A New Estimate of the Galaxy Luminosity Function, Using Machine Learning and a Mock Catalogue

SUTTIKOON KOONKOR

A New Estimate of the Galaxy Luminosity Function, Using Machine Learning and a Mock Catalogue

Suttikoon Koonkor

A thesis submitted to Durham University
in accordance with the regulations for
admittance to PhD. Astrophysics

Institute for Computational Cosmology
Durham University
United Kingdom

September 2025

A New Estimate of the Galaxy Luminosity Function, Using Machine Learning and a Mock Catalogue

Suttikoon Koonkor

Abstract

A new measurement of the galaxy luminosity function (LF) is presented over the redshift range $0.05 < z < 2.0$ using data from the Physics of the Accelerating Universe Survey (PAUS). Leveraging the high photometric redshift precision ($\sigma_z/(1+z) \sim 0.0035$) made possible by PAUS's 40 narrow-band optical filters, rest-frame magnitudes are derived and the LF is estimated in multiple redshift bins using the $1/V_{\rm max}$ method. The analysis is supported by a realistic mock catalogue constructed from the `GALFORM` semi-analytic galaxy formation model. To compute rest-frame magnitudes for observed galaxies, a Random Forest regression model was trained to predict k-corrections from observable properties. The mock was used to investigate the impact of photometric uncertainties and redshift errors. This revealed that these observational effects significantly modify the shape of the LF–especially at the bright end. A detailed error analysis revealed that photometric and photometric redshift errors dominate the LF uncertainty, contributing errors nearly an order of magnitude larger than those from large-scale structure across all redshift bins. Importantly, the observed faint-end turnover in the LF–driven by selection in the observed i-band–can still be used to constrain galaxy formation models when the same selection is applied to the simulated data. The resulting luminosity functions in the i-band and other rest-frame bands (u, g, r, z) show good agreement with theoretical predictions from `GALFORM`. The LF is also measured separately for red and blue galaxies, revealing distinct evolutionary trends. Symbolic regression models developed by collaborators are used to estimate stellar masses and measured the stellar mass function from both PAUS and the mock sample. This thesis highlights the value of combining narrow-band photometric surveys and machine learning methods to probe galaxy evolution

with precision, and demonstrates the usefulness of mocks to model selection effects.

Supervisors: Prof. Carlton Baugh and Dr. Peder Norberg

Institute for Computational Cosmology

Department of Physics

Durham University

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my primary supervisor, Prof. Carlton Baugh, for his unwavering support, patient guidance, and encouragement throughout my PhD. His insight, scientific expertise, and thoughtful feedback have shaped every part of this thesis. I'm especially grateful for the outreach opportunities he encouraged me to take part in during my time at Durham– they reminded me why I fell in love with astrophysics in the first place. (And don't worry, I'm keeping a hopeful eye on Bolton Wanderers' promotion chances–one can dream!)

I am also sincerely thankful to Prof. Peder Norberg, my second supervisor, for this valuable advice, especially in the early stages of my research, and for always offering a clear and critical perspective.

My thanks go as well to Dr. Giorgio Manzoni for his guidance and support throughout my work with both the `GALFORM` lightcone mock and PAUS data. His expertise and willingness to help at every stage made a significant different to the progress of this thesis.

I gratefully acknowledge the Office of the Civil Service Commission (OCSC), Thailand, for their financial support throughout the duration of my PhD. This opportunity would not have been possible without their sponsorship, and I am honoured to have been a recipient of their funding.

To the PAUS collaboration, thank you for the opportunity to work with such a rich dataset, and for your guidance and support throughout. It has been a privilege to contribute to the PAUS science goals.

I am also grateful to the Institute for Computational Cosmology (ICC), Durham University, for access to the COSMA supercomputer, and to the incredible COSMA support team who keep everything running smoothly behind the scenes. Much of

the analysis presented in this thesis would not have been possible without your efforts.

A special thank you goes to Shufei, who brought energy and kindness into the department. Your cheerful messages and check-ins made a big difference–thank you for always being a supportive presence. I hope Joshua and Lily do well.

To my family–my mum, dad, and sister–thank you for your unconditional love and belief in me. Even though we were separated by continents, time zones, and oceans, I never once felt alone during my time in the UK. Not a single day passed without your words of encouragement. You reminded me to believe in myself, and I carried that with me through every step.

And now, on to the wonderful chaos crew.

To my incredible friends in the department–Makun, Zoe, Emmy, Sarah, Marcus, Dom, Ciera and Tim (yes, you're definitely one of us)–thank you for keeping me sane. From spontaneous coffee walks, Halloween party's COSMA machine costumes, summer BBQs, and that one unforgettable escape to Holy Island. I couldn't have asked for a better group of friends to share this journey with–especially to make Durham feel like my second Home. Your kindness, humour, and patience to put up with my terrible puns made the toughest days brighter and the good days even better. These memories will stay with me for a very, very long time. Oh yes, thank you for actually not making the "Gym Jar" idea happen!

To my amazing Ustinov College football teammates–Dom (yes, you again), Jay, Evan, Rafe, Ellis, and Harrison–thank you for letting me be part of a squad full of passion, grit, and just the right amount of chaos on and off the pitch. Whether we were winning 18-2, losing 0-9 with formation experiment, or turning up to the pitch because we got the date wrong, it was always a good time. Jay's top-bin screamer from 40 yards out and Ellis' bicycle kick are still burned into my memory like Premier League highlights. Wishing you all the very best in the coming seasons–keep the dream alive, and keep those screamers coming!

