected total number of AGNs at $z > 2$ and the expected dual fraction at $z > 3$ as a function of the reference frequency, below which the binaries are driven by GW emission, assuming $p_{\rm BH} = 0.2$, $\alpha = 8/3$, $A_{\rm dyn} = 0.001$, and $A_{\rm sh} = 1$. The dashed and solid curves are with and without galaxy mergers and where one of the galaxies does not contain an SMBH. *Lower panel*: illustration of the timescales of the environmental energy loss mechanisms of SMBH binaries as functions of their separations for the same parameter values as in the upper panels. The vertical bands illustrate the separation ranges probed by the PTAs and by the JWST dual AGN observations. The colour coding of the curves matches that shown in the upper panels.

evolution of the binary system to be affected by interaction with the environment as well as the emission of GWs. A similar conclusion was reached by the NANOGrav Collaboration (Agazie et al. 2023b). We discuss now the extent to which the unexpectedly high fraction of dual AGNs observed with JWST (Perna et al. 2023) can be accommodated within our analysis of high-z SMBH data and NANOGrav SGWB measurements including environmental effects.

We estimated the numbers of dual AGNs and objects that we identified as single AGNs assuming that the collisions of galaxies trigger AGN activity. This is in agreement with numerical simulations (Scudder et al. 2012; Callegari et al. 2011; Capelo et al. 2015) and low-z observations (Treister et al. 2012; Comerford et al. 2015; Donley et al. 2018; Koss et al. 2012), except for a small number of objects (Liu et al. 2018; Mićić et al. 2023). In this way, the AGN observations are linked to the NANOGrav observations and to the BH mass-stellar mass relation through the galaxy merger rate. Environmental effects determine the evolution of the separations of the components in dual systems; therefore, to estimate the numbers of dual AGNs and objects seen as single AGNs, we started by characterising the effective timescales of different environmental effects.

Several such effects are expected to contribute to the binary evolution in different ranges of the binary separation (Kelley et al. 2017), as illustrated in Fig. 5. The effective timescale for the binary evolution is defined as $t_{\rm eff} \equiv |E|/\dot{E}$, where E is the binary total energy. The effective timescale can be expressed as

$$t_{\rm eff}^{-1} = t_{\rm GW}^{-1} + t_{\rm env}^{-1}, \tag{7}$$

where $t_{\rm GW}$ is the timescale of the GW driven binary evolution and $t_{\rm env}$ is the timescale of the environmental effects. For a circular orbit the binary total energy is given by $E = -m_1 m_2/d$ where d denotes the separation of the BHs, so, in terms of the effective timescale, the separation evolves as

$$\dot{d} = -d\, t_{\rm eff}^{-1}. \tag{8}$$

The dominant contribution to the binary evolution at small separations arises from GW emission, whose timescale is

$$t_{\rm GW} = \frac{5d^4}{1024\eta M^3} \approx \frac{14\,{\rm Myr}}{\eta}\left[\frac{M}{10^9 M_\odot}\right]^{-3}\left[\frac{d}{0.1\,{\rm pc}}\right]^4, \tag{9}$$

where η and M denote the symmetric mass ratio and the total mass of the binary. However, other energy-loss mechanisms come into play at separations probed by PTAs and other observations.

At kiloparsec separations, dynamical friction drives the binary evolution (Kelley et al. 2017). Its timescale is (Binney & Tremaine 2008)

$$t_{\rm dyn} \simeq \frac{20\,{\rm Myr}}{\ln\Lambda}\frac{\sigma}{200\,{\rm km/s}}\left[\frac{M}{10^9 M_\odot}\right]^{-1}\left[\frac{d}{\rm kpc}\right]^2, \tag{10}$$

where $\ln\Lambda$ is the Coulomb logarithm and σ is the velocity dispersion of stars.

Dynamical friction becomes inefficient at shorter distances when the binary binding energy becomes larger than the kinetic energy of the stars whose orbits pass close to the binary. The binary evolution is then driven by loss-cone scattering, whose timescale can be estimated as (Quinlan 1996)

$$t_{\rm sh} \simeq \frac{\sigma}{H\rho d} \approx 300\,{\rm Myr}\frac{\sigma}{200\,{\rm km/s}}\left[\frac{\rho}{10 M_\odot/{\rm pc}^3}\right]^{-1}\left[\frac{d}{\rm pc}\right]^{-1}, \tag{11}$$

where $H \approx 15$–20 is the dimensionless hardening rate, and ρ and σ denote the mass density and the velocity dispersion of stars in the neighbourhood of the binary.

At intermediate scales between the GW-driven and loss-cone scattering-driven phases, we used the parametrisation Eq. (6) that is given in terms of the binary separation as

$$t_{\rm int} = t_{\rm GW}\left[\frac{d}{d_{\rm ref}}\right]^{-\frac{3\alpha}{2}}. \tag{12}$$

The reference separation $d_{\rm ref}$ is related to the reference frequency used in Eq. (6) by $d_{\rm ref}^3 = 8M/(\pi f_{\rm ref}(\mathcal{M}/10^9 M_\odot)^{-\beta})^2$. We estimated the total timescale of the environmental effects as $t_{\rm env} = (t_{\rm int}^{-1} + t_{\rm sh}^{-1})^{-1} + t_{\rm dyn}$.

Effects that potentially determine the dynamics of binaries at intermediate scales include gas infall and viscous drag. The timescale of the binary evolution by gas infall can be estimated from the accretion timescale (Begelman et al. 1980), $t_{\rm gas} \simeq m_1/\dot{m}_1$, where m_1 is the mass of the heavier BH, and the timescale of viscous drag has been estimated in Ivanov et al. (1999) to scale as, $t_{\rm vd} \propto \sqrt{d}/M^{1/14}$. These correspond, respectively, to $(\alpha,\beta) = (8/3, 5/8)$ and $(\alpha,\beta) \approx (7/3, 3/4)$, while a transition directly from the loss-cone scattering to GW-driven evolution would correspond to $(\alpha,\beta) = (10/3, 2/5)$. For the PTA fit shown in Fig. 3 we fixed $\beta = 5/8$ after checking that the fit is not very sensitive to the value of β. We also find that the above values of $\alpha = 8/3$ and $\alpha = 10/3$ provide good fits to the data.

Given the estimates for the effective timescales, we can estimate the expected number of dual and single AGNs detected by JWST. As already noted, we assumed that AGN activity is triggered only during major mergers where the stellar mass of the heavier galaxy is not more than three times the stellar mass of the lighter one. Single AGNs are generated in major mergers in which one of the galaxies did not contain an SMBH. Moreover, dual AGNs are seen as single AGNs if one of the AGNs is not luminous enough or if their angular separation is too small to be resolved.

The expected number of detectable dual AGNs is

$$\bar{N}_{\rm AGN,2} = f_{\rm sky} \int {\rm d}\lambda_2 \, p_{\rm det}(m_1, z) p_{\rm det}(m_2, z) p_{\rm dual}(d, z), \tag{13}$$

where $f_{\rm sky}$ is the sky fraction being observed, $p_{\rm det}$ the probability that the BH is luminous enough to be detectable, and $p_{\rm dual}$ is the probability that a dual AGN is resolvable. Counting all mergers where at least one of the galaxies is occupied by an SMBH that becomes active during the merger and is sufficiently luminous to be detected, the expected number of detectable AGNs is

$$\begin{aligned} \bar{N}_{\rm AGN,tot} = {} & 2 f_{\rm sky} \int {\rm d}\lambda_1 \, p_{\rm det}(m, z) \\ & + 2 f_{\rm sky} \int {\rm d}\lambda_2 \, p_{\rm det}(m_2, z) \\ & - f_{\rm sky} \int {\rm d}\lambda_2 \, p_{\rm det}(m_1, z) p_{\rm det}(m_2, z). \end{aligned} \tag{14}$$

The merger rate and the evolution of the BH separation enter through

$${\rm d}\lambda_1 \equiv \frac{1}{1+z} \frac{{\rm d}V_c}{{\rm d}z} \frac{{\rm d}n^1_{\rm AGN}}{{\rm d}m{\rm d}d} {\rm d}m{\rm d}z{\rm d}d \tag{15}$$

and

$${\rm d}\lambda_2 \equiv \frac{1}{1+z} \frac{{\rm d}V_c}{{\rm d}z} \frac{{\rm d}n^2_{\rm AGN}}{{\rm d}m_1{\rm d}m_2{\rm d}d} {\rm d}m_1{\rm d}m_2{\rm d}z{\rm d}d, \tag{16}$$

where

$$\begin{aligned} \frac{{\rm d}n^1_{\rm AGN}}{{\rm d}m{\rm d}d} = {} & \frac{1 - \sqrt{p_{\rm BH}}}{|\dot{d}|} \int {\rm d}M_{\rm v,1} {\rm d}M_{\rm v,2} \frac{{\rm d}R_h}{{\rm d}M_{\rm v,1}{\rm d}M_{\rm v,2}} \\ & \times p_{\rm AGN}(M_{\rm v,1}, M_{\rm v,2}, d) \frac{{\rm d}p_{\rm occ}(m|M_*(M_{\rm v,2}, z))}{{\rm d}m} \end{aligned} \tag{17}$$

is the number density of AGNs formed in galaxy mergers where one of the galaxies did not contain an SMBH, and

$$\begin{aligned} \frac{{\rm d}n^2_{\rm AGN}}{{\rm d}m_1{\rm d}m_2{\rm d}d} = {} & \frac{1}{|\dot{d}|} \int {\rm d}M_{\rm v,1} {\rm d}M_{\rm v,2} \frac{{\rm d}R_h}{{\rm d}M_{\rm v,1}{\rm d}M_{\rm v,2}} \\ & \times p_{\rm AGN}(M_{\rm v,1}, M_{\rm v,2}, d) \prod_{j=1,2} \frac{{\rm d}p_{\rm occ}(m_j|M_*(M_{{\rm v},j}, z))}{{\rm d}m_j} \end{aligned} \tag{18}$$

is the number density of dual AGNs. The probability that a galaxy with stellar mass M_* is occupied by an SMBH with mass m can be expressed as

$$\frac{{\rm d}p_{\rm occ}(m|M_*)}{{\rm d}m} = \sqrt{p_{\rm BH}} \frac{{\rm d}P(m|M_*)}{{\rm d}m}, \tag{19}$$

where the distribution ${\rm d}P/{\rm d}m$, given by Eq. (1), is normalised to 1, so $\sqrt{p_{\rm BH}}$ can be interpreted as the SMBH occupation fraction,

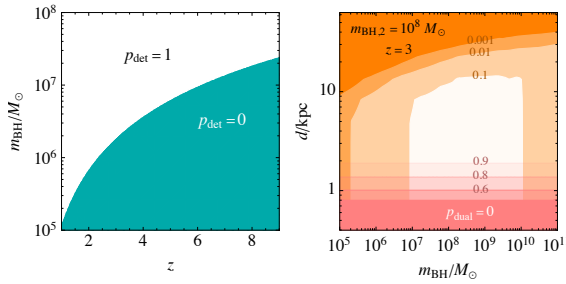


Fig. 6. *Left panel*: detectability probability, $p_{\rm det}$, arising from the JWST sensitivity threshold on the flux. *Right panel*: dual identification probability, $p_{\rm dual}$, at $z = 3$ arising from the JWST angular resolution, shown by the red regions. The orange regions show the probability that both galactic nuclei are active for $m_{\rm BH,2} = 10^8 M_\odot$ at $z = 3$.

which we took to be independent of M_*. The prefactor $1 - \sqrt{p_{\rm BH}}$ in Eq. (17) gives the probability that one of the merging galaxies does not contain an SMBH. In this case, dynamical friction drives this single SMBH towards the centre of the galaxy that resulted from the merger. So, for Eq. (17) we estimated the timescale as $t_{\rm eff} = t_{\rm dyn}$. We assumed that the AGN activity starts when the distance between the galaxies becomes smaller than the half-stellar radius of the smaller galaxy. We approximated the half-stellar radius as being 10% of the virial radius of the halo:

$$\begin{aligned} p_{\rm AGN}(M_{\rm v,1}, M_{\rm v,2}, d) = {} & \Theta(0.1 \min_j R_{{\rm v},j} - d) \\ & \times \Theta(3 \min_j M_{*,j} - \max_j M_{*,j}). \end{aligned} \tag{20}$$

In the right panel of Fig. 6 the shades of orange show the dual AGN probability ${\rm d}n^2_{\rm AGN}/{\rm d}n_{\rm tot}$ where ${\rm d}n_{\rm tot}$ is obtained from Eq. (18) with $p_{\rm AGN} = 1$.

We estimated the detection probability, $p_{\rm det}$, by setting a threshold on the flux:

$$p_{\rm det}(m_{\rm BH}, z) = \Theta\left(\frac{L}{4\pi D_L^2} - F_{\rm sens}\right), \tag{21}$$

where D_L denotes the luminosity distance of the system. We approximated the AGN luminosity optimistically as $L \simeq L_{\rm Edd}/3 \approx 4.3 \times 10^{43}\,(m_{\rm BH}/10^6 M_\odot)\,{\rm erg/s}$ and chose $F_{\rm sens} \simeq 10^{-15}\,{\rm erg/s/cm^2}$, which corresponds to the dimmest source found in Perna et al. (2023). The threshold on the flux gives a redshift-dependent lower bound on the BH mass, as shown in the left panel of Fig. 6.

We estimated the probability, $p_{\rm dual}$, by assuming that the duality of the AGNs is properly identified if its projected angular separation,

$$\theta(d, z) = \frac{d}{D_A(z)} \sqrt{1 - \sin^2 i \sin^2 \phi}, \tag{22}$$

where $D_A(z)$ denotes the angular diameter distance of the system and i and ϕ the inclination and phase of the dual, is larger than twice the angular resolution $\theta_{\rm res} \approx 0.1''$ of JWST, and we averaged over the angles i and ϕ:

$$p_{\rm dual}(d, z) = \frac{1}{4\pi} \int {\rm d}i \, {\rm d}\phi \sin i \, \Theta(\theta(d, z) - 2\theta_{\rm res}). \tag{23}$$

The probability $p_{\rm dual}$ is shown in the right panel of Fig. 6.

We write the timescales of dynamical friction and stellar loss-cone scattering as $t_{\rm dyn} = 20\,{\rm Myr}\,A_{\rm dyn}[M/10^9 M_\odot]^{-1}[d/{\rm kpc}]^2$ and $t_{\rm sh} = 300\,{\rm Myr}\,A_{\rm sh}[d/{\rm pc}]^{-1}$ (see Eqs. (10) and (11)), and study how the dual AGN fraction $\bar{N}_{\rm AGN,2}/\bar{N}_{\rm AGN,tot}$ and the total number of AGNs $\bar{N}_{\rm AGN,tot}$ depend on the parameters $A_{\rm dyn}$, $A_{\rm sh}$, $p_{\rm BH}$, $f_{\rm ref}$ and α. We considered the observations reported in Perna et al. (2023) of the GA-NIFS sample of 17 known AGNs at $z > 3$, of which $N_1 = 12$ were identified as single AGNs, $N_2 = 4$ as dual AGNs and one as a triple AGN. Based on these observations, we estimate that the dual fraction $N_2/(N_1 + N_2) = 0.25$. We also considered the JWST results of Kocevski et al. (2024) that cover 587.8 arcmin^2, corresponding to $f_{\rm sky} \approx 4 \times 10^{-6}$, revealing 341 LRDs in the redshift range $2 < z < 11$. Interpreting these LRDs as AGNs, we used this as an observation of the total AGN number, $N_{\rm tot} = 341$[1].

In the upper panels of Fig. 5 we show the dependences of the numbers of AGNs and fractions of dual AGNs on $f_{\rm ref}$ for $\alpha = 8/3$, corresponding to the gas infall timescale, $A_{\rm sh} = 1$, $A_{\rm dyn} = 0.001$ and $p_{\rm BH} = 0.2$. For these parameters, the lower panel of Fig. 5 illustrates the timescales of the environmental energy loss mechanisms, including GW emission, gas effects, stellar loss-cone scattering and dynamical friction, as functions of the binary separation. The different curves in the lower panel correspond to different values of $f_{\rm ref}$ indicated by the coloured dots in the upper panel. The horizontal bands in the upper panel show the 95% CL bands of the JWST LRD and dual AGN observations calculated assuming the Poisson and binomial distribution, respectively. The value of $p_{\rm BH}$ is fixed so that, neglecting AGNs that may form in mergers where only one of the galaxies contains an SMBH, the expected total AGN number roughly matches the LRD observations for large $f_{\rm ref}$. We see in the right panel that, in this benchmark case, the dual fraction is within the 95% range of the observed value for $f_{\rm ref} > 6\,{\rm nHz}$ if only the AGNs that may form in mergers where both of the galaxies contain an SMBH are considered. These values of $p_{\rm BH}$, α and $f_{\rm ref}$ are within the 95% CL region of the NANOGrav fit, illustrating that it is possible to obtain a coherent description of the evolution of binary systems that explains both the NANOGrav GW signal and the JWST dual AGN and LRD observations.

Figure 5 shows that the expected number of systems seen as single AGNs would be much larger than observed if $f_{\rm ref}$ is small, because in that case, the binaries would spend a lot of time orbiting with $d \sim 1\,{\rm pc}$ separation. This is related to the well-known final-parsec problem. Both the JWST and PTA observations prefer large values of $f_{\rm ref}$ for which the final-parsec problem is avoided. Then, the main contribution to the total number of AGNs comes from the same region as for dual AGNs, $d \gtrsim 1\,{\rm kpc}$, where the effective timescale is largest. In this regime, both $\bar{N}_{\rm AGN,2}$ and $\bar{N}_{\rm AGN,tot}$ are proportional to the prefactor $A_{\rm dyn}$ of the dynamical friction timescale. For the benchmark shown in Fig. 5 we have fixed $A_{\rm dyn} = 0.001$ so that the result is consistent with the LRD observations for values of $p_{\rm BH}$ consistent with the NANOGrav observations. Larger values of $A_{\rm dyn}$ or $p_{\rm BH}$ than those used in the above benchmark case would increase the total AGN number while the dual fraction would remain roughly the same. Larger values of $A_{\rm sh}$ would instead decrease the dual fraction and increase the total AGN number.