Finally, to thank my friends at work during my part-time job–Roman, Nate, Tommy, Art, and Gee–thank you for fuelling me (sometimes literally) through long days during busy shifts and deadlines from the university. Your support and flexibility gave me the space and peace of mind I needed to finish this. And your food? Absolute lifesaver.

Contents

Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, Department of Physics, University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is the sole work of the author unless referenced to the contrary in the text.

For this thesis we make use of the `GALFORM` lightcone mock catalogue, which is the work of Manzoni et al. (2024), for theoretical dataset and the Physics of the Accelerating Universe Survey (PAUS), which is the work of the PAUS collaboration (Eriksen et al., 2019; Padilla et al., 2019; Navarro-Gironés et al., 2024), for the observational dataset. We also make use of a stellar mass estimation formula, which is the work of Kumar et al. (in prep), of which I am a co-author.

The main results of this thesis, as shown in Chapter 6, will be submitted shortly for publication as a first-author paper (Koonkor et al.).

List of Figures

List of Tables

List of Acronyms

AGN Active Galactic Nuclei

BB Broad Band

CCD Charge-Coupled Device

CDF Cumulative Distribution Function

CDM Cold Dark Matter

CFHTLS Canada-France-Hawaii Telescope Legacy Survey

CMB Cosmic Microwave Background

COSMOS Cosmic Evolution Survey

DESI Dark Energy Spectroscopic Instrument

FWHM Full Width at Half Maximum

GAMA Galaxy And Mass Assembly

IMF Initial Mass Function

IGM Intergalatic Medium

ISM Interstellar Medium

KiDS Kilo-Degree Survey

LSS Large Scale Structure

LF Luminosity Function

MC Monte Carlo

ML Machine Learning

MSE Mean Squared Error

NB Narrow Band

PAUS Physics of the Accelerating Universe Survey

PMILL Planck Millennium

RFR Random Forest Regression

SED Spectral Energy Distribution

SF Star Formation

SFH Star Formation History

SFR Star Formation Rate

SPS Stellar Population Synthesis

SDSS Sloan Digital Sky Survey

SMF Stellar Mass Function

SMBH Supermassive Black Hole

SN Supernova

SSP Simple Stellar Population

UV Ultraviolet

WHT William Herschel Telescope

Chapter 1

Introduction

The luminosity function (Luminosity Function (LF)) is one of the most fundamental statistical tools in extragalactic astrophysics. It quantifies the number density of galaxies as a function of luminosity and provides insight into the underlying physical processes shaping the galaxy population, such as star formation, feedback mechanisms, and gas accretion. In theoretical models of galaxy formation, the LF serves as a critical constraint, sensitive to the efficiency of gas cooling, Supernova (SN) feedback, and the role of Active Galactic Nuclei (AGN) in suppressing star formation (e.g. Efstathiou (2000); Cole et al. (2000); Benson et al. (2003).

Early estimates of the galaxy luminosity function were limited by sample size and redshift coverage. A major breakthrough came from Loveday et al. (1992), who used the Stromlo-APM Redshift Survey. This marked a step forward in deriving a well-constrained local LF from over 1,700 galaxies that sparsely sampled over $\sim 4,300$ deg^2, and demonstrated the Schechter function's effectiveness in describing the LF across a wide luminosity range. Other important advances include the analysis by Efstathiou et al. (1988), who used the CfA Redshift Survey to refine statistical methods for LF estimation and highlighted the impact of large-scale structure on LF measurements, and Marzke et al. (1994), who further explored the CfA survey data to derive morphological-type-dependent LFs, revealing significant differences between early- and late-type galaxies. The pioneering study by Lilly

et al. (1995) then pushed LF studies to $z \sim 1$ using deep imaging and spectroscopy over five small fields (each covering a $10' \times 10'$ field) to estimate LF evolution, based on 730 galaxies. Around the same time, Zucca et al. (1997) used the ESO Slice Project (ESP) to determine a more precise local LF using over 3,000 spectroscopic redshifts. Other notable contributions include Folkes et al. (1999), who used the 2dF Galaxy Redshift Survey to explore the LF by spectral type, but over a limited range of redshift. These studies revealed both the shape and evolution of the LF, but were constrained by small volumes and limited sampling of large-scale structure.

A series of photometric and spectroscopic surveys have built upon the early work by Lilly et al. The Classifying Objects by Medium-Band Observation in 17 Filters (COMBO-17 Survey: Wolf et al. 2003) measured the luminosity function from a sample of $\sim 25,000$ galaxies using photometric redshifts derived from 12 medium-band filters (with a Full Width at Half Maximum (FWHM) varying from 140 to 310 Å and $UBVRI$ broad band filters covering a total area of 0.78 deg^2. The photometric redshift error was $\sigma_{68}/(1+z) = 0.03$ (about a hundred times larger than the error on a spectroscopic redshift). Similarly, the Advanced Large Homogeneous Area Medium-Band Redshift Astronomical (ALHAMBRA survey: Moles et al. 2008) measured $\sim 600\,000$ galaxy redshifts using 20 medium-band optical filters (with a FWHM of 310 Å) over a total solid angle of 4 deg^2. Despite the improvement in the luminosity function estimates from these larger samples which results in a reduced sensitivity to the effect of large scale structure, the photometric redshifts are still substantially less accurate than typical spectroscopic estimates. This can lead to the evolution in the measured luminosity function being misestimated, when photometric redshift outliers are assigned to the wrong redshift bin in the analysis. Large spectroscopic surveys have provided high-precision measurements of the local luminosity function, including the 2-degree galaxy redshift survey (Norberg et al., 2002) and the main Sloan Digital Sky Survey (SDSS) (Blanton et al., 2003), both with median redshifts around $z \sim 0.1$. The deeper Galaxy

And Mass Assembly (GAMA) survey measured the luminosity function in several broad band filters out to $z \sim 0.5$ (Loveday et al., 2012). The Dark Energy Spectroscopic Instrument (DESI) Bright Galaxy Survey reaches a similar depth to GAMA but over a much larger solid angle, pushing the luminosity function estimates out to $z \sim 0.6$ (Moore et al in prep). Other surveys have probed intermediate redshifts, but typically adopt a colour selection to target a particular redshift interval without covering $z = 0$ (e.g. Davidzon et al. 2013). Some spectroscopic surveys have sampled a wide baseline in redshift ($0 < z < 2$) but at the expense of covering a relatively small solid angle, of order one square degree (Ilbert et al., 2005).

The Physics of the Accelerating Universe Survey (PAUS; (Eriksen et al., 2019; Padilla et al., 2019) aims to bridge this gap. PAUS uses 40 narrow-band optical filters to achieve photometric redshift precision of $\sigma_{68}/(1+z) \sim 0.0035$, approaching the regime of low-resolution spectroscopy* (Navarro-Gironés et al. 2024). Covering over ~ 50 square degrees, PAUS provides an unprecedented opportunity to measure the LF with high redshift accuracy over a large cosmic volume and a wide baseline in redshift. In addition, it enables studies of galaxy properties–including rest-frame colours and stellar populations–with finer spectral resolution than traditional broad-band surveys.

Complementing observational advances, recent years have also seen a surge in Machine Learning (ML) applications across astronomy. ML models are now routinely used to estimate stellar masses (e.g. Chu et al. 2024), metallicity (e.g. Acquaviva 2016), classify galaxy morphologies (e.g. Domínguez Sánchez et al. 2018), and improve photo-z estimates (e.g. Bonfield et al. 2010; Pasquet et al. 2019; Daza-Perilla et al. 2025). In this thesis, we apply machine learning techniques not only to infer rest-frame galaxy properties but also to understand their implications for the observed galaxy luminosity and galaxy stellar mass functions.

This thesis presents a new measurement of the galaxy luminosity function using

*For reference, the target random measurement error in redshift for the Dark Energy Spectroscopic Instrument is $\sigma_{68}/(1+z) = 1.4 \times 10^{-4}$ (DESI Collaboration et al., 2024).

PAUS data, interpreted through the lens of the GALFORM semi-analytic galaxy formation model (Cole et al., 2000; Lacey et al., 2016). We exploit a realistic lightcone mock catalogue (Manzoni et al. 2024) to account for selection effects, photometric redshift uncertainties, and flux errors. In particular, we emphasise how incompleteness at the faint end–due to the selection band shifting with redshift–can still be informative when forward-modelling the same selection in the theoretical framework. This work builds upon decades of LF studies, offering a modern perspective using state-of-the-art observational and theoretical tools.

1.1 Thesis Outline

This thesis is structured as follows. Several chapters include material that was not part of my original research but are included for completeness, to provide context for the core results presented in this work. Where applicable, I clearly indicate which aspects of each chapter reflect my direct contributions.

- **Chapter 2** provides an overview of the Physics of the Accelerating Universe Survey (PAUS), including the survey design and its photometric redshift performance. While the PAUS photo-z calibration and analysis are not part of my original work, I independently measured the effective survey areas used for luminosity function estimation after applying different selection cuts.

- **Chapter 3** outlines the theoretical background of galaxy formation and the GALFORM semi-analytic model. This chapter is also included for completeness and was not part of my original research. It describes the key physical processes implemented in GALFORM, including gas accretion, star formation, feedback, mergers, and chemical enrichment.

- **Chapter 4** describes the construction of a mock galaxy catalogue using GALFORM outputs. Although I did not develop GALFORM or the lightcone itself, I applied photometric flux and photometric redshift uncertainties to the cata-

logue to match PAUS-like observational conditions. These perturbed mocks form the basis for my comparison with data. I have added some analysis of the mock that was not considered by Manzoni et al. (2024).

- **Chapter 5** is entirely my work and focuses on the use of machine learning to estimate rest-frame magnitudes and the background knowledge of the k-correction.

- **Chapter 6** contains the core results of this thesis. I present a new measurement of the galaxy luminosity function in the i-band and other rest-frame bands using PAUS observations. I quantify uncertainties from photometric errors, photometric redshift outliers, and sample incompleteness, and compare the measured LFs with `GALFORM` predictions. I also analyse the luminosity function by colour and redshift and explore its redshift evolution.

- **Chapter 7** revisits the application of machine learning-based stellar mass estimation using the perturbed catalogue. The machine learning model was developed by Adarsh Kumar. I use this model to predict stellar masses for the PAUS dataset and to measure the stellar mass function, which is compared to theoretical predictions.

- **Chapter 8** concludes the thesis with a summary of the main findings and their implications for galaxy formation models. I also discuss possible directions for future work.

Chapter 2

Observational Data: The Physics of the Accelerating Universe Survey (PAUS)

This chapter provides an overview of the observational dataset used throughout this thesis: the Physics of the Accelerating Universe Survey (PAUS). Designed to bridge the gap between traditional broadband photometry and spectroscopic surveys, PAUS offers a unique combination of wide-area coverage and low-resolution spectra through its use of 40 narrow-band (Narrow Band (NB)) filters. This capability enables high-precision photometric redshift measurements essential for studying galaxy evolution and large-scale structure.