[1] The origin of the LRDs is debated because they are not detected in X-rays (Ananna et al. 2024). However, the non-detection of X-rays may be caused by high obscuration or intrinsic X-ray weakness of the AGNs (Maiolino et al. 2024b). Our objective is to determine whether the surprisingly high number density of LRDs can be accounted for under the assumption that they are AGNs whose nuclear activity is triggered during mergers. Determining, whether this mechanism could be responsible for triggering local AGN activity, exceeds the scope of this letter and is left for future work.

5. Conclusions

We have presented a global analysis of the available data on high-z SMBHs detected with JWST (Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024a; Yue et al. 2023) and in previous observations from I21, in conjunction with results from previous samples of SMBHs in low-z AGNs and IGs from RV15 and measurements of the SGWB background by NANOGrav (Agazie et al. 2023c), which are compatible with data from other PTAs (Agazie et al. 2023a). The JWST data are compatible with previous high-z SMBH data, and both are compatible with a sample of low-z IGs but incompatible with a low-z AGN sample (see Fig. 2 and Table 1). Moreover, we have shown that our global fit to the high-z SMBH and local IG data also fits the NANOGrav data on the nanohertz SGWB, whereas alternative interpretations of the high-z SMBH data that evoke an evolution of the $M_{\rm BH}/M_*$ relationship evolution with z (Pacucci & Loeb 2024) or selection biases and measurement uncertainties (Li et al. 2024) provide only poor fits to the SGWB data, as seen in Fig. 4.

We have also shown how the dual AGN fraction reported in Perna et al. (2023) can be accommodated within our global analysis. Our interpretation of the data on the dual AGN fraction (Perna et al. 2023) is related to the need for environmental effects to accommodate the spectral shape of the NANOGrav data (see Fig. 3 and the upper panel of Fig. 4 as well as Agazie et al. 2023b and Ellis et al. 2024a). Figure 5 shows that an extrapolation of these environmental effects to larger scales can accommodate the fraction of dual AGNs detected with JWST, can be seen in Fig. 5. Moreover, we also find that our global analysis is compatible with the abundance of LRDs in JWST observations reported recently (Matthee et al. 2024; Kocevski et al. 2024).

Observations of high-z SMBHs are still in their infancy, with much more data to come from JWST and other sources. Likewise, PTA observations of the nanohertz SGWB will develop rapidly in the near future. Doubtless, both sets of data will provide new puzzles. The central point of our paper is that these datasets should be considered together and that these puzzles may share common features. The resolution of these puzzles may also require inputs from other observational programmes, such as measurements of GWs at higher frequencies, interpreted within an integrated approach.

Acknowledgements. This work was supported by the European Regional Development Fund through the CoE program grant TK202 and by the Estonian Research Council grants PRG803, PSG869, RVTT3 and RVTT7. The work of J.E. and M.F. was supported by the United Kingdom STFC Grants ST/T000759/1 and ST/T00679X/1. The work of V.V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736.

References

Agazie, G., Antoniadis, J., Anumarlapudi, A., et al. 2023a, ApJ, submitted [arXiv:2309.00693]
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023b, ApJ, 952, L37
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023c, ApJ, 951, L8
Aird, J., Coil, A. L., & Georgakakis, A. 2018, MNRAS, 474, 1225
Akins, H. B., Casey, C. M., Allen, N., et al. 2023, ApJ, 956, 61
Amaro-Seoane, P., Audley, H., Babak, S., et al. 2017, arXiv e-prints [arXiv:1702.00786]
Ananna, T. T., Bogdán, A., Kovács, O. E., Natarajan, P., & Hickox, R. C. 2024, ApJ, submitted [arXiv:2404.19010]

Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2024, A&A, 685, A94
Badurina, L., Bentine, E., Blas, D., et al. 2020, JCAP, 05, 011
Barro, G., Pérez-González, P. G., Kocevski, D. D., et al. 2024, ApJ, 963, 128
Begelman, M. C., Blandford, R. D., & Rees, M. J. 1980, Nature, 287, 307
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition (Princeton: Princeton University Press)
Bogdan, A., Goulding, A. D., Natarajan, P., et al. 2024, Nat. Astron., 8, 126
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Callegari, S., Kazantzidis, S., Mayer, L., et al. 2011, ApJ, 729, 85
Capelo, P. R., Volonteri, M., Dotti, M., et al. 2015, MNRAS, 447, 2123
Comerford, J. M., Pooley, D., Barrows, R. S., et al. 2015, ApJ, 806, 219
Ding, X., Onoue, M., Silverman, J. D., et al. 2023, Nature, 621, 51
Donley, J. L., Kartaltepe, J., Kocevski, D., et al. 2018, ApJ, 853, 63
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2023, A&A, 676, A38
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2024a, Phys. Rev. D, 109, L021302
Ellis, J., Fairbairn, M., Urrutia, J., & Vaskonen, V. 2024b, ApJ, 964, 11
Ellis, J., Fairbairn, M., Franciolini, G., et al. 2024c, Phys. Rev. D, 109, 023522
El-Neaj, Y. A., Alpigiani, C., Amairi-Pyka, S., et al. 2020, EPJ Quant. Technol., 7, 6
Georgakakis, A., Aird, J., Schulze, A., et al. 2017, MNRAS, 471, 1976
Girelli, G., Pozzetti, L., Bolzonella, M., et al. 2020, A&A, 634, A135
Harikane, Y., Zhang, Y., Nakajima, K., et al. 2023, ApJ, 959, 39
Ivanov, P. B., Papaloizou, J. C. B., & Polnarev, A. G. 1999, MNRAS, 307, 79
Izumi, T., Matsuoka, Y., Fujimoto, S., et al. 2021, ApJ, 914, 36
Kawamura, S., Ando, M., Seto, N., et al. 2021, Prog. Theor. Exp. Phys., 2021, 05A105
Kelley, L. Z., Blecha, L., & Hernquist, L. 2017, MNRAS, 464, 3131
Kocevski, D. D., Finkelstein, S. L., Barro, G., et al. 2024, arXiv e-prints [arXiv:2404.03576]
Koss, M., Mushotzky, R., Treister, E., et al. 2012, ApJ, 746, L22
Labbé, I., Greene, J. E., Bezanson, R., et al. 2023a, arXiv e-prints [arXiv:2306.07320]
Labbé, I., van Dokkum, P., Nelson, E., et al. 2023b, Nature, 616, 266
Lacey, C. G., & Cole, S. 1993, MNRAS, 262, 627
Larson, R. L., Finkelstein, S. L., Kocevski, D. D., et al. 2023, ApJ, 953, L29
Li, J., Silverman, J. D., Shen, Y., et al. 2024, arXiv e-prints [arXiv:2403.00074]
Liu, X., Guo, H., Shen, Y., Greene, J. E., & Strauss, M. A. 2018, ApJ, 862, 29
Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024a, A&A, 691, A145
Maiolino, R., Risaliti, G., Signorini, M., et al. 2024b, arXiv e-prints [arXiv:2405.00504]
Matthee, J., Naidu, R. P., Brammer, G., et al. 2024, ApJ, 963, 129
Mićić, M., Wells, B. N., Holmes, O. J., & Irwin, J. A. 2023, arXiv e-prints [arXiv:2310.09945]
Pacucci, F., & Loeb, A. 2024, arXiv e-prints [arXiv:2401.04159]
Pacucci, F., Nguyen, B., Carniani, S., Maiolino, R., & Fan, X. 2023, arXiv e-prints [arXiv:2308.12331]
Padmanabhan, H., & Loeb, A. 2024, A&A, 684, L15
Perna, M., Arribas, S., Lamperti, I., et al. 2023, arXiv e-prints [arXiv:2310.03067]
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Quinlan, G. D. 1996, New Astron., 1, 35
Reardon, D. J., Zic, A., Shannon, R. M., et al. 2023, ApJ, 951, L6
Reines, A. E., & Volonteri, M. 2015, ApJ, 813, 82
Ruan, W.-H., Guo, Z.-K., Cai, R.-G., & Zhang, Y.-Z. 2020, Int. J. Mod. Phys. A, 35, 2050075
Scudder, J. M., Ellison, S. L., Torrey, P., Patton, D. R., & Mendel, J. T. 2012, MNRAS, 426, 549
Treister, E., Schawinski, K., Urry, C. M., & Simmons, B. D. 2012, ApJ, 758, L39
Übler, H., Maiolino, R., Curtis-Lake, E., et al. 2023, A&A, 677, A145
Wang, H.-T., Jiang, Z., Sesana, A., et al. 2019, Phys. Rev. D, 100, 043003
Xu, H., Chen, S., Guo, Y., et al. 2023, arXiv e-prints [arXiv:2306.16216]
Yue, M., Eilers, A.-C., Simcoe, R. A., et al. 2023, arXiv e-prints [arXiv:2309.04614]

Appendix A: Compilation of JWST data and corner plots

In Fig. A.1 we present corner plots that exhibit the posteriors for the parameters a, b and σ that were introduced in (1) to characterise the BH mass-stellar mass relation for fits to various datasets discussed in the main text. These include the JWST high-z SMBH data compiled in Fig. A.2 (upper left), the previous high-z SMBH data compiled in I21 (upper right), the RV15 data on local AGNs (lower left), and the RV15 data on dynamically measured SMBH (lower right). The latter two fits agree with the fits found by RV15.

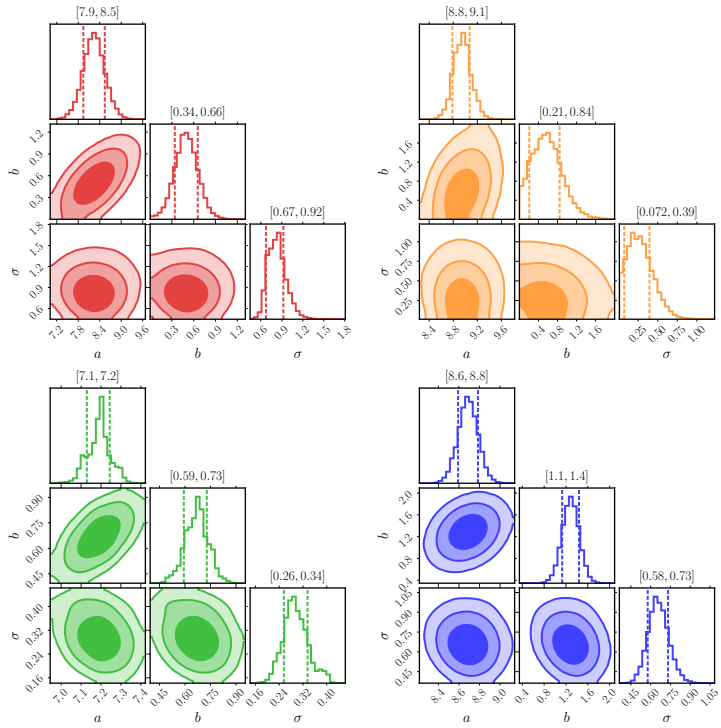


Fig. A.1. Posteriors for fits to the JWST high-z SMBH data compiled in Fig. A.2 *(upper left)*, previous high-z SMBH data compiled in I21 *(upper right)*, the RV15 data on local AGNs *(lower left)*, and the RV15 data on the dynamically measured SMBH *(lower right)*.

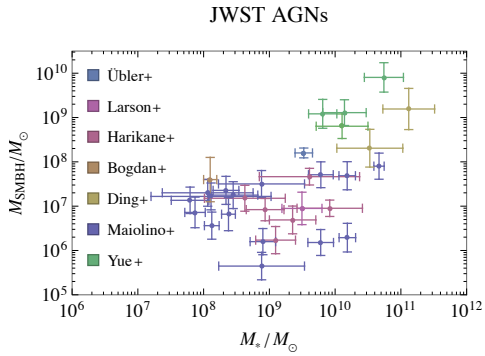


Fig. A.2. Scatter plot of the JWST data analysed in this paper. The following are the papers listed in the legend: Übler+ (Übler et al. 2023), Larson+ (Larson et al. 2023), Harikane+ (Harikane et al. 2023), Bogdan+ (Bogdan et al. 2024), Ding+ (Ding et al. 2023), Maiolino+ (Maiolino et al. 2024a), and Yue+ (Yue et al. 2023).

Appendix 10

X
J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. The origin of the JWST supermassive black holes. *Astron. Astrophys.*, 703:A138, 2025

A&A, 703, A138 (2025)
https://doi.org/10.1051/0004-6361/202555389
© The Authors 2025

Astronomy & Astrophysics

The origin of the JWST supermassive black holes

Juan Urrutia[1,2,⋆], John Ellis[3,4], Malcolm Fairbairn[3], and Ville Vaskonen[1,5,6]

[1] Keemilise ja Bioloogilise Füüsika Instituut, Rävala pst. 10, 10143 Tallinn, Estonia
[2] Department of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia
[3] King's College London, Strand, London WC2R 2LS, United Kingdom
[4] Theoretical Physics Department, CERN, Geneva, Switzerland
[5] Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy
[6] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy

Received 5 May 2025 / Accepted 6 August 2025

ABSTRACT

We present a new semi-analytical model for the evolution of galaxies and supermassive black holes (SMBHs) that is based on the extended Press-Schechter formalism and phenomenological modelling of star formation. The model yields black-hole-mass – stellar mass relations that reproduce both the JWST and pre-JWST observations. If the efficiency for BH mergers is high, the JWST data prefer light seeds, while the pre-JWST data prefer heavy seeds. The fit improves for a smaller merger efficiency, $O(0.1)$, for which both data prefer heavy seeds, while also accommodating the pulsar-timing-array gravitational-wave background data.

Key words. cosmology: theory

1. Introduction

The origin of supermassive black holes (SMBHs) is (literally) one of the biggest problems in astrophysics and cosmology (Volonteri et al. 2021). It has been thrown into the limelight recently by two sets of observations: the discovery of a (surprisingly strong) stochastic background of nanoHertz gravitational waves (GWs) by pulsar timing arrays (PTAs) (Agazie et al. 2023a; EPTA Collaboration and InPTA Collaboration 2024; Reardon et al. 2023; Xu et al. 2023) and the discovery of a (surprisingly heavy) population of SMBHs at high redshifts, $z \gtrsim 4$ by the JWST (Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023a; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024; Yue et al. 2024; Kocevski et al. 2023; Stone et al. 2024). The PTA signal is commonly thought to have been emitted by SMBH binary systems and points towards relatively high SMBH merger efficiency (EPTA Collaboration and InPTA Collaboration 2023; Agazie et al. 2023b; Ellis et al. 2024a,b; Bi et al. 2023; Raidal et al. 2024), although fundamental physics sources are also allowed by the available data (EPTA Collaboration and InPTA Collaboration 2024; Figueroa et al. 2024; Afzal et al. 2023; Ellis et al. 2024c). It has been suggested that the JWST observations may indicate a bias in either the high-z or local active galactic nuclei (AGN) observations (Li et al. 2025; Ellis et al. 2024b) and a preference for heavy SMBH seed scenarios (Bhowmick et al. 2024a; Pacucci et al. 2023; Natarajan et al. 2024). However, it has also been shown that the JWST observations can be explained by light seeds, as they may grow efficiently through mergers Bhowmick et al. (2024b). An ongoing debate centres on whether these SMBHs are accreting at super-Eddington rates. If so, their masses may actually have been overestimated (Lupi et al. 2024; Lambrides et al. 2024; Madau 2025).

In this paper, we address whether the puzzling JWST observations can be understood within the EPS formalism and, if so, the preferred range of BH masses seeding the SMBH assembly process. The co-evolution of SMBHs and galaxies has been studied in the literature with semi-analytical approaches using merger trees (Marulli et al. 2008; Dayal et al. 2019; Piana et al. 2020, 2024; Trinca et al. 2022; Toubiana et al. 2025; Volonteri et al. 2008; Bhowmick et al. 2022) and with large-scale N-body and hydrodynamic simulations (Croton et al. 2006; Schaye et al. 2015; Sijacki et al. 2015; Tremmel et al. 2017). Both of these approaches have a limited mass resolution and are computationally expensive. To overcome these limitations, we directly computed the growth rate of BHs and stellar masses in halo mergers in the EPS formalism. This allowed us to track the growth of BHs in small halos and explore the possible importance of light seeds. We also modelled the accretion of baryonic matter, accounting for supernova (SN) feedback that expels gas from light halos and AGN feedback that heats the gas and prevents star formation. We constrained the star formation rate to accommodate the measured galactic UV luminosity function. Our approach is computationally inexpensive, allowing for systematic scans of the SMBH seeding models[1].

The combined model of mergers and accretion yields a stellar mass-BH mass relation that is at high masses similar to those of the empirical relation found in a global fit to the high-z, PTA GW, and local IG data. Consequently, it is in agreement with the PTA GW observations (Ellis et al. 2024b), with a preferred merger efficiency[2] $p_{\rm BH} \sim 0.1$ as also suggested by modelling the PTA background (Ellis et al. 2024a). The predicted relation depends on the assumed masses of the BH seeds. We find that the JWST data are best described by a scenario with

⋆ Corresponding author: juan.urrutia@kbfi.ee

[1] The full evolution from $z = 20$ to $z = 0$, with $\Delta z = 0.1$, takes ≈2 min on an eight-core M2 Apple processor, resolving the halos down to 1 $M_\odot$.
[2] The merger efficiency $p_{\rm BH}$ describes the probability of a halo merger leading to a merger of the central BHs.

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

light seeds weighing $\sim 10^2 - 10^3\,M_\odot$ if the BH merger efficiency is high, $p_{\rm BH} \sim 1$, and the rest of the SMBH observations can be explained by SMBH seed masses ranging from $10^4 - 10^5\,M_\odot$. If the efficiency of merging is lower, $p_{\rm BH} \sim 0.1$, then heavier seeds with masses of $10^4 - 10^5\,M_\odot$ are preferred by both the JWST and the pre-JWST data. We also find that our model fits are better than power-law fits (Ellis et al. 2024b), and we discuss the effect of a scatter in the model predictions in this paper: the fit quality is improved, but the conclusions are not affected.

2. Co-evolution of SMBHs and galaxies

The evolution of the average BH mass and the average stellar mass in the dark-matter (DM) halos of mass M is governed by

$$\begin{aligned} \dot{M}_{\rm BH}(M) &= \dot{M}_{\rm BH}^{\rm merg}(M) + \dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M), \\ \dot{M}_{*}(M) &= \dot{M}_{*}^{\rm merg}(M) + \dot{M}_{*}^{\rm sf}(M), \end{aligned} \tag{1}$$

where $\dot{M}_J^{\rm merg}(M)$: $J = {\rm BH}, *$, denote the growth rates by mergers, $\dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M)$ is the BH growth rate by accretion, and $\dot{M}_*^{\rm sf}(M)$ is the stellar mass growth rate by star formation. The baryonic gas sources the latter two terms, and, consequently, the evolution of the BH and stellar masses is coupled through the SN and AGN feedback processes that can heat or eject parts of the baryonic gas (Bower et al. 2006, 2017; Dubois et al. 2015; Habouzit et al. 2017). We utilised a phenomenological fit to the star formation rate shown by the yellow curve in Fig. 1 (see Appendix B for details). This accounts for the SN and AGN feedback and is based on the results of Harikane et al. (2022), which has been tested successfully with new JWST data (Harikane et al. 2023b). Following Dayal et al. (2019), we consistently accounted for the effect of the SN feedback (Dayal et al. 2014), which strongly suppresses the BH accretion in halos lighter than $M \approx 5 \times 10^{11}\,M_\odot$ so that only the fraction $f_{\rm rem}$ shown by the red curve in Fig. 1 is available for accretion. Moreover, we limited the average accretion to the Eddington rate[3]. According to Piana et al. (2020) and Bower et al. (2017), these are approximately the conditions necessary to explain the UV AGN luminosity function, but rather than an ad hoc cut on the SMBH mass where accretion is started, it is a consequence of the interpretation of the star formation fit.

We used the EPS formalism to model the growth by mergers. The growth rate by mergers is given by the $\Delta z \equiv z' - z \to 0$ limit of (see Appendix A)

$$\begin{aligned} \Delta M_J^{\rm merg}(M, z, z') = &\int_0^M {\rm d}M' \left|\frac{{\rm d}S}{{\rm d}M'}\right| \frac{M}{M'} M_J(M', z') \\ &\times \frac{\delta_c(z') - \delta_c(z)}{\sqrt{2\pi[S(M') - S(M)]^3}} e^{-\frac{[\delta_c(z') - \delta_c(z)]^2}{2[S(M') - S(M)]}} - M_J(M, z'), \end{aligned} \tag{2}$$

where $J = {\rm BH}, *$, $S(M)$ denotes the variance of the DM perturbations and $\delta_c(z)$ is the threshold DM density contrast for halo formation. We assumed the cold DM model and computed $S(M)$ and $\delta_c(z)$ following Dodelson (2003) and Eisenstein & Hu (1998), with the cosmological parameters inferred from the CMB observations (Planck Collaboration VI 2020).

We also considered the possibility that not all mergers are efficient at forming a tight BH binary that merges quickly. We parametrised this with an efficiency parameter, $0 \le p_{\rm BH} \le 1$, which multiplies the BH growth rate by mergers if it is positive[4]. High merger efficiency, $p_{\rm BH} \simeq 1$, might be a good approximation if the SMBH is surrounded by a nuclear star cluster, which is very effective at lowering the orbital timescale (Mukherjee et al. 2025). However, it might be too optimistic for a more general case and should be considered an effective upper value. Studies of the efficiency of mergers induced by dark matter or three-body effects indicate a reduced BH merger efficiency $p_{\rm BH} \sim O(0.1)$ (Zhou et al. 2025), although with a big uncertainty[5].

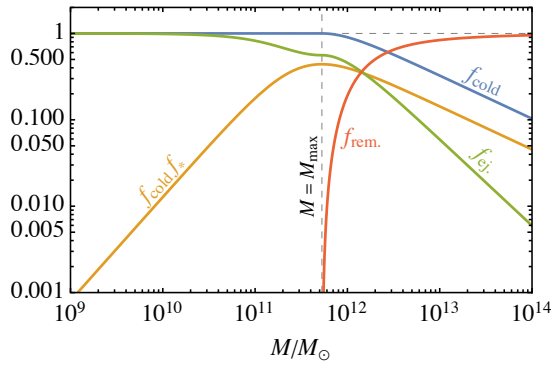


Fig. 1. Baryon fractions as functions of DM halo mass M. $f_{\rm ej}$ is the cold gas fraction that is ejected from the halo by SNe, $f_{\rm cold}$ is the fraction of the remaining gas that is cold, $f_{\rm rem}$ is the fraction of gas available for accretion, and f_* is the fraction of cold gas that is used for star formation.

We initiated the evolutionary process by planting a seed of mass $m_{\rm seed}$ at a given redshift ($z_{\rm seed}$) in every halo that is heavier than a given minimal mass, $M_{\rm seed}$, and we evolve the BH masses and stellar masses by solving the coupled Equations (1) numerically. We note that we only evolved the expected BH mass, $M_{\rm BH} = \int {\rm d}P(m_{\rm BH})\, m_{\rm BH}$, and do not include the inevitable scatter in the present computations. We therefore estimated the occupation fractions with Dirac delta functions:

$$\frac{{\rm d}P(m_{\rm BH}|M_*, z)}{{\rm d}m_{\rm BH}} = \frac{M_{\rm BH}(M_*, z)}{m_{\rm BH}} \delta\left[m_{\rm BH} - m_{\rm BH}(M_*, z)\right], \tag{3}$$

where

$$m_{\rm BH}(M_*, z) = \max\left[m_{\rm seed}, M_{\rm BH}(M_*, z)\right] \tag{4}$$

corresponds to the proper BH mass. To explore how a spread in the results could affect the comparison with the data, we can replace the Dirac delta function in Eq. (3) with a log-normal distribution, $\mathcal{N}(\log_{10} m_{\rm BH} | \log_{10} m_{\rm BH}(M_*, z), \sigma)/m_{\rm BH}$, where $\mathcal{N}(x|\bar{x}, \sigma)$ denotes the Gaussian distribution with mean $\bar{x}$ and the variances σ^2, as discussed below.

[3] We note that the BH accretion can include periods of super-Eddington accretion during galaxy mergers and periods of less efficient accretion. In our model, the growth through both mergers and accretions is described as an average over a long time.

[4] This approach has been used in, for example, Ricarte & Natarajan (2018). We emphasise that it is a simplification and could be improved by including a time delay with some spread between the halo merger and the merger of the central BHs, among other things. We leave such improvements for future work.

[5] This is also consistent, within large uncertainties, with the merger efficiency needed to explain the PTA GW background with inspiralling SMBH binaries (Ellis et al. 2024a).

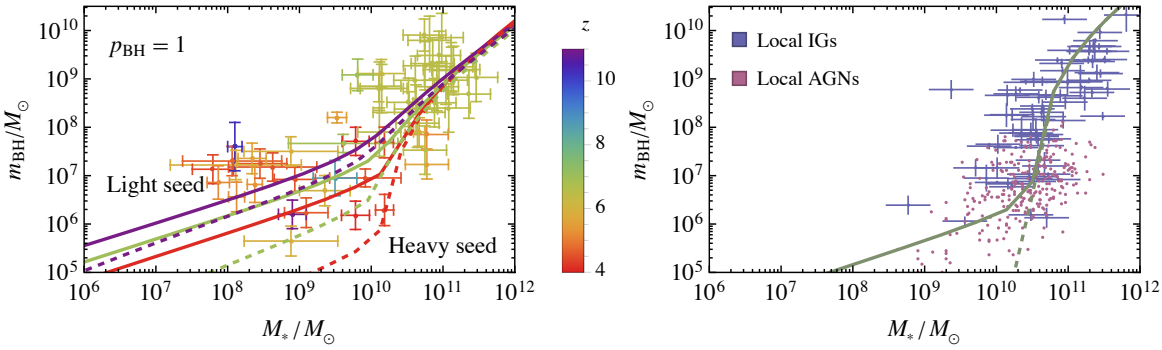


Fig. 2. Comparisons of model predictions for stellar mass-BH mass relation with high-z observations on the left and low-z observations on the right. The solid lines show the evolution of light seeds ($m_{\rm seed} = 100\,M_\odot$ with $M_{\rm seed} = 3\times10^4\,M_\odot$), and dashed lines show the evolution of heavy seeds ($m_{\rm seed} = 10^5\,M_\odot$ with $M_{\rm seed} = 3\times10^7\,M_\odot$), both for $p_{\rm BH} = 1$. The right panel shows the predictions of the same two models at $z \approx 0$ as full and dashed lines, respectively, compared to the local observations.

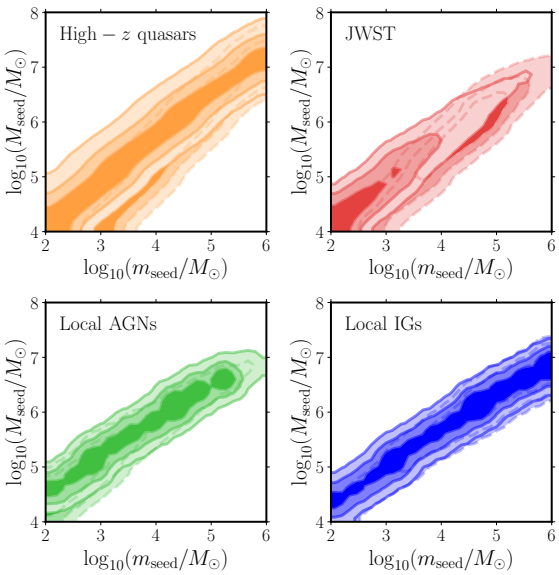


Fig. 3. Fits to stellar mass-BH mass data as functions of seed BH mass $m_{\rm seed}$ and minimal halo mass, $M_{\rm seed}$, where the seeds are inserted at $z_{\rm seed} = 20$ with $p_{\rm BH} = 1$ in solid and $p_{\rm BH} = 0.1$ in dashed. The contours indicate the 68%, 95%, and 99% credible regions. Upper left: High-z quasar data compiled by Izumi et al. (2021). Upper right: High-z observations with JWST (Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023a; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024; Yue et al. 2024). Lower left: Low-z AGNs as compiled by Reines & Volonteri (2015). Lower right: Low-z IGs from Reines & Volonteri (2015).

3. Implications of the data

We used the stellar mass-BH mass data from previous works (Übler et al. 2023; Larson et al. 2023; Harikane et al. 2023a; Bogdan et al. 2024; Ding et al. 2023; Maiolino et al. 2024; Yue et al. 2024; Kocevski et al. 2023; Stone et al. 2024) for the JWST AGNs, the high-redshift quasar masses compiled by Izumi et al. (2021), and the data concerning local AGNs and IGs (where SMBH masses are measured dynamically) compiled by Reines & Volonteri (2015). Examples of our model predictions for the stellar mass-BH mass correlation for $p_{\rm BH} = 1$ are shown alongside the data in Fig. 2. The benchmark cases shown correspond to examples with light seeds ($m_{\rm seed} = 100\,M_\odot$ and $M_{\rm seed} = 3\times10^4\,M_\odot$) and heavy seeds ($m_{\rm seed} = 10^5\,M_\odot$ with $M_{\rm seed} = 3\times10^7\,M_\odot$). The bend in the model predictions at $M_* \sim 10^{10}\,M_\odot$ arises because SN feedback expels all of the gas remaining after star formation below a critical halo mass, preventing efficient BH accretion. Therefore, growth via mergers is the dominant growth channel for BHs in such light galaxies. Light-seed scenarios have enough light BHs to fuel the growth of the AGNs observed by the JWST, but the heavy-seed model predictions undershoot the JWST observations for $M_* \lesssim 10^9\,M_\odot$. The high-$z$ quasar data with $M_* \gtrsim 10^{10}\,M_\odot$ compiled by Izumi et al. (2021) are, instead, fitted well by either light or heavy seeds. In the right panel of Fig. 2, we see that both the low-z IGs and the low-z AGN data are fitted well thanks to the steep fall of the model predictions at $M_* \sim 10^{11}-10^{10}\,M_\odot$ arising from the SN feedback.

We compare the qualities of our model predictions to the stellar mass-BH mass observations using the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_j \int \mathrm{d}\log_{10} m_{\rm BH}\,\mathrm{d}\log_{10} M_* \frac{\mathrm{d}P(m_{\rm BH}|M_*, z_j, \boldsymbol{\theta})}{\mathrm{d}\log_{10} m_{\rm BH}} \times \mathcal{N}(\log_{10} m_{\rm BH}|\log_{10} m_{{\rm BH},j}, \sigma_{{\rm BH},j}) \times \mathcal{N}(\log_{10} M_*|\log_{10} M_{*,j}, \sigma_{*,j}), \quad (5)$$

where the product is over the data points and the Gaussian probability densities $\mathcal{N}(x|\bar{x},\sigma)$ account for the uncertainties in the measurements of the BH masses and the galaxy stellar masses. Fixing all the parameters associated with star formation by observations, the free parameters are those associated with the formation of the first generation of massive BHs and the efficiency factor of the BH growth by mergers:

$$\boldsymbol{\theta} = (m_{\rm seed}\,, M_{\rm seed}\,, z_{\rm seed}, p_{\rm BH})\,. \quad (6)$$

We find that varying $z_{\rm seed} \in (15-20)$ does not significantly affect the likelihood. The following results were calculated for fixed $z_{\rm seed} = 20$.

We performed scans over $m_{\rm seed}$ and $M_{\rm seed}$ for fixed values of $p_{\rm BH} = 1$ and $p_{\rm BH} = 0.1$, as mentioned above. In Fig. 3, we

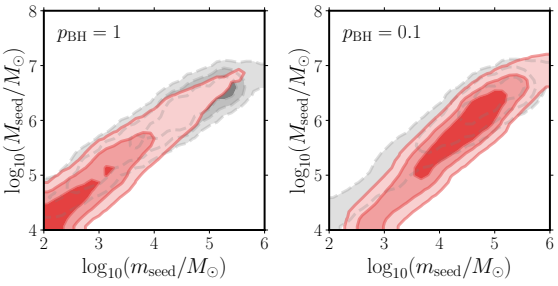


Fig. 4. Fits to stellar mass-BH mass data; the red posteriors show the fit to the JWST data, and the grey posteriors show the fit combining the low-z data and the high-z quasar data.

show the posterior distributions for each of the different data sets, ignoring a possible spread in the mass relation. The upper left plot displays results for fits to the high-z quasar data compiled by Izumi et al. (2021). We see that, whilst these data favour a strong correlation between $m_{\rm seed}$ and $M_{\rm seed}$, they do not favour any particular range of seed masses. This reflects the observation that the predictions of both the light- and heavy-seed models pass through the high-z SMBH data compiled by Izumi et al. (2021). On the other hand, as seen in the upper right plot, the JWST data show a preference for light seeds with masses of $m_{\rm seed} \lesssim 10^3\,M_\odot$ if $p_{\rm BH} = 1$ and heavier seeds with masses of $10^4\,M_\odot \lesssim m_{\rm seed} \lesssim 10^5\,M_\odot$ if $p_{\rm BH} = 0.1$. The low-z observations, shown in the lower panels, also show strong correlations between the seed halo mass and the seed BH mass, with the contours scaling roughly as $M_{\rm seed} \propto m_{\rm seed}^{0.6}$, but they do not constrain the seed BH mass significantly by themselves.

In Fig. 4, we compare the fit combining the low-z data and the high-z quasar data with the fit to the JWST high-z data. We note that the trend of the fit to the high-z quasar data is slightly steeper than for the low-z data, $M_{\rm seed} \propto m_{\rm seed}^{0.8-1}$, which is why the fit combining the low-z data and the high-z quasar data shown in grey in Fig. 4 has a preferred seed mass range. The JWST fit shows a preference for lighter seeds than the combined fit. This preference is stronger for $p_{\rm BH} = 1$ than for $p_{\rm BH} = 0.1$, but even for $p_{\rm BH} = 1$ the fits overlap significantly within the 2σ confidence level. In Fig. 5, we show the evolution of the $(m_{\rm seed}, M_{\rm seed})$ relation for the best fits to JWST data and the rest of the SMBH data for $p_{\rm BH} = 0.1$: parameter values are given in Table 1. We see how the JWST-only fit, shown as solid lines, traverses the data and has the correct z dependence. On the other hand, the fit to the rest of the data, shown as dashed lines, reproduces the local data better.

The JWST and other datasets exhibit significant scatter, with objects both above and below the model predictions, while the fits presented in Figs. 3 and 4. assume zero spread in the model predictions. To explore whether our results are sensitive to spread in our model predictions, we repeated the analysis including a log-normal spread with $\sigma = 0.5$ in the predicted SMBH mass. The fits are shown in the supplemental material. Our overall conclusions do not change, but the preference of the fit moves to slightly smaller seed masses and the quality of the fit improves. We collected best fits from the different scans in Table 1, which also shows the best fits obtained for a power-law fit as in Ellis et al. (2024b). Impressively, our model provides significantly better fits than the power-law fit with the same number of degrees of freedom. Moreover, we see that a spread in the model predictions clearly improves the quality of the fit, and the data prefer a lower merger efficiency of $p_{\rm BH} = 0.1$ over $p_{\rm BH} = 1$, independently of the spread. However, we also see that, in particular for $p_{\rm BH} = 0.1$, the seed BHs need to be relatively heavy compared to the minimal halo masses where they are planted; on the other hand, the JWST data become more compatible with the rest of the SMBH data.

Lambrides et al. (2024) (see also Pacucci & Narayan 2024) raised the possibility that the masses of at least some of the JWST SMBHs (Maiolino et al. 2024) may have been overestimated. It has been argued that the non-detection of measurable X-rays or UV emission lines suggests that these SMBHs may be lower mass objects accreting at super-Eddington rates. As seen in Fig. 2 of Lambrides et al. (2024), this argument would reduce the estimated masses of some SMBHs observed with the JWST (Maiolino et al. 2024) by perhaps an order of magnitude. This would shift the preference of the JWST data towards seeds planted in heavier halos.