We begin in § 2.1 by introducing PAUS and its scientific goals. In § 2.2, we describe the PAUCam instrument and its NB filter system. § 2.3 outlines the survey design, including field selection and overlap with ancillary datasets. The methodology for measuring photometric redshifts is presented in § 2.4, followed by an assessment of redshift performance and data quality in § 2.5. In § 2.6, we describe the Monte Carlo technique used to estimate the effective survey area and its generalisation to various selection cuts. Finally, we summarise the key aspects of the dataset in § 2.7.

2.1 Introduction to PAUS

The Physics of the Accelerating Universe Survey (PAUS; (Eriksen et al., 2019; Padilla et al., 2019)) is a photometric survey designed to tackle some of the most pressing questions in cosmology–most notably, the nature of dark energy. PAUS achieves this by providing highly precise photometric redshifts through an innovative combination of Narrow Band (NB) and Broad Band (BB) imaging. By employing a set of 40 NB filters spanning the optical wavelength from 450 nm to 850 nm, PAUS attains a redshift precision of approximately $\sigma_{68}/(1+z) = 0.0035$ for galaxies brighter than $i_{AB} = 22.5$ (Eriksen et al. 2019; Alarcon et al. 2021). This precision bridges the gap between traditional BB imaging–which offers high depth and wide area coverage but limited spectral resolution–and spectroscopy, which, despite its high accuracy, is far more observationally intensive. PAUS will be able to provide cosmological constraints in its own right, for example, by using combined lensing and clustering analyses (e.g. Eriksen and Gaztañaga 2015c). PAUS will also provide the best constraints to date on the intrinsic alignment of galaxies, which is a key systematic in weak gravitational lensing studies. Hence the PAUS survey has the potential to enhance the scientific value of data extracted from much larger lensing surveys, such as Euclid (Hoekstra et al. 2017) and Roman (Wenzl et al. 2022).

Formerly installed at the prime focus of the 4.2m William Herschel Telescope (WHT) on La Palma, the PAU Camera (PAUCam) is a uniquely designed instrument featuring a large field of view (approximately 1° in diameter in total with 40' unvignetted) and a sophisticated filter system. This system, with its 40 NB filters arranged to cover a broad wavelength range with minimal overlap, enables the survey to capture low-resolution spectral energy distributions (Spectral Energy Distribution (SED)s) for millions of galaxies. The innovative design of PAUCam, which includes technical advancements such as movable filter trays, cryogenic operations, and a lightweight carbon fibre housing for the camera itself, underpins the

survey's capability to deliver high-quality photometric data (Padilla et al. 2019).

The observational strategy of PAUS is specifically optimised to maximise both area coverage and redshift precision through the number of filters each field is observed in. Each filter tray holds eight NB filters and so each field needs to be observed five times by moving filter trays in and out. By carefully selecting survey fields–such as the well-studied Cosmic Evolution Survey (Cosmic Evolution Survey (COSMOS)) field and various Canada-France-Hawaii Telescope Legacy Survey (Canada-France-Hawaii Telescope Legacy Survey (CFHTLS)) fields, the survey ensures substantial overlap with other deep photometric datasets, which supply the photometry in other bands. This complementary coverage not only aids in cross-calibration but also allows the implementation of forced photometry* techniques, further improving the accuracy of the measured fluxes and redshifts (Castander et al. 2012). The high-density and precise redshift measurements provided by PAUS are crucial for studies that require an accurate three-dimensional mapping of galaxies, such as clustering analyses, intrinsic alignment measurements, and investigations of galaxy evolution (Eriksen and Gaztañaga 2015a,b,c).

In my thesis, the PAUS dataset plays a central role in measuring the galaxy luminosity function. Accurate determination of photometric redshifts is essential in this context as errors in redshifts directly translate into uncertainties in distance and, consequently, in the intrinsic luminosities of galaxies. The exceptional redshift precision achieved by PAUS minimises these uncertainties, thereby providing a robust foundation for statistical analyses of galaxy populations. Furthermore, the ability to obtain reliable redshifts for a large number of galaxies over wide areas makes PAUS an excellent resource for probing the evolution of galaxies across cosmic time.

*Forced photometry is a technique where the flux is measured within an aperture at a fixed pre-determined position rather than relying on a source detection in the target image. This approach ensures consistent flux measurements across multiple bands, improves flux estimates for faint sources, and allows photometry to be extracted even when the source is undetected in a given band. This process is required for PAUS because narrow-band imaging is noisier than broad-band imaging.

2.2 The PAU Camera (PAUCam)

PAUCam is the core instrument of the PAU survey, engineered to deliver high-precision photometry over a wide field of view. Fig. 2.1 shows PAUCam when it was installed at the prime focus of the 4.2m William Herschel Telescope (WHT) based in La Palma, Spain. PAUCam is designed to image an approximately 1° diameter field, with the central $\sim 40'$ remaining unvignetted (see Fig. 2.2(b) and see also Castander et al. 2012).