4. Conclusions

In this paper, we presented a new semi-analytical model of the co-evolution of SMBHs and their host galaxies. This model is based on estimates of the growth rate of BH masses and the galaxy stellar masses derived from the EPS formalism and on phenomenological modelling of star formation, SN, and AGN feedback processes and BH accretion. In particular, we included the suppression of BH accretion in halos lighter than $\approx 5\times10^{11}\,M_\odot$ due to SN feedback and a parameter that describes how efficient the halo mergers are at forming a tight BH binary. Our approach involves solving numerically coupled differential equations for the evolution of the SMBH masses and the stellar masses and is much faster than approaches based on merger trees or numerical simulations. This allowed us to perform a systematic study of different SMBH seeding scenarios, including scans of the model parameter space.

Our model predicts BH mass-stellar mass relations that are compatible with the observations. We find that the heavy seed scenarios where BHs of mass $\sim10^5\,M_\odot$ are planted in halos heavier than $\sim10^7\,M_\odot$ undershoot the JWST observations because the BHs in halos lighter than $\approx 5\times10^{11}\,M_\odot$ grow predominantly by mergers, but there are not enough BHs to merge with. Scenarios with lighter seeds planted into lighter halos are preferred by the JWST as they allow for more efficient growth of the BHs by mergers.

We performed a likelihood analysis scanning over the seed BH masses and the masses of the smallest halos into which they are planted. As seen in Table I, our analysis suggests that the JWST data prefer lighter seeds, $m_{\rm seed} \lesssim 10^3\,M_\odot$ if the merging efficiency is high, $p_{\rm BH} = 1$, but the preference moves to higher masses, $10^4\,M_\odot \lesssim m_{\rm seed} \lesssim 10^5\,M_\odot$, for a lower merger efficiency of $p_{\rm BH} = 0.1$. The other data from local observations and observations of high-z quasars instead prefer the higher seed-mass range, $10^4\,M_\odot \lesssim m_{\rm seed} \lesssim 10^5\,M_\odot$, for both $p_{\rm BH} = 1$ and $p_{\rm BH} = 0.1$. The fits have two non-trivial successes. First, the fits are better for $p_{\rm BH} = 0.1$, which aligns with a fit to the PTA GW background (Ellis et al. 2024a). Secondly, our model provides a much better fit than those obtained with a simple power-law parametrisation (Ellis et al. 2024b), as also seen in Table I.

Our model does not resolve the spread in the BH masses, but we explored the effects of the spread by parametrising it. We find that including the spread significantly improves the fit to the observations, as also seen in Table I. This is expected given that the data exhibit significant scatter. The inclusion of the spread in

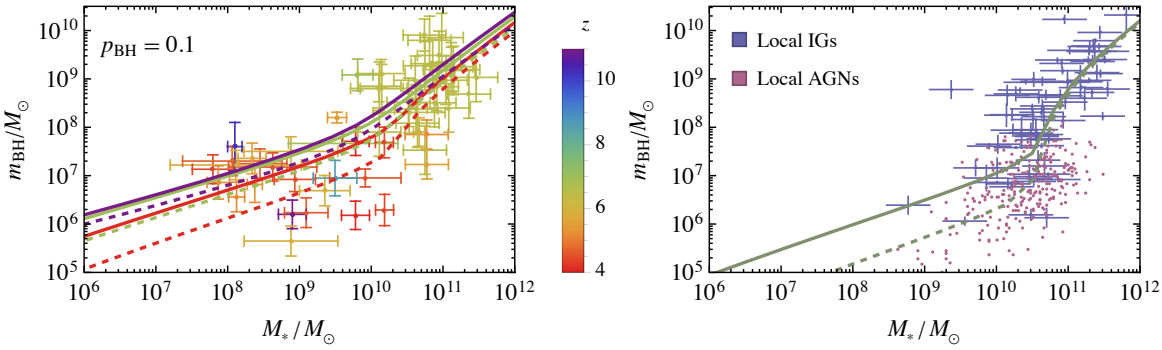


Fig. 5. Comparisons of model predictions for stellar mass-BH mass relation with high-z observations on the left and with low-z observations on the right. The solid lines show the best fit for $p_{\rm BH} = 0.1$ to the JWST data, and the dashed lines show the best fit to the rest of the SMBH data. The best-fit parameter values are given in Table 1.

Table 1. Likelihood comparison.

	$p_{\rm BH} = 1$ no spread	$p_{\rm BH} = 0.1$ no spread	$p_{\rm BH} = 1$ $\sigma = 0.5$	$p_{\rm BH} = 0.1$ $\sigma = 0.5$	power-law no spread	power-law $\sigma = 0.5$
JWST	$m_{\rm seed} = 10^2\,M_\odot$	$3\times10^4\,M_\odot$	$10^2\,M_\odot$	$2\times10^4\,M_\odot$	$a = 8.5$	8.5
	$M_{\rm seed} = 10^4\,M_\odot$	$5\times10^5\,M_\odot$	$10^4\,M_\odot$	$3\times10^5\,M_\odot$	$b = 0.64$	0.64
	$-2\log\mathcal{L} = 200$	200	70	60	$-2\log\mathcal{L} = 230$	140
pre-JWST	$3\times10^4\,M_\odot$	$6\times10^4\,M_\odot$	$5\times10^3\,M_\odot$	$4\times10^3\,M_\odot$	$a = 8.3$	8.0
	$10^6\,M_\odot$	$2\times10^6\,M_\odot$	$3\times10^5\,M_\odot$	$2\times10^5\,M_\odot$	$b = 1.8$	1.4
	1080	990	780	650	1100	970
all data	$3\times10^4\,M_\odot$	$2\times10^5\,M_\odot$	$5\times10^3\,M_\odot$	$10^5\,M_\odot$	$a = 8.3$	8.0
	$10^6\,M_\odot$	$3\times10^6\,M_\odot$	$3\times10^5\,M_\odot$	$2\times10^6\,M_\odot$	$b = 1.4$	1.0
	1290	1190	850	720	1820	1290

Notes. Four leftmost columns: Compilation of the best-fit and log-likelihood values of the model parameters described in the text. Two rightmost columns: Best fits for a power-law mean BH mass-stellar mass relation of $m_{\rm BH} = 10^a\,M_\odot(M_*/10^{11}M_\odot)^b$ with a log-normal spread defined as in Ellis et al. (2024b) are shown for comparison.

the model predictions also prefers slightly lower seed masses but does not change the overall conclusions (see Fig. C.3).

Building on the success of our model framework to study the scatter in the range of SMBH masses for any given value of the stellar mass is clearly a priority. Another interesting extension of this work would be to calculate the magnitude of the stochastic GW background due to BH binaries to higher frequencies than those explored by the PTAs that could be explored by LISA (Amaro-Seoane et al. 2017) and deciHz experiments (Badurina et al. 2020; El-Neaj et al. 2020; Kawamura et al. 2021). However, these explorations lie beyond the scope of this paper.

Acknowledgements. This work was supported by the Estonian Research Council grants PRG803, PSG869, RVTT3 and RVTT7 and the Center of Excellence program TK202. The work of J.E. and M.F. was supported by the United Kingdom STFC Grants ST/T000759/1 and ST/X000753/1. The work of V.V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101065736.

References

Afzal, A., Agazie, G., Anumarlapudi, A., et al. 2023, ApJ, 951, L11 [Erratum: ApJ 971, L27 (2024)]
Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023a, ApJ, 951, L8
Agazie, G., Alam, Md F, Anumarlapudi, A., et al. 2023b, ApJ, 951, L9
Amaro-Seoane, P., Audley, H., Babak, S., et al. 2017, arXiv e-prints [arXiv:1702.00786]
Badurina, L., Bentine, E., Blas, D., et al. 2020, JCAP, 05, 011
Bhowmick, A. K., Blecha, L., Ni, Y., et al. 2022, MNRAS, 516, 138
Bhowmick, A. K., Blecha, L., Torrey, P., et al. 2024a, MNRAS, 533, 1907
Bhowmick, A. K., Blecha, L., Torrey, P., et al. 2024b, MNRAS, 531, 4311
Bi, Y.-C., Wu, Y.-M., Chen, Z.-C., & Huang, Q.-G. 2023, Sci. China Phys. Mech. Astron., 66, 120402
Bogdan, A., Goulding, A. D., Natarajan, P., et al. 2024, Nat. Astron., 8, 126
Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, MNRAS, 370, 645
Bower, R. G., Schaye, J., Frenk, C. S., et al. 2017, MNRAS, 465, 32
Correa, C. A., Wyithe, J. S. B., Schaye, J., & Duffy, A. R. 2015, MNRAS, 450, 1514
Croton, D. J., Springel, V., White, S. D. M., et al. 2006, MNRAS, 365, 11 [Erratum: MNRAS 367, 864 (2006)]
Dayal, P., Ferrara, A., Dunlop, J., & Pacucci, F. 2014, MNRAS, 445, 2545
Dayal, P., Rossi, E. M., Shiralilou, B., et al. 2019, MNRAS, 486, 2336
Ding, X., Onoue, M., Silverman, J. D., et al. 2023, Nature, 621, 51
Dodelson, S. 2003, Modern Cosmology (Amsterdam: Academic Press)
Dubois, Y., Volonteri, M., Silk, J., et al. 2015, MNRAS, 452, 1502
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2024a, Phys. Rev. D, 109, L021302
Ellis, J., Fairbairn, M., Hütsi, G., et al. 2024b, A&A, 691, A270
Ellis, J., Fairbairn, M., Franciolini, G., et al. 2024c, Phys. Rev. D, 109, 023522
El-Neaj, Y. A., Alpigiani, C., Amairi-Pyka, S., et al. 2020, EPJ Quant. Technol., 7, 6
EPTA Collaboration and InPTA Collaboration (Antoniadis, J., et al.) 2023, A&A, 678, A50
EPTA Collaboration and InPTA Collaboration (Antoniadis, J., et al.) 2024, A&A,

685, A94
Figueroa, D. G., Pieroni, M., Ricciardone, A., & Simakachorn, P. 2024, Phys. Rev. Lett., 132, 171002
Habouzit, M., Volonteri, M., & Dubois, Y. 2017, MNRAS, 468, 3935
Harikane, Y., Ono, Y., Ouchi, M., et al. 2022, ApJS, 259, 20
Harikane, Y., Zhang, Y., Nakajima, K., et al. 2023a, ApJ, 959, 39
Harikane, Y., Ouchi, M., Oguri, M., et al. 2023b, ApJS, 265, 5
Izumi, T., Matsuoka, Y., Fujimoto, S., et al. 2021, ApJ, 914, 36
Kawamura, S., Ando, M., Ando, M., et al. 2021, PTEP, 2021, 05A105
Klessen, R. S., & Glover, S. C. O. 2023, ARA&A, 61, 65
Kocevski, D. D., Onoue, M., Inayoshi, K., et al. 2023, ApJ, 954, L4
Kulkarni, M., Visbal, E., & Bryan, G. L. 2021, ApJ, 917, 40
Lacey, C. G., & Cole, S. 1993, MNRAS, 262, 627
Lambrides, E., Garofali, K., Larson, R., et al. 2024, arXiv e-prints [arXiv:2409.13047]
Larson, R. L., Finkelstein, S. L., Kocevski, D. D., et al. 2023, ApJ, 953, L29
Latif, M. A., & Ferrara, A. 2016, PASA, 33, e051
Li, J., Silverman, J. D., Shen, Y., et al. 2025, ApJ, 981, 19
Lupi, A., Trinca, A., Volonteri, M., Dotti, M., & Mazzucchelli, C. 2024, A&A, 689, A128
Madau, P. 2025, ApJ, submitted [arXiv:2501.09854]
Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024, A&A, 691, A145
Marulli, F., Bonoli, S., Branchini, E., Moscardini, L., & Springel, V. 2008, MNRAS, 385, 1846
Mukherjee, D., Zhou, Y., Chen, N., Di Carlo, U. N., & Di Matteo, T. 2025, ApJ, 981, 203
Naoz, S., & Barkana, R. 2007, MNRAS, 377, 667
Natarajan, P., Pacucci, F., Ricarte, A., et al. 2024, ApJ, 960, L1
Pacucci, F., & Narayan, R. 2024, ApJ, 976, 96
Pacucci, F., Nguyen, B., Carniani, S., Maiolino, R., & Fan, X. 2023, ApJ, 957, L3
Piana, O., Dayal, P., Volonteri, M., & Choudhury, T. R. 2020, MNRAS, 500, 2146
Piana, O., Pu, H.-Y., & Wu, K. 2024, MNRAS, 530, 1732
Planck Collaboration VI. 2020, A&A, 641, A6 [Erratum: A&A 652, C4 (2021)]
Raidal, J., Urrutia, J., Vaskonen, V., & Veermäe, H. 2024, A&A, 691, A212
Reardon, D. J., Zic, A., Shannon, R. M., et al. 2023, ApJ, 951, L6
Reines, A. E., & Volonteri, M. 2015, ApJ, 813, 82
Ricarte, A., & Natarajan, P. 2018, MNRAS, 474, 1995
Schauer, A. T. P., Glover, S. C. O., Klessen, R. S., & Clark, P. 2021, MNRAS, 507, 1775
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
Sijacki, D., Vogelsberger, M., Genel, S., et al. 2015, MNRAS, 452, 575
Stone, M. A., Lyu, J., Rieke, G. H., Alberts, S., & Hainline, K. N. 2024, ApJ, 964, 90
Toubiana, A., Sberna, L., Volonteri, M., et al. 2025, A&A, 700, A135
Tremmel, M., Karcher, M., Governato, F., et al. 2017, MNRAS, 470, 1121
Trinca, A., Schneider, R., Valiante, R., et al. 2022, MNRAS, 511, 616
Übler, H., Maiolino, R., Curtis-Lake, E., et al. 2023, A&A, 677, A145
Volonteri, M., Lodato, G., & Natarajan, P. 2008, MNRAS, 383, 1079
Volonteri, M., Habouzit, M., & Colpi, M. 2021, Nat. Rev. Phys., 3, 732
Xu, H., Chen, S., Guo, Y., et al. 2023, Res. Astron. Astrophys., 23, 075024
Yue, M., Eilers, A.-C., Simcoe, R. A., et al. 2024, ApJ, 966, 176
Zhou, Y., Mukherjee, D., Chen, N., et al. 2025, ApJ, 980, 79

Appendix A: Going beyond halo merger trees

Although the EPS formalism is a convenient way to derive halo merger trees and thereby BH histories, it has its drawbacks, mainly that halos with masses below $\sim 10^8\,M_\odot$ are too abundant and computationally expensive to include (although this can be improved by an adaptive resolution (Toubiana et al. 2025)). In this paper, we aim to overcome these drawbacks by applying the EPS formalism directly to the evolution of SMBHs without invoking the halo merger tree step. The key insight behind this approach is that *a SMBH inside a halo on average, has had the same history as the average mass element of the halo.* This implies that the statistics from the random paths for the mass elements of a halo at some (M, z) are the same as for the SMBH inside that halo, which is the key assumption used when their evolution is calculated using halo merger trees.

In the cold dark matter model, the variance of matter perturbations $S(M)$ is a monotonically decreasing function of the halo mass M: $S(M) \to 0$ as $M \to \infty$. Thus, we can replace M by $S \equiv S(M)$ and the matter overdensity field δ around each spatial point evolves along a path in the $(S(M), \delta)$ plane that originates from $(0, 0)$. The first crossing of the path across a threshold value $\delta_c(z)$ determines the halo mass M around that point at redshift z. We compute the threshold $\delta_c(z)$ as in Dodelson (2003) and the variance $S(M)$ following Eisenstein & Hu (1998), with the cosmological parameters fixed by the CMB observations (Planck Collaboration VI 2020).

Assuming that the fluctuations are Gaussian, the probability a path to get to the point (S, δ) is $P(\delta|S) = e^{-\delta^2/2S}/\sqrt{2\pi S}$ and the probability that the first crossing is at variance larger than S is given by

$$\begin{aligned} P_{\rm FC}(S, z) &= 1 - \int_{-\infty}^{\delta_c(z)} {\rm d}\delta\, [P(\delta|S) - P(2\delta_c(z) - \delta|S)] \\ &= 1 - {\rm erf}\left[\frac{\delta_c(z)}{\sqrt{2S}}\right]. \end{aligned} \tag{A.1}$$

This is the cumulative mass fraction in halos lighter than M and the EPS halo mass function (HMF) is simply

$$\begin{aligned} \frac{{\rm d}n(z)}{{\rm d}M} &= \frac{\rho_{\rm DM}}{M}\left|\frac{{\rm d}S}{{\rm d}M}\right|\frac{{\rm d}P_{\rm FC}(z)}{{\rm d}S} \\ &= \frac{\rho_{\rm DM}}{M}\left|\frac{{\rm d}S}{{\rm d}M}\right|\frac{\delta_c(z)}{\sqrt{2\pi S^3}} e^{-\frac{\delta_c(z)^2}{2S}}, \end{aligned} \tag{A.2}$$

where $\rho_{\rm DM}$ is the present average DM energy density.