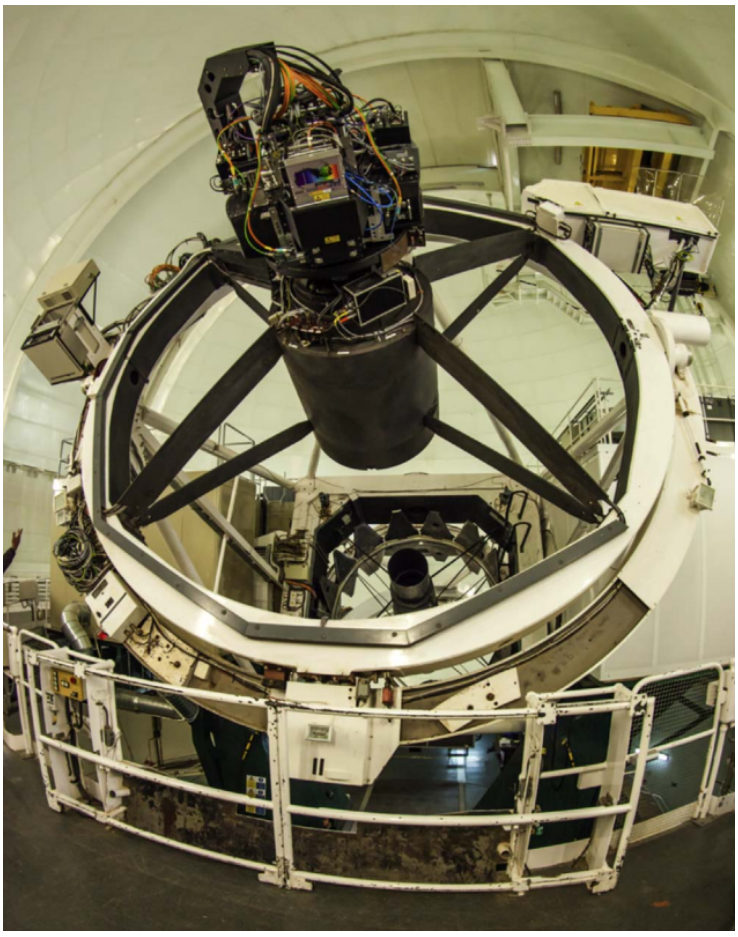
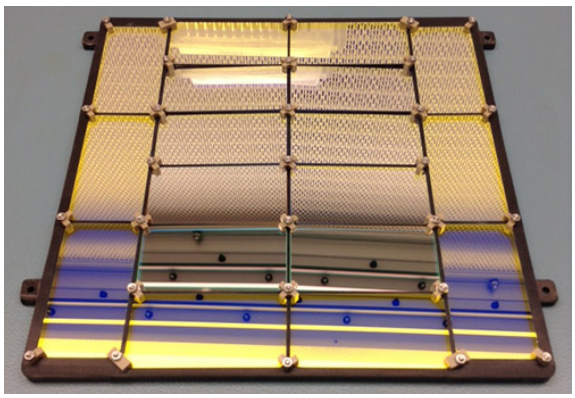
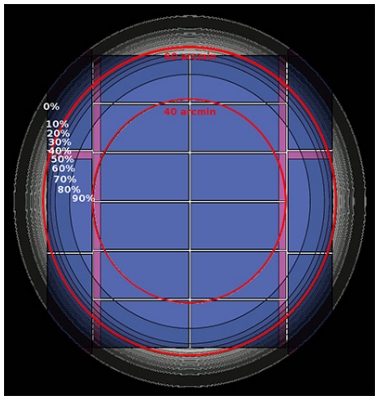


Figure 2.1: The PAUCam mounted at the prime focus of the William Herschel Telescope. Figure taken from Padilla et al. (2019).

The focal plane of PAUCam is populated by 18 Hamamatsu fully depleted Charge-Coupled Device (CCD)s, each with a resolution of 2k x 4k pixels (see Fig. 2.2 (a)). These detectors provide high quantum efficiency (more than 75%



(a) The PAUCam NB tray. The 8 central detectors are assigned to the NB filters, while the 10 external CCDs are covered with BB filters.



(b) The 8 central CCDs are unvignetted, whereas the 10 external detectors experience varying degrees of vignetting.

Figure 2.2: The PAUCam filter tray and their optics performance. Images taken from `https://pausurvey.org/paucam/`.

across the entire covered wavelength range) and are optimised to achieve a pixel scale of $\sim 0.265''$ per pixel. This mosaic arrangement of multiple CCD detectors on the focal plane is essential for capturing a high density of sources simultaneously within a single pointing. Such a layout underpins the survey's ability to measure accurate photometric redshifts and perform statistical studies, such as estimating the galaxy luminosity function.

PAUCam is a multi-filter system. Fig. 2.3 shows the camera filter system featuring a set of 40 NB filters designed to cover the optical wavelength range from 4500 Å to 8500 Å. Each NB filter has a full-width at half maximum (FWHM) of approximately 130 Å and the filters are spaced by roughly 100 Å. This configuration results in nearly contiguous spectral coverage, allowing for the reconstruction of low-resolution spectral energy distributions of observed objects. In addition to the NB filters, a complementary set of 6 BB filters ($ugrizY$) is incorporated to enhance calibration and improve redshift measurements (Castander et al. 2012; Padilla et al. 2019)