If the path is not initiated in $(0, 0)$ but instead at (S, δ), the probability to get to the point (S', δ'), $S' > S$ becomes $P(\delta'|\delta, S, S') = P(\delta' - \delta|S' - S)$. The probability for a path starting at (S, δ) to cross the threshold $\delta' = \delta_c(z') > \delta = \delta_c(z)$ at variance larger than S' is, in the same way as in (A.1), given by

$$P_{\rm FC}(\delta', S'|\delta, S) = 1 - {\rm erf}\left[\frac{\delta' - \delta}{\sqrt{2(S' - S)}}\right], \tag{A.3}$$

and the probability for the path to cross the threshold δ' in the range $(S', S' + {\rm d}S')$ is

$$\begin{aligned} p_{\rm FC}(\delta', S'|\delta, S){\rm d}S' &= \frac{{\rm d}P_{\rm FC}(\delta', S'|\delta, S)}{{\rm d}S'}{\rm d}S' \\ &= \frac{\delta' - \delta}{\sqrt{2\pi(S' - S)^3}} e^{-\frac{(\delta' - \delta)^2}{2(S' - S)}}{\rm d}S'. \end{aligned} \tag{A.4}$$

From this one can compute what fraction of the mass M of a halo at z on average comes from a halo of mass M' in the past, $z' > z$. If instead, we want to compute how the mass of the halo M' at z' is going to be distributed in heavier halos $M > M'$ in some later time $z < z'$, we need to invert the conditional probability (A.4). This gives the probability that a path that starts at (δ', S') crosses the threshold δ in the range $(S, S + {\rm d}S)$ (see Eq. (2.16) in Lacey & Cole (1993)):

$$\begin{aligned} p_{\rm FC}(\delta, S|\delta', S'){\rm d}S &= p_{\rm FC}(\delta', S'|\delta, S)\frac{p_{\rm FC}(\delta, S|0, 0)}{p_{\rm FC}(\delta', S'|0, 0)}{\rm d}S \\ &= p_{\rm FC}(\delta', S'|\delta, S)\left(\frac{S'}{S}\right)^{\frac{3}{2}}\left(\frac{\delta}{\delta'}\right)e^{\frac{\delta'^2}{2S'} - \frac{\delta^2}{2S}}{\rm d}S. \end{aligned} \tag{A.5}$$

A change of variables then gives the probability that the halo whose mass at z' is M' ends up being a part of a halo in the mass range $(M, M + {\rm d}M)$ at $z < z'$:

$${\rm d}P(M, z|M', z') \equiv p_{\rm FC}(\delta, S|\delta', S')\left|\frac{{\rm d}S}{{\rm d}M}\right|{\rm d}M. \tag{A.6}$$

This allows us to compute the evolution of the HMF forward in time from $z' > z$ to z as

$$\frac{{\rm d}n(z)}{{\rm d}M} = \int_0^M {\rm d}M' \frac{{\rm d}n(z')}{{\rm d}M'}\frac{M'}{M}\frac{{\rm d}P(M, z|M', z')}{{\rm d}M}. \tag{A.7}$$

It is easy to check that this formula holds by inserting the HMF (A.2) in the right-hand side and performing the integral. We estimate the growth rate of the DM halos by computing the expected growth of the halo through mergers with smaller halos:

$$\dot{M} = (1 + z)H(z)\lim_{\Delta z \to 0}\frac{1}{\Delta z}\left[\frac{\int_M^{2M}{\rm d}\tilde{M}\,\tilde{M}\frac{{\rm d}P(\tilde{M}, z|M, z+\Delta z)}{{\rm d}\tilde{M}}}{\int_M^{2M}{\rm d}\tilde{M}\,\frac{{\rm d}P(\tilde{M}, z|M, z+\Delta z)}{{\rm d}\tilde{M}}} - M\right]. \tag{A.8}$$

We show in Fig. A.1 DM halo growth rates calculated using Eq. (A.8) (solid lines) and, for comparison, estimates derived in Correa et al. (2015), which is based on finding the average mass of the main progenitor. The agreement is generally good over the range of halo masses M relevant for our study.

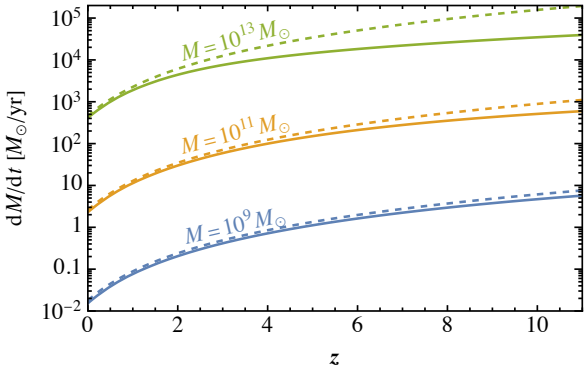


Fig. A.1. Calculations of the growth rates of DM halos. The solid curves are obtained using Eq. (A.8) and used in our analysis. The dashed curves show for comparison the estimates derived in Correa et al. (2015), which are based on finding the average mass of the main progenitor.

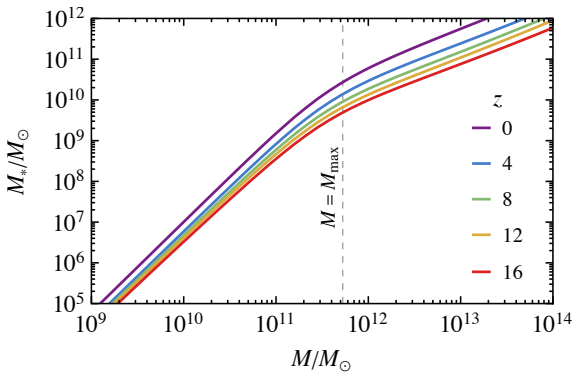


Fig. B.1. The evolution of the halo mass-stellar mass relation given by the star formation rate (B.5), the EPS estimate for the halo growth (A.8) and the growth of stellar masses by mergers (B.3).

Appendix B: SMBHs and baryonic physics

The average BH mass and the stellar mass evolve according to the coupled equations

$$\begin{aligned}\dot{M}_{\rm BH}(M, p_{\rm BH}) &= \dot{M}_{\rm BH}^{\rm merg}(M, p_{\rm BH}) + \dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M)\,,\\ \dot{M}_*(M) &= \dot{M}_*^{\rm merg}(M) + \dot{M}_*^{\rm sf}(M)\,,\end{aligned} \tag{B.1}$$

where $\dot{M}_J^{\rm merg}(M)$, $J = ({\rm BH}, *, {\rm gas})$, denotes the growth rate by mergers, $\dot{M}_{\rm BH}^{\rm acc}$ denotes BH accretion rate, and $\dot{M}_*^{\rm sf}$ is the star formation rate. The baryonic matter (stars and gas) and the SMBH have the same progenitors as the DM halo that hosts them. So, considering only mergers, their expected masses, M_J, evolve from z' to z in a halo of mass M as

$$\begin{aligned}&M_J(M, z') + \Delta M_J^{\rm merg}(M, z, z') = \\ &\left[\frac{{\rm d}n(z)}{{\rm d}M}\right]^{-1}\int_0^M {\rm d}M'\,\frac{{\rm d}n(z')}{{\rm d}M'}\frac{{\rm d}P(M, z|M', z')}{{\rm d}M}M_J(M', z') = \\ &\int_0^M {\rm d}M'\left|\frac{{\rm d}S}{{\rm d}M'}\right|\times \\ &\frac{M}{M'}M_J(M', z')\frac{\delta_c(z') - \delta_c(z)}{\sqrt{2\pi[S(M') - S(M)]^3}}e^{-\frac{[\delta_c(z') - \delta_c(z)]^2}{2[S(M') - S(M)]}}\,.\end{aligned} \tag{B.2}$$

The growth rate by mergers is given by the limit

$$\dot{M}_J^{\rm merg}(M) = (1+z)H(z)\lim_{\Delta z\to 0}\frac{\Delta M_J^{\rm merg}(M, z, z+\Delta z)}{\Delta z}\,. \tag{B.3}$$

For the growth rate of BHs by mergers, $\dot{M}_{\rm BH}^{\rm merg}(M, p_{\rm BH})$, we considered the possibility that not all mergers are efficient at forming a tight binary and the consequent BH is left orbiting at very high separations and it is no longer efficient at accreting or merging. In that case, we considered

$$\dot{M}_{\rm BH}^{\rm merg}(M, p_{\rm BH}) = \begin{cases} M_{\rm BH}^{\rm merg}(M) & <0 \\ p_{\rm BH}\,M_{\rm BH}^{\rm merg}(M) & \geq 0\end{cases}\,. \tag{B.4}$$

The part $1 - p_{\rm BH}$ of the accreted SMBH mass can be thought of as an SMBH that is left wandering in the galaxy or becomes ejected and does not play any significant role in future mergers.

Our aim in the following is to utilise phenomenological fits as much as possible while combining the different components in a physically consistent manner. We start from the star formation rate found in Harikane et al. (2022), based on fitting the UV luminosity function

$$\begin{aligned}\dot{M}_*^{\rm sf}(M, z) &= \frac{6.4\times 10^{-2}[0.53\tanh(0.54(2.9 - z)) + 1.53]}{(M/M_{\rm crit})^{-1.2} + (M/M_{\rm crit})^{0.5}}\dot{M}\\ &= f_*(M, z)f_{\rm cold}(M)f_B\dot{M}\end{aligned} \tag{B.5}$$

with $M_{\rm crit} = 10^{11.5}\,M_\odot$. Stars are formed from the cold baryonic gas. In the second line of (B.5) we have factorised the fit into the fraction of cold gas,

$$f_{\rm cold}(M) \equiv \begin{cases}1\,, & M < M_{\rm max}\\ \frac{1.83}{(M/M_{\rm crit})^{-1.2} + (M/M_{\rm crit})^{0.5}}\,, & M \geq M_{\rm max}\end{cases}\,, \tag{B.6}$$

and the effective star formation efficiency, which determines the fraction of cold gas that is converted to stars,

$$\begin{aligned}&f_*(M, z) \equiv [0.33 + 0.12\tanh(0.54(2.9 - z))]\times\\ &\begin{cases}\frac{1.83}{(M/M_{\rm crit})^{-1.2} + (M/M_{\rm crit})^{0.5}}\,, & M < M_{\rm max}\\ 1\,, & M \geq M_{\rm max}\end{cases}\,.\end{aligned} \tag{B.7}$$

The mass $M_{\rm max} = 5.3\times 10^{11}\,M_\odot$ corresponds to the maximum of $\dot{M}_*^{\rm sf}(M, z)$. Finally, we define the fraction of the cold gas that is ejected from the halo or pushed away from the central region so it is not available for the SMBH to accrete, by supernovae (SNe) as

$$f_{\rm ej}(M, z) \equiv (1 - f_*(M, z))\times\begin{cases}1\,, & M < M_{\rm max}\\ f_{\rm cold}(M, z)^2\,, & M \geq M_{\rm max}\end{cases}\,. \tag{B.8}$$

This means that the cold gas used for star formation is sufficient for the feedback from the resulting SNe to expel the remaining gas. Star formation reaches its maximum at $M = M_{\rm max}$, and the SNe are no longer able to eject all the gas at $M > M_{\rm max}$. Therefore, for $M > M_{\rm max}$ there is still gas available to fuel AGN activity, which has the effect of heating the gas in the galaxy, leading to a downturn in the star formation, as illustrated in Fig. 1. The evolution of stellar masses is fixed by the star formation rate (B.5), the EPS estimate for the halo growth (A.8) and the growth of stellar masses by mergers (B.3). We show the resulting evolution of the halo mass-stellar mass relation in Fig. B.1.

The amount of gas left in the halo after the star formation period is available for accretion by the SMBH. Following Dayal et al. (2019), we estimate the BH accretion rate as

$$\begin{aligned}&\dot{M}_{\rm BH}^{\rm acc}(M_{\rm BH}, M) = \min\{f_{\rm rem}(M, z)f_B(f_1^{\rm acc}\dot{M} + f_2^{\rm acc}M),\\ &f_{\rm Edd}\dot{M}_{\rm Edd}(M_{\rm BH})\}\,,\end{aligned} \tag{B.9}$$

where $f_1^{\rm acc} = 5.5\times 10^{-4}$ and $f_2^{\rm acc} = 5\times 10^{-5}$/Myr is the fraction of the available gas that can be accreted by the SMBH,

$$f_{\rm rem}(M, z) \equiv 1 - (f_{\rm ej}(M, z) + f_*(M, z))f_{\rm cold}(M) \tag{B.10}$$

is the total fraction of gas left after star formation and ejection of gas by SNe, $f_{\rm Edd}$ is the fraction of the Eddington rate at which the SMBHs can maximally accrete and the Eddington accretion rate is given by

$$\dot{M}_{\rm Edd}(M_{\rm BH}) = \frac{4\pi G M_{\rm BH} m_p}{\epsilon_r \sigma_T} \approx 2.2\times 10^{-3} m_{\rm BH}(M_{\rm BH})/{\rm Myr}\,, \tag{B.11}$$

where m_p is the proton mass and σ_T is the Thomson cross section. The physical interpretation is that in low-mass halos, the SN feedback prevents efficient accretion, and in heavy halos, the accretion is limited by the Eddington rate and the amount of available gas.

We initiate the evolution by planting a seed of mass $m_{\rm seed}$ at some redshift $z_{\rm seed}$ in all halos that are heavier than some minimal mass $M_{\rm seed}$. Note that we evolve only the expected BH mass, $M_{\rm BH} = \int {\rm d}P(m_{\rm BH})\, m_{\rm BH}$, and do not include the inevitable scatter in the present computations. We therefore estimate the occupation fractions with Dirac delta functions

$$\frac{{\rm d}P(m_{\rm BH}|M_*,z)}{{\rm d}m_{\rm BH}} = \frac{M_{\rm BH}(M_*,z)}{m_{\rm BH}}\delta\left[m_{\rm BH} - m_{\rm BH}(M_*,z)\right]\,, \tag{B.12}$$

where

$$m_{\rm BH}(M_*,z) = \max\left[m_{\rm seed}, M_{\rm BH}(M_*,z)\right] \tag{B.13}$$

corresponds to the proper BH mass. This is the BH mass that enters the accretion rate in Eq. (B.11) and that we compare with observations in the main text. For clarity, Table B.1 compiles the parameters related to the accretion and growth of the BHs.

Appendix C: Effects of spread and astrophysical implications for the seeds

As mentioned above, our model calculations are for means in the distribution of SMBH masses, and do not include the inevitable spread. Including this within the growth formalism we have described would cause the simple additive process to become a nonlinear convolution. In order to explore how a spread could change the results, we include an ad-hoc spread in the distribution of BH masses only in the final comparison with the data, but the evolution is computed only using the mean of the distribution. We use the lognormal distribution

$$\frac{{\rm d}P(m_{\rm BH}|M_*,z)}{{\rm d}\log_{10}m_{\rm BH}} = \frac{M_{\rm BH}(M_*,z)}{m_{\rm BH}}\times \\ \mathcal{N}(\log_{10}m_{\rm BH}|\log_{10}m_{\rm BH}(M_*,z),\sigma) \tag{C.1}$$

for this purpose, and assume $\sigma = 0.5$. Results from modifying the distribution in this way are shown in Fig. C.1 for $p_{\rm BH} = 1$, in Fig. C.2 for $p_{\rm BH} = 0.1$ and in Fig. C.3 for the JWST and the combination of all other data, and included in Table 1, and are commented in the main text.

We also compare the results from our fit to astrophysically motivated models for the seeds. In particular, we focus on $M_{\rm seed}$, the minimal halo mass to form SMBHs. The best fit to all the data was achieved at $p_{\rm BH} = 0.1$ and, as seen in Fig. C.4, the halo mass where the seeds are embedded aligns nicely with the simulations from Schauer et al. (2021), Kulkarni et al. (2021), and is compatible at the 2σ CL with the analytical estimation of atomic cooling halos at $z = 20$. It is also compatible at the 3σ CL with a very light halo mass, which has been found by Naoz & Barkana (2007) to be able to host Pop-III stars in simulations. This wide range of results from the simulations is what has motivated our prior range for $M_{\rm seed}$. The SMBH mass is the most uncertain part of the seeding process. The best fit seems to be achieved for "heavy" halos, close to the atomic cooling halos range, that require heavy seeds with masses $\sim 10^5\,M_\odot$, which could be associated with a direct collapse scenario, see Latif & Ferrara (2016) for a review. The typical light-seed scenario, in which heavy Pop-III stars collapse to BHs of $O(10^2\,M_\odot)$ (Klessen & Glover 2023) for a review, seems to be viable only if the Pop-III stars can form in very light minihalos of $O(10^4\,M_\odot)$. For heavier values of the minimal halo mass to form the Pop-III stars, either the Pop-III stars are very heavy $O(10^4\,M_\odot)$ or there are $O(10^2)$ per halo that can merge efficiently, otherwise the direct collapse scenario seems to be a more viable option.

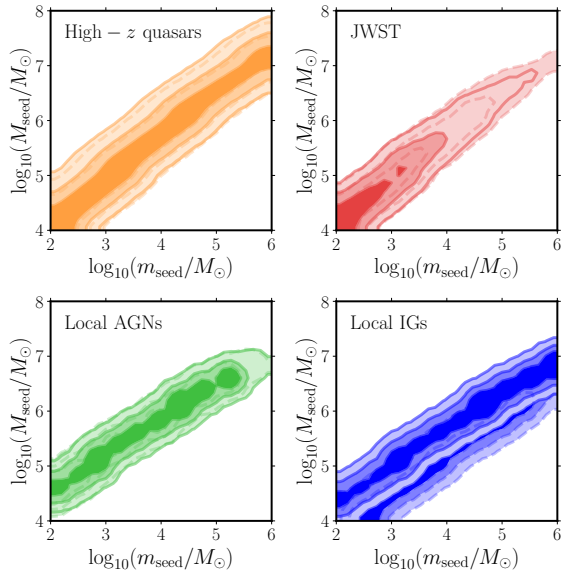


Fig. C.1. Fits to the stellar mass-BH mass data as functions of the seed BH mass $m_{\rm seed}$ and minimal halo mass $M_{\rm seed}$ where the seeds are inserted at $z_{\rm seed} = 20$, for $p_{\rm BH} = 1$ without spread in the predicted relation in solid and with a log-normal spread ($\sigma = 0.5$) in dashed.