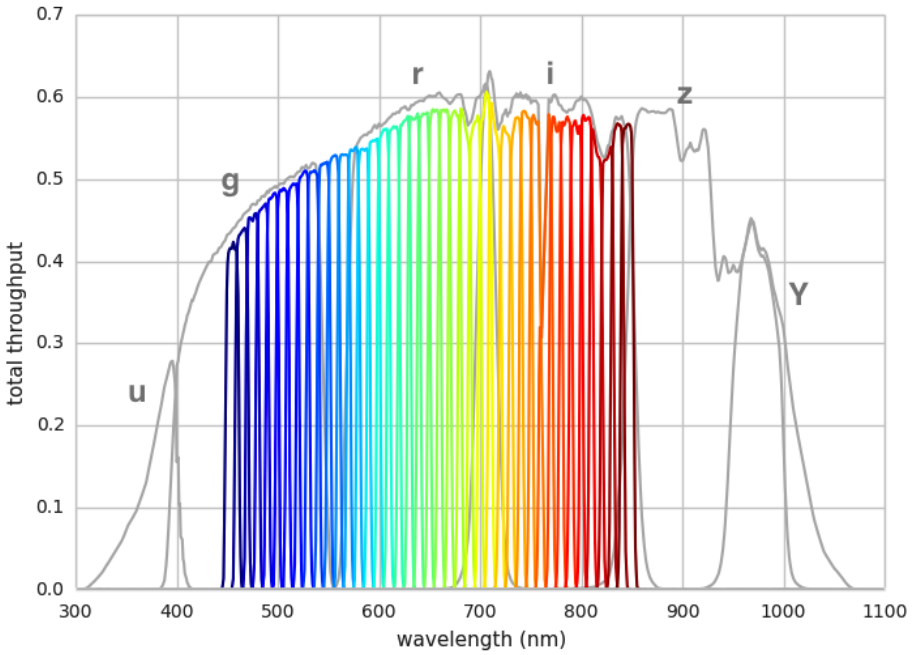


Figure 2.3: The filter system and throughput of PAUCam showing the 40 NB filters (130 Å wide in steps of 100 Å) covering the wavelength range from 4500 Å to 8500 Å together with the 6 BB filters. The filter curves include atmospheric transmission, telescope optics, and CCD quantum efficiency. Image taken from `https://pausurvey.org/paucam/filters/`

2.3 Survey Design

The PAUS is designed to achieve the high photometric precision required for cosmological studies while covering a statistically significant area. By targeting fields that have extensive multi-wavelength data, the survey is optimised for both deep photometric analysis and robust cross-calibration with spectroscopic samples. In particular, the survey focuses on well-established regions, including COSMOS (Scoville et al. 2007), CFHTLS-W1 and W3 (Heymans et al. 2012; Erben et al. 2013), and the GAMA G09 fields (de Jong et al. 2013). Fig. 2.4 shows the positions of these fields and their overlap with ancillary surveys on the sky map. This strategic field selection ensures that PAUS can benefit from the high-quality broadband photometry and detailed spectroscopic data already available in these regions.

The overall design of the PAUS is structured to balance area and depth. Table 2.1 shows that the deep wide fields collectively cover an area of approximately

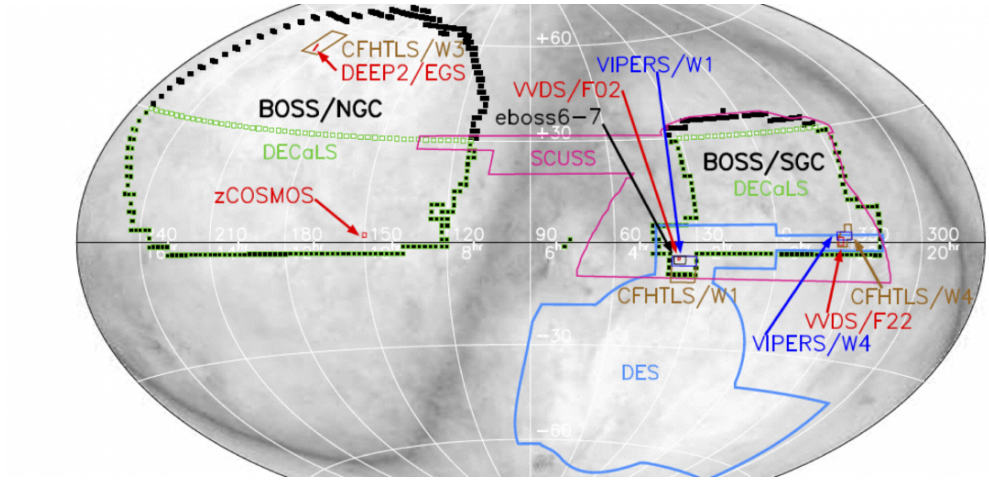


Figure 2.4: A sky map showing observed field positions overlapping with ancillary surveys on the equatorial projected grid. The grey shaded regions show the MW galactic plane. PAUS observed in the fields labelled with CFHTLS/W1, CFHTLS/W3, CFHTLS/W4, and zCOSMOS. Image taken from `https://pausurvey.org/pausurvey/fields/.`

51 deg^{2*} and yield a sample of roughly 1.8 million objects down to $i_{AB} < 23$ (Navarro-Gironés et al. 2024). Such extensive coverage provides a high galaxy number density–reaching up to 3×10^4 galaxies per square degree–thereby enable precise statistical studies, including analyses of galaxy clustering (Eriksen and Gaztañaga 2015a,b,c) and the galaxy luminosity function (this work).

Fields	W1	W3	G09	COSMOS	Total
Area (deg^2)	12.04	22.64	15.7	1	51.38
# Galaxies	401 815	792 664	663 535	13 380	1 871 394

Table 2.1: The survey-covered area along with the number of objects with PAUS NB photometry in the COSMOS, CFHTLS, and Kilo-Degree Survey (KiDS) fields, to a depth of $i_{AB} = 23.0$ (Navarro-Gironés et al. 2024).

An important aspect of the survey design is its deliberate overlap with existing deep photometric surveys. For instance, the COSMOS field offers high-resolution imaging and comprehensive multi-band data, while the CFHTLS and KiDS fields provide additional calibration support. This overlap not only facilitates the validation of photometric redshift measurements but also enhances the overall data quality by allowing for forced photometry techniques that incorporate precise broadband information (see the next section for more details).

*we use the different covered areas due to specific masks and selection effects applied for estimating the luminosity function. See §2.6 for details.

2.4 Measuring Photometric Redshifts

Measuring photometric redshifts has evolved considerably over the past decades. Traditionally, redshift estimates in large imaging surveys relied on broadband photometry, where the SED of a galaxy was sampled coarsely. Early methods employed template-fitting algorithms or empirical calibrations using spectroscopic training sets, which, although effective for many purposes, were limited by the low spectral resolution of broadband filters. This resulted in typical redshift uncertainties of several percent (Ilbert et al. 2006; Moles et al. 2008), which, while acceptable for many cosmological studies (Hildebrandt et al. 2017; Abbott et al. 2018), were not sufficient for applications requiring fine redshift discrimination (Eriksen and Gaztañaga 2015c).

The advent of the PAUS marked a significant improvement by employing 40 NB filters spanning 4500 Å to 8500 Å. This filter set provides a nearly continuous sampling of the galaxy SED, allowing for the detection of subtle spectral features and sharper SED breaks. Stothert et al. (2018) shows that the PAUS NB filters are sensitive to some spectral features including the 4000 Å break or prominent emission lines (OII, OIII, and Hα).

Spectral Features	Rest-frame wavelength (Å)	Redshift range
OII	3727	0.21 - 1.28
OIII	4959/5009	0.0 - 0.70
Hα	6563	0.0 - 0.29
D4000$_N$	3850-3950, 4000-4100	0.17 - 1.07
D4000$_W$	3750-3950, 4050-4250	0.20 - 1.00

Table 2.2: The redshift ranges over the PAUS NB filters can detect some common spectral features. Table adapted from Table 1 in Stothert et al. (2018).

The increased wavelength resolution translates into a dramatic improvement in photometric redshift precision, with typical uncertainties reaching $\sigma_{68}(\Delta z)/(1+z) \simeq 0.0035$ for galaxies brighter than $i_{AB} = 22.5$.

PAUS leverages this detailed spectral information by using a Bayesian SED template-fitting algorithm called the `BCNz2` code to derive photometric redshifts.

The `BCNz2` code models the SED of each galaxy as a linear combination of continuum templates and an additional emission line component. By marginalising over the template coefficients, `BCNz2` computes a full probability distribution $p(z)$ for each galaxy, rather than a single best-fit redshift value. This probabilistic approach not only provides a measure of the redshift precision but also helps in identifying potential outliers in the redshift estimation (Eriksen et al. 2019).

2.5 Photometric Redshift Performance and Data Quality

This section reviews the photometric redshift performance of the PAU Survey. PAUS delivers significantly higher redshift precision compared to traditional broad-band photometric surveys. Eriksen et al. (2019) evaluated the photometric redshift performance of PAUS by comparing redshifts derived with `BCNz2` against zCOSMOS bright spectroscopic data (zCOSMOS DR3 bright: Lilly et al. 1995). Recently, Navarro-Gironés et al. (2024) evaluated the performance against spectroscopic measurements from established surveys including the Sloan Digital Sky Survey (SDSS DR16: Ahumada et al. 2020), the Galaxy and Mass Assembly (GAMA DR3: Baldry et al. 2017), the VIMOS Public Extragalactic Redshift Survey (VIPERS: Scodeggio et al. 2018), the DEEP2 redshift survey (Davis et al. 2003; Newman et al. 2013), KiDs-COSMOS (KiDS DR5: Davies et al. 2014), the 2-degree Field Galaxy Redshift Suvey (2dFGRS: Colless et al. 2001, the VIMOS VLT DEEP Survey (VVDS: Fèvre et al. 2013), and the 3D-HST (Brammer et al. 2012). Fig. 2.5 shows the photometric redshift measured using `BCNz2` code for data from PAUS W1, W3, and G09 fields vs. the spectroscopic redshifts available.

Eriksen et al. (2019) provides a full description of the metrics used to evaluate the photometric redshift performance. Here, we briefly review these metrics for completeness. The photometric redshift performance from PAUS is quantified using several statistics based on the quantity Δ_z, defined as:

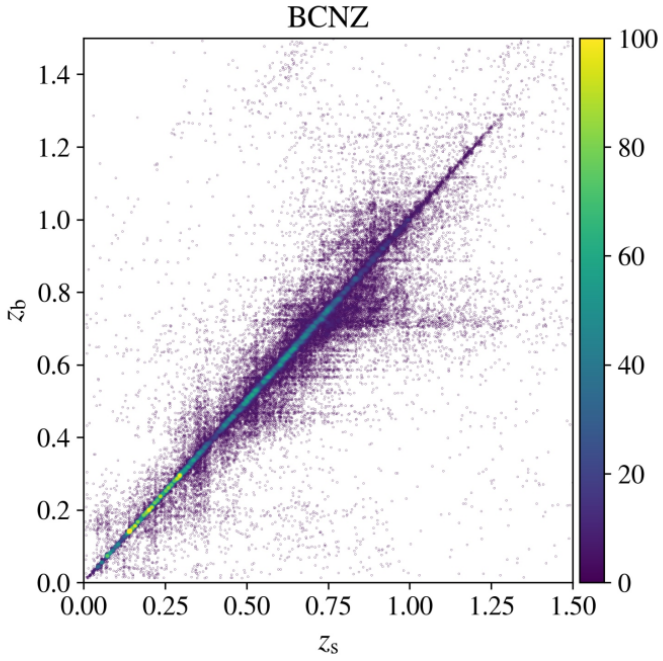


Figure 2.5: `BCNZ2` photometric redshift against spectroscopic redshifts for the PAUS W1, W3 and G09 fields. The photometric redshift is plotted on the y-axis and the spectroscopic is on the x-axis. The colour bar represents the number of galaxies. Plot is taken from Navarro-Gironés et al. (2024).