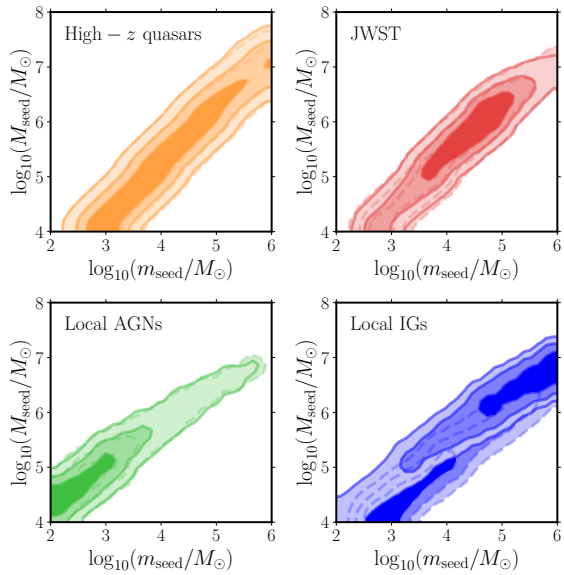


Fig. C.2. As in Fig. C.1 for $p_{\rm BH} = 0.1$.

Table B.1. Parameters of the model

Parameter	Value/Values	Physical interpretation
$p_{\rm BH}$	$(0.1\,,1)$	Probability that the merger of halos leads to the merger of the central SMBHs.
$f_{\rm rem}(M,z)$	$1-(f_{\rm ej}(M,z)+f_*(M,z))f_{\rm cold}(M)$	Fraction of baryons accreted by the DM halo that are not used for star formation or ejected by supernovae, see Fig.1.
$f_{\rm Edd}$	1	Upper limit on the Eddington ratio of the SMBH accretion rate.
$f_1^{\rm acc}$	5.5×10^{-4}	Fraction of baryons accreted by the halo that fall into the central SMBH (Dayal et al. 2019).
$f_2^{\rm acc}$	5.5×10^{-5}/Myr	Fraction of free baryons in the galaxy that are accreted by the central SMBH per unit time (Dayal et al. 2019).

Notes. Compilation of the parameters related to the accretion and growth of the BHs and their physical interpretations.

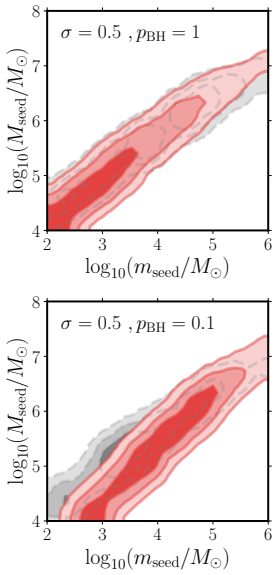


Fig. C.3. As in Fig. 4, allowing for a lognormal spread in the model predictions with $\sigma = 0.5$ for $p_{\rm BH} = 1$ in the left panel and for $p_{\rm BH} = 0.1$ in the right panel.

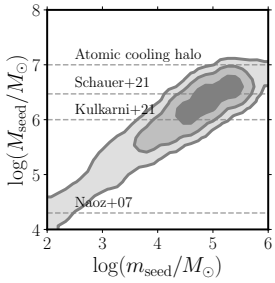


Fig. C.4. Comparison of our best fit to all the data, with $p_{\rm BH} = 0.1$, with respect to different values from simulations on $M_{\rm seed}$ (Schauer et al. 2021; Kulkarni et al. 2021; Naoz & Barkana 2007), as well as the atomic cooling halo estimation at $z = 20$. The implications of the fit for the viability of posteriors in astrophysical terms are discussed in Sec. C.

Appendix 11

XI
J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Constraints on Dark Matter Models from Supermassive Black Hole Evolution. *Phys. Rev. Lett.*, 136(13):131001, 2026

Constraints on Dark Matter Models from Supermassive Black Hole Evolution

John Ellis,[1,2,*] Malcolm Fairbairn,[1,†] Juan Urrutia,[3,4,‡] and Ville Vaskonen[3,5,6,§]
[1]*King's College London, Strand, London WC2R 2LS, United Kingdom*
[2]*Theoretical Physics Department, CERN, Geneva, Switzerland*
[3]*Keemilise ja Bioloogilise Füüsika Instituut, Rävala puiestee 10, 10143 Tallinn, Estonia*
[4]*Department of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia*
[5]*Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, Via Marzolo 8, 35131 Padova, Italy*
[6]*Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 3 October 2025; revised 27 January 2026; accepted 18 February 2026; published 30 March 2026)

A semianalytical model for the evolution of galaxies and supermassive black holes within the cold dark matter paradigm has been shown to yield stellar-black hole mass relations that reproduce both the James Webb Space Telescope and pre-Webb telescope observations. Either fuzzy or warm dark matter would suppress the formation of the smaller galactic halos that play important roles in the cold dark matter fit to the high-redshift supermassive black hole data. Our analysis of the stellar-black hole mass relation disfavors fuzzy dark matter fields with masses $< 2.0 \times 10^{-20}$ eV and warm dark matter particles with masses < 7.2 keV, both at the 95% confidence level.

DOI: 10.1103/njd2-3vdl

Introduction—The advent of the James Webb Space Telescope (JWST) has enabled surveys to probe the infrared universe with unprecedented accuracy [1–8], facilitating exploration of the reionization era and before (see Ref. [9] for a review). One of the most interesting developments has been new insight into proposed mechanisms for the formation of supermassive black holes (SMBHs) [10–12]. It had been suggested that they could have been seeded by black holes (BHs) with masses $\mathcal{O}(10^2\text{–}10^3)M_\odot$ that could be the remnants of population III (Pop-III) [13] stars that collapsed at redshifts $z \gtrsim 10$ (light-seed scenario), or by heavier BHs with masses $\mathcal{O}(10^5)M_\odot$ in the cores of protogalaxies that subsequently merged to form the larger galaxies seen at lower redshifts (heavy-seed scenario) [14,15].

Using JWST, a new population of high-z BHs has been discovered [16–24] that represent a challenge for the conventional picture of SMBH formation [25–28]. The high-z SMBHs seem to require either a heavy-seed scenario [29–31] or a very high efficiency for mergers of light seeds [32,33] in order for them to be a viable option. The possibility of growth through mergers inside nuclear clusters has also been tested against the JWST results [34]. An active topic of debate has been whether these SMBHs are accreting at super-Eddington rates, in which case their masses may have been overestimated [35–39]. In parallel, the discovery of "little red dots" [40–43] has defied many astrophysical assumptions about galaxies at high z [44–46]. There is an active debate centered around the nature of these objects and the role that SMBHs play in them [47–51].

This Letter explores the evolution of SMBH seeds in cold dark matter (CDM), fuzzy dark matter (FDM), and warm dark matter (WDM) universes, building on our earlier results for the star formation rate [52] and our semianalytical model for the coevolution of galaxies and SMBHs [32]. We compare model predictions to the stellar-black hole (BH) mass observations across a broad redshift range and conclude that, despite its uncertainties, SMBH evolution is a sensitive probe of dark matter (DM) properties, indeed more powerful than the stellar UV luminosity functions that we studied previously [52]. The reason is that the bright stars that dominate the UV emission are short-lived, so the UV luminosity function serves as a cosmic "snapshot" in which the lightest halos are too dim to be probed. In contrast, SMBHs are long-lived, carrying information from as early as $z \sim 20$ when they were first formed in light halos, which are very sensitive to the nature of DM.

We show here that the best-fit parameters for the evolution of the SMBH population in the CDM model are well within the ranges expected from simulations and physical arguments, making the observed SMBH population less surprising than previously thought. We further conclude that the current population of SMBHs disfavors FDM fields with masses $< 2.0 \times 10^{-20}$ eV and WDM particles with masses < 7.2 keV, both at the 95% confidence level. Finally, for the light-seed scenario to be viable,

*Contact author: john.ellis@cern.ch
†Contact author: malcolm.fairbairn@kcl.ac.uk
‡Contact author: juan.urrutia@kbfi.ee
§Contact author: ville.vaskonen@pd.infn.it

 © 2026 American Physical Society

TABLE I. List of the parameters in our model and their interpretations.

Parameter	Interpretation
$m_{\rm seed}$	Mass of the SMBH seed
$M_{\rm seed}$	Lowest halo mass that is planted with seeds
$p_{\rm BH}$	The probability that a halo merger ends in the merger of the central SMBHs
$f_{\rm SN}$	The fraction of gas that is not expelled after SN feedback
$f_{\rm Edd}$	The maximal allowed accretion as a fraction of the Eddington limit
$f_1^{\rm acc}$	The fraction of incoming baryons accreted onto the SMBH
$f_2^{\rm acc}$	The rate of accretion of baryons in the halo
$m_{\rm FDM/WDM}$	DM mass, if not CDM

it requires the buildup of light halos, making the light-seed case more sensitive to the nature of DM. This correlation between the nature of DM and the origin of SMBHs will provide an interesting phenomenology for future gravitational wave detectors like those of the Laser Interferometer Space Antenna [53] or Atom Interferometer Observatory and Network [54] projects, which we plan to explore in future work.

Methodology—For each DM model, we begin by embedding a seed of mass $m_{\rm seed}$ at some redshift $z_{\rm seed}$ in every halo that is heavier than some minimal mass $M_{\rm seed}$, and evolve the BH masses and stellar masses numerically using the formalism derived in Refs. [32,55] and summarized in Supplemental Material [56]. Our model includes the eight parameters described in Table I. We resolve halos in the range $M_{\rm v} = 10^2$–$10^{16} M_\odot$ and plant the seeds at $z_{\rm seed} = 20$. We have checked that our results are insensitive to this choice, and that the redshift steps and halo mass bins are small enough that the solutions are not affected. (The complete SMBH-galaxy coevolution from $z_{\rm seed}$ to $z = 0$ takes a few seconds in *Mathematica* on an M2 MacBook Air.)

In Fig. 1 we show the resulting stellar-BH mass relation at different redshifts for a benchmark case with $\log_{10}(m_{\rm seed}/M_\odot) = 5$, $\log_{10}(M_{\rm seed}/M_\odot) = 7$, $\log_{10} f_{\rm SN} = -3.0$, $\log_{10} f_1^{\rm acc} = -19.5$, $\log_{10}(f_2^{\rm acc}\ {\rm Myr}) = -3.0$, $\log_{10} f_{\rm Edd} = 0.5$, and $p_{\rm BH} = 0.1$. We display results for two different cosmologies, CDM and FDM, with $m_{\rm FDM} = 10^{-21}$ eV. We see in the lower panel that both DM models can approximately reproduce the observations at $z = 0$, but at high z, shown in the upper panel, the prediction is very sensitive to the DM model. There are two reasons for this. First, the abundance of low-mass halos is suppressed at the time of seeding for the non-CDM models [57], and, second, at these masses, the supernova (SN) feedback during reionization makes accretion inefficient, so mergers at very high redshift can be the main channel for the growth of these SMBHs [64], which is further suppressed. The

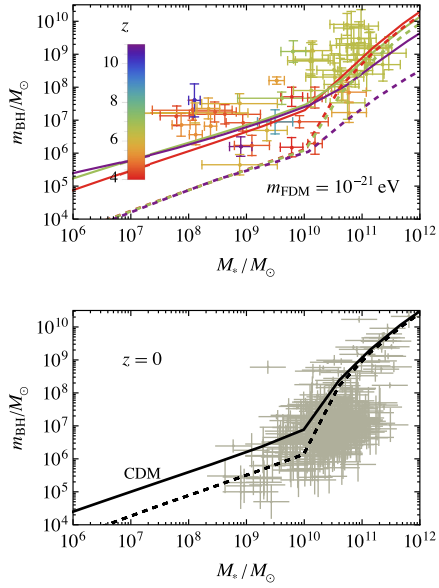


FIG. 1. Stellar-BH mass relation in a heavy-seed scenario. Solid lines show the CDM results and dashed lines the FDM result with $m_{\rm FDM} = 10^{-21}$ eV. The different colors in the top panel represent the different redshifts, while the lower panel represents the local universe.

effect of WDM (not shown) is qualitatively similar to that of FDM. This example illustrates how the suppression of small-scale structure abundance affects the evolution of SMBHs.

In Fig. 2 we show the average evolution of an SMBH in a DM halo whose mass today is $M = 10^{13} M_\odot$. The blue curves show a heavy-seed case, with the same values as for Fig. 1, and the orange curves are for a light-seed case where

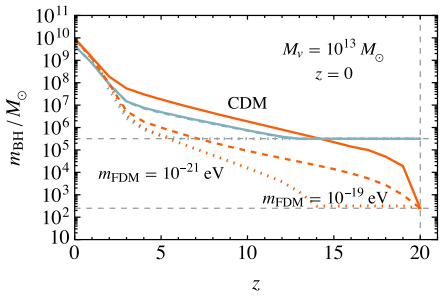


FIG. 2. Growth of individual SMBHs in both the light- and heavy-seed scenarios (orange and blue, respectively). CDM calculations are shown with solid curves and FDM with dashed ($m_{\rm FDM} = 10^{-19}$ eV) and dotted ($m_{\rm FDM} = 10^{-21}$ eV) curves. The host DM halo mass at $z = 0$ is fixed to $M_{\rm v} = 10^{13} M_\odot$.

we have taken $\log_{10}(m_{\rm seed}/M_\odot) = 2.5$, $\log_{10}(M_{\rm seed}/M_\odot) = 4.5$, and $p_{\rm BH} = 1$. The solid, dashed, and dotted curves correspond, respectively, to CDM, FDM with $m_{\rm FDM} = 10^{-19}$ eV and FDM with $m_{\rm FDM} = 10^{-21}$ eV. We see that the growth curves are similar at low redshifts and that the heavy-seed case grows steadily from $z = 12$, whereas in the CDM model, the light-seed case shows more rapid growth at the beginning. The reason is that a large number of light halos are simultaneously trapped in a collapsing overdensity that becomes a heavy halo [65]. When looking at the dashed and dotted curves, we see that the FDM model leads to a slower SMBH growth at high z than the CDM model, but that the predictions for low z are very similar. The light-seed case is more sensitive to the DM model since it relies more on the merging of light halos for the buildup of the seeds. If lighter FDM masses are considered, the same problem affects the heavy-seed case, as shown by the dotted curve.

Results—We perform a Markov-chain Monte Carlo analysis over the model parameters $\boldsymbol{\theta}$ (see Table I), comparing the predictions in the $m_{\rm BH}, M_*$ plane at each redshift to the observations [16–24,66,67]. In each case, we generate four Markov-chain Monte Carlo chains of 2×10^4 steps with 5×10^3 steps burn in, and check the convergence of the chains using the Gelman-Rubin statistic [68]. The likelihood is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) \propto \prod_j \int & \mathrm{d}\log_{10} m_{\rm BH} \mathrm{d}\log_{10} M_* \frac{\mathrm{d}P(m_{\rm BH}|M_*, z_j, \boldsymbol{\theta})}{\mathrm{d}\log_{10} m_{\rm BH}} \\ & \times \mathcal{N}\left(\log_{10} m_{\rm BH} | \log_{10} m_{{\rm BH},j}, \sigma_{{\rm BH},j}\right) \\ & \times \mathcal{N}\left(\log_{10} M_* | \log_{10} M_{*,j}, \sigma_{*,j}\right), \end{aligned} \tag{1}$$

where j labels the data points, the Gaussian probability densities $\mathcal{N}(x|\bar{x}, \sigma)$ account for the uncertainties in the measurements of the BH masses and the galaxy stellar masses, and $\mathrm{d}P(m_{\rm BH}|M_*, z_j)/\mathrm{d}\log_{10} m_{\rm BH}$ denotes the occupation fraction of BH masses in galaxies with stellar mass M_* at redshift z_j, which we obtain from the numerical computation described in the previous section. For each parameter, we consider log-uniform priors.

When analyzing the stellar-BH mass relation in CDM we see a strong correlation between $m_{\rm seed}$ and $M_{\rm seed}$, which is qualitatively compatible with the findings of our previous work [32]. The full posteriors can be seen in Supplemental Material [56] in Fig. S2. Our results for the CDM case are similar also to those of [69]: a wide range of model parameters can reproduce the low-z SMBH population, while the discrimination between different scenarios becomes significantly more pronounced at redshifts $z \gtrsim 10$. Both the light- and heavy-seed scenarios can provide good fits, but for light seeds of $m_{\rm seed} < 10^4 M_\odot$ they need to be planted in very light halos, $M_{\rm seed} \lesssim 10^6 M_\odot$. Whether such initial conditions are physical is up for debate [70]. Some studies have found such light halos to be able to host Pop-III stars [71] that could provide the BH seeds, although the majority of studies find higher halo mass thresholds [72,73]. The range of predictions from these studies falls within our 68.5% CL region.

We found that light seeds require high merger efficiency $p_{\rm BH} \simeq 1$, whereas for heavy seeds, the preference is $p_{\rm BH} \simeq 0.1$. Low values of $p_{\rm BH} \lesssim 0.1$ are excluded at the 68.5% CL. This range, $p_{\rm BH} \gtrsim 0.1$, is consistent with that required to reproduce the observed pulsar timing array (PTA) gravitational-wave background [74]. If SMBH binaries are driven solely by gravitational wave emission, the PTA data prefer $p_{\rm BH} \approx 0.17$ [74], which, as mentioned above, would indicate a preference for heavy seeds.

We find an upper bound of $f_{\rm SN} < 10^{-2.21}$ at the 68.5% CL on the efficiency of the SN feedback. If $f_{\rm SN}$ is too large, the SMBHs can accrete efficiently at all masses. This results in a power-law stellar-SMBH mass relation that does not give a good fit for the local active galactic nuclei, as has been seen in simulations that suppress the SN feedback [75]. The other accretion parameters $f_1^{\rm acc}$, $f_2^{\rm acc}$, and $f_{\rm Edd}$ are less constrained. We find that, as expected, there is a preference for either Eddington or super-Eddington growth. The correlation with the seed mass is not as strong as with $p_{\rm BH}$, but we find that the heavy-seed case does not have a preference for super-Eddington growth.

We perform similar analyses of the FDM and WDM cases, in which we consider the effect on structure evolution. We do not consider changes in the accretion model [76]. The resulting posteriors are shown in Fig. 3. We find lower bounds on the FDM and WDM masses that depend on the SMBH seed mass: the lighter the seeds are, the higher the FDM or WDM mass has to be. The same correlation between $m_{\rm seed}$ and $p_{\rm BH}$ that we found in the CDM case holds here, and the smaller the merging efficiency, the less stringent the DM mass bounds are. Marginalizing all parameters, we find that $m_{\rm FDM} < 2.0 \times 10^{-20}$ eV and $m_{\rm WDM} < 7.2$ keV are excluded at the 95% CL. If the correct seeding scenario is light seeds, something that could be tested with future gravitational wave detectors [54], then the bounds would become much more stringent.

In the case of FDM, the bound is as stringent as the bound coming from Lyman-α clouds, which implies $m_{\rm FDM} > 2 \times 10^{-20}$ eV [79]. Other cosmological and astrophysical bounds include $m_{\rm FDM} > 0.6 \times 10^{-19}$ eV from Eridanus II [80], $m_{\rm FDM} \gtrsim 3 \times 10^{-19}$ eV from Segue 1 and 2 [81], $m_{\rm FDM} \gtrsim 2.2 \times 10^{21}$ eV from Cetus-II [82], and $m_{\rm FDM} > 4.4 \times 10^{-21}$ eV from gravitational lensing [83]. Some works suggest evidence for FDM with $m_{\rm FDM} \lesssim 10^{-22}$ eV [84,85], which conflicts with these limits. As a function of the SMBH seed mass, the lower bounds are approximately $m_{\rm FDM} \gtrsim 10^{-20}\ \mathrm{eV}(m_{\rm seed}/10^6 M_\odot)^{-3/4}$.

As seen in Fig. 3, we find a very similar behavior in the WDM case. The current SMBH population implies

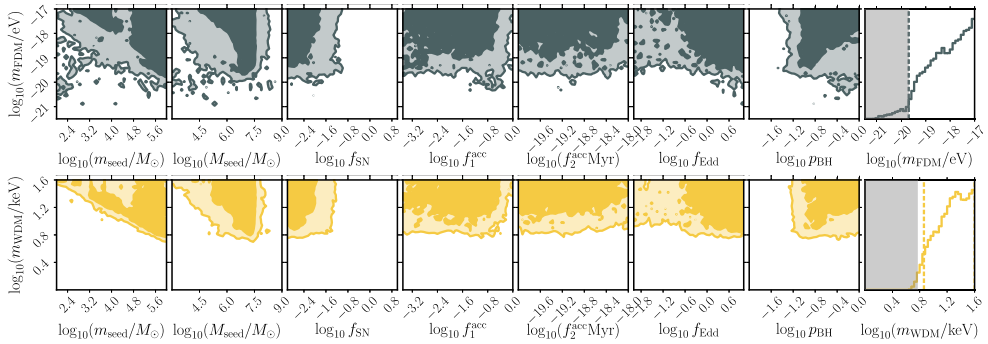


FIG. 3. Posteriors of the stellar-BH mass fit parameters as functions of the FDM mass (upper panels) and WDM mass (lower panels). The vertical dashed lines are our 95% CL lower bounds on these masses. The gray band in the $\log_{10}(m_{\rm FDM}/{\rm eV})$ posterior shows the most stringent 95%CL exclusion of $m_{\rm FDM}$ from a Lyman-α analysis [77], namely $< 2 \times 10^{-20}$ eV, and the gray band in the $\log_{10}(m_{\rm WDM}/{\rm keV})$ posterior shows the most stringent 95%CL exclusion of $m_{\rm WDM}$ from a joint analysis of strong gravitational lensing, the Lyman-α forest, and Milky Way satellites [78], namely < 6.0 keV. The full corner plots are shown in Supplemental Material [56].

$m_{\rm WDM} > 7.2$ keV, which is slightly more stringent than the bound coming from a combined analysis of strong gravitational lensing, Lyman-α clouds, and Milky Way satellites [78] of $m_{\rm WDM} > 6.0$ keV. Again we see a correlation with the seeding mass of the form $m_{\rm WDM} \gtrsim 10^{0.7}\ {\rm keV}(m_{\rm seed}/10^6 M_\odot)^{-0.225}$. Other constraints include: $m_{\rm WDM} > 5.3$ keV from Lyα data [86], $m_{\rm WDM} > 0.8$ keV from lensing at $z \sim 0.2$ [87], $m_{\rm WDM} > 5.6$ keV from quasar lensing [88], $m_{\rm WDM} > 3.0$ keV from quadruple-image lenses [89], $m_{\rm WDM} > 3.6$ keV from Milky Way satellites [90], and $m_{\rm WDM} > 3.2$ keV from JWST UV luminosity data [91].

Conclusions—We have found that SMBHs could be better discriminators of DM scenarios than stars [52]. The reason is that, while stars are short-lived, SMBHs live practically forever. Consequently, the SMBH population retains imprints of their seeding scenario. Since seeding occurred at very high redshifts in low-mass halos, these imprints are ideal for testing deviations from CDM.

Leveraging this insight, we have used the latest data from the JWST on both the stellar population, which informs the internal galactic processes, and the masses of the galaxies that host the SMBHs, in conjunction with previous data, to put new bounds on DM. We have found that FDM masses of $m_{\rm FDM} < 2.0 \times 10^{-20}$ eV and WDM masses of $m_{\rm WDM} < 7.2$ keV are excluded at the 95% CL.

We have also studied the plausibility of the SMBH growth model against the observed SMBH population. We have found that the JWST population can be accommodated and that the best-fit parameters fall within the range of expected values from simulations, physical arguments, and the PTA gravitational-wave background strength. We also find that the feasibility of light seeds does not depend only on their ability to grow at super-Eddington rates (which is favored but not required), but on whether they are planted in sufficiently light halos. These protogalaxies will eventually be trapped in an overdensity, which will cause the small galaxies to merge into a bigger one. If the SMBHs that take part can merge efficiently enough, they can become viable seeding candidates. This makes the bounds on DM dependent on the seeding scenario, and light seeds provide more stringent bounds on deviations from CDM. This is an interesting phenomenon that could be explored with future gravitational-wave detectors such as those of the Laser Interferometer Space Antenna [53] and Atom Interferometer Observatory and Network [54] projects.

Acknowledgments—This work was supported by the Estonian Research Council Grants No. PRG803, No. PSG869, No. RVTT3, and No. RVTT7 and the Center of Excellence program TK202. The work of J. E. and M. F. was supported by the United Kingdom STFC Grants No. ST/T000759/1 and No. ST/T00679X/1. The work of V. V. was partially supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101065736.

Data availability—The data are not publicly available. The data are available from the authors upon reasonable request.

[1] S. L. Finkelstein, M. B. Bagley, P. Arrabal Haro, M. Dickinson, H. C. Ferguson, J. S. Kartaltepe *et al.*, The cosmic evolution early release science survey (CEERS), Astrophys. J. Lett. **983**, L4 (2025).

[2] C. T. Donnan, R. J. McLure, J. S. Dunlop, D. J. McLeod, D. Magee, K. Z. Arellano-Córdova *et al.*, JWST PRIMER: A new multifield determination of the evolving galaxy UV luminosity function at redshifts $z = 9 - 15$, Mon. Not. R. Astron. Soc. **533**, 3222 (2024).

[3] M. B. Bagley, N. Pirzkal, S. L. Finkelstein, C. Papovich, D. A. Berg, J. M. Lotz *et al.*, The next generation deep extragalactic exploratory public (NGDEEP) survey, Astrophys. J. Lett. **965**, L6 (2024).

[4] C. M. Casey, J. S. Kartaltepe, N. E. Drakos, M. Franco, S. Harish, L. Paquereau *et al.*, COSMOS-web: An overview of the JWST cosmic origins survey, Astrophys. J. **954**, 31 (2023).

[5] D. J. Eisenstein, C. Willott, S. Alberts, S. Arribas, N. Bonaventura, A. J. Bunker *et al.*, Overview of the JWST advanced deep extragalactic survey (JADES), arXiv:2306.02465.

[6] M. J. Rieke, B. Robertson, S. Tacchella, K. Hainline, B. D. Johnson, R. Hausen *et al.*, JADES initial data release for the Hubble ultra deep field: Revealing the faint infrared sky with deep JWST NIRCam imaging, Astrophys. J. Suppl. Ser. **269**, 16 (2023).

[7] G. Östlin *et al.*, MIRI deep imaging survey (MIDIS) of the Hubble ultra deep field—survey description and early results for the galaxy population detected at 5.6 μm, Astron. Astrophys. **696**, A57 (2025).

[8] R. Bezanson, I. Labbe, K. E. Whitaker, J. Leja, S. H. Price, M. Franx *et al.*, The JWST UNCOVER treasury survey: Ultradeep NIRSpec and NIRCam observations before the epoch of reionization, Astrophys. J. **974**, 92 (2024).

[9] A. Adamo, H. Atek, M. B. Bagley, E. Bañados, K. S. S. Barrow, D. A. Berg *et al.*, The first billion years according to JWST, arXiv:2405.21054 [Nat. Astron. (to be published)].

[10] T. E. Woods *et al.*, Titans of the early universe: The Prato statement on the origin of the first supermassive black holes, Publ. Astron. Soc. Aust. **36**, e027 (2019).

[11] K. Inayoshi, E. Visbal, and Z. Haiman, The assembly of the first massive black holes, Annu. Rev. Astron. Astrophys. **58**, 27 (2020).

[12] M. Volonteri, M. Habouzit, and M. Colpi, The origins of massive black holes, Nat. Rev. Phys. **3**, 732 (2021).

[13] V. Cammelli, P. Monaco, J. C. Tan, J. Singh, F. Fontanot, G. De Lucia *et al.*, The formation of supermassive black holes from Population III.1 seeds. III. Galaxy evolution and black hole growth from semi-analytic modelling, Mon. Not. R. Astron. Soc. **536**, 851 (2025).

[14] V. Bromm and A. Loeb, Formation of the first supermassive black holes, Astrophys. J. **596**, 34 (2003).

[15] G. Lodato and P. Natarajan, Supermassive black hole formation during the assembly of pre-galactic discs, Mon. Not. R. Astron. Soc. **371**, 1813 (2006).

[16] X. Ding, M. Onoue, J. D. Silverman, Y. Matsuoka, T. Izumi, M. A. Strauss *et al.*, Detection of stellar light from quasar host galaxies at redshifts above 6, Nature (London) **621**, 51 (2023).

[17] H. Übler, R. Maiolino, E. Curtis-Lake, P. G. Pérez-González, M. Curti, M. Perna *et al.*, GA-NIFS: A massive black hole in a low-metallicity AGN at z ∼ 5.55 revealed by JWST/NIRSpec IFS, Astron. Astrophys. **677**, A145 (2023).

[18] R. L. Larson, S. L. Finkelstein, D. D. Kocevski, T. A. Hutchison, J. R. Trump, P. Arrabal Haro *et al.*, A CEERS discovery of an accreting supermassive black hole 570 myr after the big bang: Identifying a progenitor of massive $z > 6$ quasars, Astrophys. J. Lett. **953**, L29 (2023).

[19] Y. Harikane, Y. Zhang, K. Nakajima, M. Ouchi, Y. Isobe, Y. Ono, S. Hatano, Y. Xu, and H. Umeda, A JWST/NIRSpec first census of broad-line AGNs at $z = 4 - 7$: Detection of 10 faint AGNs with $M_{BH}10^6 - 10^8$ $M_\odot$ and their host galaxy properties, Astrophys. J. **959**, 39 (2023).

[20] A. Bogdan *et al.*, Evidence for heavy-seed origin of early supermassive black holes from a $z \approx 10$ x-ray quasar, Nat. Astron. **8**, 126 (2024).

[21] R. Maiolino *et al.*, JADES—The diverse population of infant black holes at $4 < z < 11$: Merging, tiny, poor, but mighty, Astron. Astrophys. **691**, A145 (2024).

[22] M. Yue, A.-C. Eilers, R. A. Simcoe, R. Mackenzie, J. Matthee, D. Kashino, R. Bordoloi, S. J. Lilly, and R. P. Naidu, EIGER. V. Characterizing the host galaxies of luminous quasars at $z \gtrsim 6$, Astrophys. J. **966**, 176 (2024).

[23] D. D. Kocevski, M. Onoue, K. Inayoshi, J. R. Trump, P. Arrabal Haro, A. Grazian *et al.*, Hidden little monsters: Spectroscopic identification of low-mass, broad-line AGNs at $z > 5$ with CEERS, Astrophys. J. Lett. **954**, L4 (2023).

[24] M. A. Stone, J. Lyu, G. H. Rieke, S. Alberts, and K. N. Hainline, Undermassive host galaxies of five z ∼ 6 luminous quasars detected with JWST, Astrophys. J. **964**, 90 (2024).

[25] F. Pacucci, B. Nguyen, S. Carniani, R. Maiolino, and X. Fan, JWST CEERS and JADES active galaxies at $z = 4 - 7$ violate the local $M_{\rm BH} - M_\star$ relation at $> 3\sigma$: Implications for low-mass black holes and seeding models, Astrophys. J. Lett. **957**, L3 (2023).

[26] J. Ellis, M. Fairbairn, G. Hütsi, J. Urrutia, V. Vaskonen, and H. Veermäe, Consistency of JWST black hole observations with NANOGrav gravitational wave measurements, Astron. Astrophys. **691**, A270 (2024).

[27] J. Li, J. D. Silverman, Y. Shen, M. Volonteri, K. Jahnke, M.-Y. Zhuang *et al.*, Tip of the iceberg: Overmassive black holes at $4 < z < 7$ found by JWST are not inconsistent with the local relation, Astrophys. J. **981**, 19 (2025).

[28] P. Dayal and R. Maiolino, The properties of primordially-seeded black holes and their hosts in the first billion years: Implications for JWST, arXiv:2506.08116.

[29] P. Natarajan, F. Pacucci, A. Ricarte, A. Bogdan, A. D. Goulding, and N. Cappelluti, First detection of an over-massive black hole galaxy UHZ1: Evidence for heavy black hole seed formation from direct collapse, Astrophys. J. Lett. **960**, L1 (2024).

[30] A. Toubiana, L. Sberna, M. Volonteri, E. Barausse, S. Babak, R. Enficiaud, D. Izquierdo–Villalba, J. R. Gair, J. E. Greene, and H. Q. Leclere, Reconciling PTA and JWST and preparing for LISA with POMPOCO: A parametrisation of the massive black hole population for comparison to observations, Astron. Astrophys. **700**, A135 (2025).

[31] D. D. Kocevski, S. L. Finkelstein, G. Barro, A. J. Taylor, A. Calabrò, B. Laloux *et al.*, The rise of faint, red active galactic nuclei at $z > 4$: A sample of little red dots in the JWST extragalactic legacy fields, Astrophys. J. **986**, 126 (2025).

[32] J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen, The origin of the JWST supermassive black holes, Astron. Astrophys. **703**, A138 (2025).

[33] A. K. Bhowmick, L. Blecha, P. Torrey, L. Z. Kelley, R. Weinberger, M. Vogelsberger, L. Hernquist, R. S. Somerville, and A. E. Evans, Introducing the BRAHMA simulation suite: Signatures of low-mass black hole seeding models in cosmological simulations, Mon. Not. R. Astron. Soc. **531**, 4311 (2024).
[34] K. Kritos, R. S. Beckmann, J. Silk, E. Berti, S. Yi, M. Volonteri, Y. Dubois, and J. Devriendt, Supermassive black hole growth in hierarchically merging nuclear star clusters, Astrophys. J. **991**, 58 (2025).
[35] O. Piana, H.-Y. Pu, and K. Wu, Super-Eddington accretion in high-redshift black holes and the emergence of jetted AGN, Mon. Not. R. Astron. Soc. **530**, 1732 (2024).
[36] A. Lupi, A. Trinca, M. Volonteri, M. Dotti, and C. Mazzucchelli, Size matters: Are we witnessing super-Eddington accretion in high-redshift black holes from JWST?, Astron. Astrophys. **689**, A128 (2024).
[37] E. Lambrides *et al.*, The case for super-Eddington accretion: Connecting weak x-ray and UV line emission in JWST broad-line AGN during the first Gyr of cosmic time, arXiv:2409.13047.
[38] P. Madau, Chasing the light: Shadowing, collimation, and the super-Eddington growth of infant black holes in JWST-discovered AGNs, arXiv:2501.09854.
[39] T. Zana, P. R. Capelo, M. Boresta, R. Schneider, A. Lupi, A. Trinca *et al.*, Super-Eddington accretion in protogalactic cores, arXiv:2508.21114.
[40] J. Yang, F. Wang, X. Fan, J. F. Hennawi, A. J. Barth, E. Bañados *et al.*, A spectroscopic survey of biased halos in the reionization era (ASPIRE): A first look at the rest-frame optical spectra of $z > 6.5$ quasars using JWST, Astrophys. J. Lett. **951**, L5 (2023).
[41] A. J. Taylor, S. L. Finkelstein, D. D. Kocevski, J. Jeon, V. Bromm, R. O. Amorín *et al.*, Broad-line AGNs at $3.5 < z < 6$: The black hole mass function and a connection with little red dots, Astrophys. J. **986**, 165 (2025).
[42] J. E. Greene, I. Labbe, A. D. Goulding, L. J. Furtak, I. Chemerynska, V. Kokorev *et al.*, UNCOVER spectroscopy confirms the surprising ubiquity of active galactic nuclei in red sources at $z > 5$, Astrophys. J. **964**, 39 (2024).
[43] J. Matthee *et al.*, Little red dots: An abundant population of faint active galactic nuclei at $z \sim 5$ revealed by the EIGER and FRESCO JWST surveys, Astrophys. J. **963**, 129 (2024).
[44] X. Shen, M. Vogelsberger, M. Boylan-Kolchin, S. Tacchella, and R. P. Naidu, Early galaxies and early dark energy: A unified solution to the Hubble tension and puzzles of massive bright galaxies revealed by JWST, Mon. Not. R. Astron. Soc. **533**, 3923 (2024).
[45] F. Gentile *et al.*, Not-so-little red dots: Two massive and dusty starbursts at $z \sim 5$–7 pushing the limits of star formation discovered by JWST in the COSMOS-Web survey, Astrophys. J. Lett. **973**, L2 (2024).
[46] M. Kokubo and Y. Harikane, Challenging the active galactic nucleus scenario for JWST/NIRSpec little red dot and non–little red dot broad Hα emitters in light of nondetection of NIRCam photometric variability and x-ray, Astrophys. J. **995**, 24 (2025).
[47] T. T. Ananna, Á. Bogdán, O. E. Kovács, P. Natarajan, and R. C. Hickox, X-ray view of little red dots: Do they host supermassive black holes?, Astrophys. J. Lett. **969**, L18 (2024).
[48] P. Madau and F. Haardt, X-ray weak active galactic nuclei from super-Eddington accretion onto infant black holes, Astrophys. J. Lett. **976**, L24 (2024).
[49] F. Pacucci and R. Narayan, Mildly super-Eddington accretion onto slowly spinning black holes explains the x-ray weakness of the little red dots, Astrophys. J. **976**, 96 (2024).
[50] R. P. Naidu *et al.*, A "black hole star" reveals the remarkable gas-enshrouded hearts of the little red dots, arXiv:2503.16596.
[51] I. Juodžbalis, C. Marconcini, F. D'Eugenio, R. Maiolino, A. Marconi, H. Übler *et al.*, A direct black hole mass measurement in a little red dot at the epoch of reionization, arXiv:2508.21748.
[52] J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen, Starlight from JWST: Implications for star formation and dark matter models, Astron. Astrophys. **702**, A109 (2025).
[53] LISA Collaboration, LISA definition study report, arXiv:2402.07571.
[54] AION Collaboration, A prototype atom interferometer to detect dark matter and gravitational waves, arXiv:2504.09158.
[55] P. Dayal, E. M. Rossi, B. Shiralilou, O. Piana, T. R. Choudhury, and M. Volonteri, The hierarchical assembly of galaxies and black holes in the first billion years: Predictions for the era of gravitational wave astronomy, Mon. Not. R. Astron. Soc. **486**, 2336 (2019).
[56] See Supplemental Material at http://link.aps.org/supplemental/10.1103/njd2-3vdl for [brief description].
[57] Our implementations of the halo mass function cutoffs for FDM and WDM are described in detail in Sec. 3.1 of [52]. The implementation for FDM is based on [58] and [59], whose fitting formula we use. The implementation for WDM is based on [60–62]. The DM properties can also affect the formation of seeds. In particular, direct collapse SMBH seeds can form in WDM models at lower redshifts than in the CDM model [63]. We do not account for this effect but use a generic parametrization of the seeds.
[58] W. Hu, R. Barkana, and A. Gruzinov, Cold and fuzzy dark matter, Phys. Rev. Lett. **85**, 1158 (2000).
[59] S. Passaglia and W. Hu, Accurate effective fluid approximation for ultralight axions, Phys. Rev. D **105**, 123529 (2022).
[60] P. Bode, J. P. Ostriker, and N. Turok, Halo formation in warm dark matter models, Astrophys. J. **556**, 93 (2001).
[61] S. H. Hansen, J. Lesgourgues, S. Pastor, and J. Silk, Constraining the window on sterile neutrinos as warm dark matter, Mon. Not. R. Astron. Soc. **333**, 544 (2002).
[62] M. Viel, J. Lesgourgues, M. G. Haehnelt, S. Matarrese, and A. Riotto, Constraining warm dark matter candidates including sterile neutrinos and light gravitinos with WMAP and the Lyman-alpha forest, Phys. Rev. D **71**, 063534 (2005).
[63] P. Dayal, T. R. Choudhury, F. Pacucci, and V. Bromm, Warm dark matter constraints from high-z direct collapse black holes using the JWST, Mon. Not. R. Astron. Soc. **472**, 4414 (2017).
[64] A. K. Bhowmick, L. Blecha, L. Z. Kelley, A. Sivasankaran, P. Torrey, R. Weinberger *et al.*, Dynamics of low-mass black

hole seeds in the BRAHMA simulations using subgrid-dynamical friction: Impact on merger-driven black hole growth in the high redshift Universe, Astrophys. J. **991**, 81 (2025).

[65] We treat the mergers as instantaneous. In a more realistic model, this process will take time. We plan to explore this further in future work.

[66] A. E. Reines and M. Volonteri, Relations between central black hole mass and total galaxy stellar mass in the local Universe, Astrophys. J. **813**, 82 (2015).

[67] T. Izumi, Y. Matsuoka, S. Fujimoto, M. Onoue, M. A. Strauss, H. Umehata *et al.*, Subaru high-z exploration of low-luminosity quasars (SHELLQs). XIII. Large-scale feedback and star formation in a low-luminosity quasar at $z = 7.07$ on the local black hole to host mass relation, Astrophys. J. **914**, 36 (2021).

[68] A. Gelman and D. B. Rubin, Inference from iterative simulation using multiple sequences, Stat. Sci. **7**, 457 (1992).

[69] J. Jeon, B. Liu, A. J. Taylor, V. Kokorev, J. Chisholm, D. D. Kocevski, S. L. Finkelstein, and V. Bromm, The emerging black hole mass function in the high-redshift universe, Astrophys. J. **988**, 110 (2025).

[70] The results in this Letter are rather insensitive to the star formation rate at high z and, in particular, would not depend strongly on any change in Pop-III production.

[71] S. Naoz and R. Barkana, The formation and gas content of high-redshift galaxies and minihaloes, Mon. Not. R. Astron. Soc. **377**, 667 (2007).

[72] A. T. P. Schauer, S. C. O. Glover, R. S. Klessen, and P. Clark, The influence of streaming velocities and Lyman-Werner radiation on the formation of the first stars, Mon. Not. R. Astron. Soc. **507**, 1775 (2021).

[73] M. Kulkarni, E. Visbal, and G. L. Bryan, The critical dark matter halo mass for Population III star formation: Dependence on Lyman-Werner radiation, baryon-dark matter streaming velocity, and redshift, Astrophys. J. **917**, 40 (2021).

[74] J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe, Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data, Phys. Rev. D **109**, L021302 (2024).

[75] R. G. Bower, J. Schaye, C. S. Frenk, T. Theuns, M. Schaller, R. A. Crain, and S. McAlpine, The dark nemesis of galaxy formation: Why hot haloes trigger black hole growth and bring star formation to an end, Mon. Not. R. Astron. Soc. **465**, 32 (2017).

[76] H. H. S. Chiu, H.-Y. Schive, H.-Y. K. Yang, H. Huang, and M. Gaspari, Boosting supermassive black hole growth in the early universe by fuzzy dark matter solitons, Phys. Rev. Lett. **134**, 051402 (2025).

[77] V. Iršič, M. Viel, M. G. Haehnelt, J. S. Bolton, and G. D. Becker, First constraints on fuzzy dark matter from Lyman-α forest data and hydrodynamical simulations, Phys. Rev. Lett. **119**, 031302 (2017).

[78] W. Enzi *et al.*, Joint constraints on thermal relic dark matter from strong gravitational lensing, the Lyα forest, and Milky Way satellites, Mon. Not. R. Astron. Soc. **506**, 5848 (2021).

[79] K. K. Rogers and H. V. Peiris, Strong bound on canonical ultralight axion dark matter from the Lyman-alpha forest, Phys. Rev. Lett. **126**, 071302 (2021).

[80] D. J. E. Marsh and J. C. Niemeyer, Strong constraints on fuzzy dark matter from ultrafaint dwarf galaxy Eridanus II, Phys. Rev. Lett. **123**, 051103 (2019).

[81] N. Dalal and A. Kravtsov, Excluding fuzzy dark matter with sizes and stellar kinematics of ultrafaint dwarf galaxies, Phys. Rev. D **106**, 063517 (2022).

[82] M. Benito, G. Hütsi, K. Müürsepp, J. S. Almeida, J. Urrutia, V. Vaskonen, and H. Veermäe, Fuzzy dark matter fails to explain dark matter cores, Phys. Dark Universe **49**, 102010 (2025).

[83] D. M. Powell, S. Vegetti, J. P. McKean, S. D. M. White, E. G. M. Ferreira, S. May, and C. Spingola, A lensed radio jet at milli-arcsecond resolution—II. Constraints on fuzzy dark matter from an extended gravitational arc, Mon. Not. R. Astron. Soc. **524**, L84 (2023).

[84] A. Amruth *et al.*, Einstein rings modulated by wavelike dark matter from anomalies in gravitationally lensed images, Nat. Astron. **7**, 736 (2023).

[85] K. K. Rogers and V. Poulin, 5σ tension between Planck cosmic microwave background and eBOSS Lyman-alpha forest and constraints on physics beyond ΛCDM, Phys. Rev. Res. **7**, L012018 (2025).

[86] V. Iršič *et al.*, New constraints on the free-streaming of warm dark matter from intermediate and small scale Lyman-α forest data, Phys. Rev. D **96**, 023522 (2017).

[87] S. Vegetti, G. Despali, M. R. Lovell, and W. Enzi, Constraining sterile neutrino cosmologies with strong gravitational lensing observations at redshift $z \sim 0.2$, Mon. Not. R. Astron. Soc. **481**, 3661 (2018).

[88] J.-W. Hsueh, W. Enzi, S. Vegetti, M. Auger, C. D. Fassnacht, G. Despali, L. V. E. Koopmans, and J. P. McKean, SHARP—VII. New constraints on the dark matter free-streaming properties and substructure abundance from gravitationally lensed quasars, Mon. Not. R. Astron. Soc. **492**, 3047 (2020).

[89] D. Gilman, S. Birrer, A. Nierenberg, T. Treu, X. Du, and A. Benson, Warm dark matter chills out: Constraints on the halo mass function and the free-streaming length of dark matter with eight quadruple-image strong gravitational lenses, Mon. Not. R. Astron. Soc. **491**, 6077 (2020).

[90] A. Dekker, S. Ando, C. A. Correa, and K. C. Y. Ng, Warm dark matter constraints using Milky Way satellite observations and subhalo evolution modeling, Phys. Rev. D **106**, 123026 (2022).

[91] B. Liu, H. Shan, and J. Zhang, New galaxy UV luminosity constraints on warm dark matter from JWST, Astrophys. J. **968**, 79 (2024).

Curriculum Vitae

1. Personal data

Name	Juan Urrutia
Date and place of birth	1 April 1998, Barcelona, Spain
Nationality	Spanish

2. Contact information

Address	Estonia pst, 7 53, 19086 Tallinn, Estonia
Phone	+34 637918774
E-mail	juan.urrutia@kbfi.ee

3. Education

2022–2026	PhD studies, TalTech
2020–2021	High Energy Physics Master's Degree, UAB
2016–2020	Physics undergraduate, Autonomous University of Barcelona (UAB)

4. Language competence

Spanish	native
Catalan	native
English	fluent

5. Professional employment

2022– ...	Junior Researcher, KBFI

6. Computer skills

- Operating systems: Windows, Macintosh
- Programming languages: Python, Mathematica
- Scientific packages: Numpy, Scipy, Matplotlib

7. Honours and awards

- 2023, Fourth prize in the Gravity Research Essay competition
- 2025, Humboldt Postdoc Fellowship

8. Defended theses

- 2021, Probes of compact Dark Matter with Gravitational Wave Lensing, MSc, supervisor Dr. Ville Vaskonen, UAB

9. Fields of research

- Physics phenomenology, cosmology, high energy physics, astrophysics

10. Scientific work
Papers

1. J. Urrutia and V. Vaskonen. Lensing of gravitational waves as a probe of compact dark matter. *Mon. Not. Roy. Astron. Soc.*, 509(1):1358–1365, 2021
2. M. Fairbairn, J. Urrutia, and V. Vaskonen. Microlensing of gravitational waves by dark matter structures. *JCAP*, 07:007, 2023
3. G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Did JWST observe imprints of axion miniclusters or primordial black holes? *Phys. Rev. D*, 107(4):043502, 2023
4. J. Ellis, M. Fairbairn, G. Hütsi, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Prospects for future binary black hole gravitational wave studies in light of PTA measurements. *Astron. Astrophys.*, 676:A38, 2023
5. J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational wave microlensing by dressed primordial black holes. *Phys. Rev. D*, 108(2):023507, 2023
6. J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data. *Phys. Rev. D*, 109(2):L021302, 2024
7. J. Ellis, M. Fairbairn, G. Franciolini, G. Hütsi, A. Iovino, M. Lewicki, M. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. What is the source of the PTA GW signal? *Phys. Rev. D*, 109(2):023522, 2024
8. J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Probing Supermassive Black Hole Seed Scenarios with Gravitational-wave Measurements. *Astrophys. J.*, 964(1):11, 2024
9. J. Urrutia and V. Vaskonen. Dark timbre of gravitational waves. *Phys. Rev. D*, 111(12):123047, 2025
10. J. Urrutia. Using gravitational waves to detect axion miniclusters and axion stars. *PoS*, COSMICWISPers:046, 2024
11. J. Ellis, M. Fairbairn, G. Hütsi, J. Urrutia, V. Vaskonen, and H. Veermäe. Consistency of JWST black hole observations with NANOGrav gravitational wave measurements. *Astron. Astrophys.*, 691:A270, 2024

12. J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. In search of the biggest bangs since the big bang. *Int. J. Mod. Phys. D*, 33(15):2440004, 2024

13. J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Eccentricity effects on the supermassive black hole gravitational wave background. *Astron. Astrophys.*, 691:A212, 2024

14. J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. The origin of the JWST supermassive black holes. *Astron. Astrophys.*, 703:A138, 2025

15. J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe. Statistics of supermassive black hole gravitational wave background anisotropy. *Astron. Astrophys.*, 706:A159, 2026

16. A. Abdalla et al. Terrestrial Very-Long-Baseline Atom Interferometry: summary of the second workshop. *EPJ Quant. Technol.*, 12(1):42, 2025

17. M. Benito, G. Hütsi, K. Müürsepp, J. S. Almeida, J. Urrutia, V. Vaskonen, and H. Veermäe. Fuzzy dark matter fails to explain dark matter cores. *Phys. Dark Univ.*, 49:102010, 2025

18. J. Urrutia, J. Ellis, M. Fairbairn, and V. Vaskonen. Starlight from JWST: Implications for star formation and dark matter models. *Astron. Astrophys.*, 702:A109, 2025

19. I. D. Gialamas, G. Hütsi, M. Raidal, J. Urrutia, M. Vasar, and H. Veermäe. Quintessence and phantoms in light of DESI 2025. *Phys. Rev. D*, 112(6):063551, 2025

20. J. Ellis, M. Fairbairn, J. Urrutia, and V. Vaskonen. Constraints on Dark Matter Models from Supermassive Black Hole Evolution. *Phys. Rev. Lett.*, 136(13):131001, 2026

Conference presentations

1. May 2022: IBERICOS, Barcelona — Gravitational wave lensing by compact dark matter

2. May 2022: EUCAPT Second Annual Symposium, CERN — Probing low-mass halos with gravitational wave lensing

3. June 2022: EUCAPT Workshop on Gravitational Waves Probes of Black Hole Environments, Rome — Probing low-mass halos with gravitational wave lensing

4. February 2023: COST Action Kick-Off Meeting “Cosmic Whispers”, Frascati — How to probe axion mini clusters with gravitational waves

5. April 2023: IBERICOS, Porto — What are the implications for LISA of the NANOGrav 12.5 yr data in terms of supermassive black hole binaries

6. May 2023: Planck Conference, Warsaw — Lensing by dressed primordial black holes

7. April 2024: Terrestrial Very-Long-Baseline Atom Interferometer — GWs with atom interferometers

8. April 2024: Massive Black Holes in the First Billion Years, York — How to tell different seeding scenarios by recovering the occupation fraction using LISA and proposed atom interferometers

9. June 2024: LISA Cosmology Meeting, Porto — About the dark timbre of gravitational waves

10. October 2024: LISA Spanish Meeting — The importance of light Dark Matter halos in LISA
11. May 2025: 5th EUCAPT Annual Symposium — Supermassive Black Holes in light of the PTA background
12. June 2025: CosmoFoundue, Geneva — Deviations from CDM in SMBH, bounds from JWST and predictions for LISA
13. August 2025: GR Amaldi, Glasgow — Deviations from CDM in SMBH, bounds from JWST and predictions for LISA
14. May 2026: EUCAPT annual meeting, CERN — Contributed Talk: Constraints on Dark Matter from the Evolution of Supermassive Black Holes
15. June 2026: LISA Cosmology meeting, Barcelona — Plenary Talk: Wave Optics effects in GWs
16. July 2026: ATHEXIS, Athens — Contributed Talk: Wave Optics effects in GWs

Elulookirjeldus

1. Isikuandmed

Nimi	Juan Urrutia
Sünniaeg ja -koht	1. aprill 1998, Barcelona, Hispaania
Kodakondsus	Hispaania

2. Kontaktandmed

Aadress	Estonia pst, 7 53, 19086 Tallinn, Eesti
Telefon	+34 637918774
E-post	juan.urrutia@kbfi.ee

3. Haridus

2022–2026	PhD füüsikas, TalTech
2020–2021	MSc kõrge energia füüsikas, UAB
2016–2020	BSc füüsikas, UAB

4. Keeleoskus

Hispaania	emakeel
Katalaani	emakeel
Inglise	väga hea

5. Töökogemus

2022– ...	Nooremteadur, KBFI

6. Arvutioskused

- Operatsioonisüsteemid: Windows, Macintosh
- Programmeerimiskeeled: Python, Mathematica
- Teaduspaketid: Numpy, scipy, matplotlib

7. Tunnustused ja auhinnad

- 2023, Neljas koht Gravity Research Essay konkursil
- 2025, Humboldti järeldoktorantuuri stipendium

8. Kaitstud lõputööd

- 2021, Probes of Compact Dark Matter with Gravitational Wave Lensing, magistritöö, juhendaja prof. Ville Vaskonen, UAB

9. Uurimisvaldkonnad

- Füüsika fenomenoloogia, kosmoloogia, kõrge energia füüsika, astrofüüsika

10. Teadustegevus

- Teaduspublikatsioonide ja esinemiste nimekiri on ära toodud ingliskeelse CV juures.

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-564-0 (PDF)