$$\Delta_z = \frac{z_b - z_s}{1 + z_s}, \tag{2.1}$$

where z_b and z_s are the photometric and spectroscopic redshifts, respectively. The photometric precision is defined using the centralised scatter σ_{68}, given by:

$$\sigma_{68} \equiv \frac{P[84] - P[16]}{2}, \tag{2.2}$$

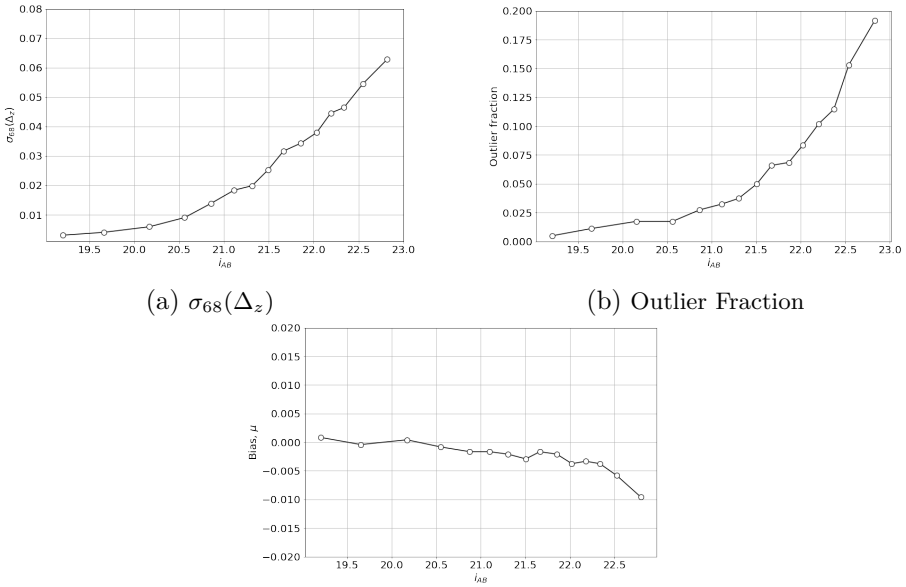
where $P[84]$ and $P[16]$ are the 84^{th} and 16^{th} percentile of the Δ_z distribution, respectively. The systematic offset between the photometric and spectroscopic redshifts is described by the bias, μ, defined as the median of the residual distribution:

$$\mu = \text{median}(z_b - z_s). \tag{2.3}$$

Finally, following the most recent definition from Navarro-Gironés et al. (2024), a galaxy is considered an outlier if it satisfies $|\Delta_z| > 0.1$*.

*Eriksen et al. (2019) uses $|\Delta_z| > 0.02$ in the early demonstration of photometric redshift performance in the COSMOS field.

For the data used for measuring the galaxy luminosity function in this work, the metrics of the photometric redshift performance are shown in Fig. 2.6. The precision of the photometric redshifts is approximately $\sigma_{68}(\Delta_z) \sim 0.003$ for the brightest objects in the sample ($i_{AB} \sim 19$) and $\sigma_{68}(\Delta_z) \sim 0.06$ for the galaxies with $i_{AB} \sim 23$. The outlier fraction is approximately 20% at the faintest magnitude bin. Overall, the photometric redshifts show negligible systematic offsets (biases) compared to previous studies (Navarro-Gironés et al. 2024). This result shows a good agreement with the early performance found in Eriksen et al. (2019).



(a) $\sigma_{68}(\Delta_z)$ (b) Outlier Fraction

(c) Bias, μ

Figure 2.6: The performance metrics of the photometric redshift from the PAUS wide fields (W1, W3, and G09 fields) measured using the `BCNz2` code, shown as a function of $i-$band magnitude. These metrics include (a) centralised scatter (σ_{68}), (b) outlier fraction, and (c) bias (μ). Plot adapted from Navarro-Gironés et al. (2024).

In this work, we use the PAUS W1 and W3 fields* for measuring the luminosity function. Fig. 2.7 shows the area-weighted number count distribution as a function of i-band magnitude for both fields with a magnitude cut at $i_{\rm AB} = 23.0$† and with different additional quality selections. The weighting is based on the area covered

*Production ID 1044 for W1 field and 1045 for W3 field.

†Despite the survey reaching a depth of $i_{\rm AB} = 24$, reliable redshifts are only available for galaxies brighter than $i_{\rm AB} = 23$, as discussed in Castander et al. (2012); Navarro-Gironés et al. (2024). The samples I accessed from the PAUS database (Prod. ID 1044 and 1045) therefore include only galaxies brighter than this magnitude limit.

by all galaxies in the respective fields. The galaxy number counts from both W1 and W3 fields show a good agreement. However, when selecting only galaxies with redshifts measured using more than 30 NB filters (out of 40), the number counts are suppressed by approximately 5% to 20% across all magnitude bins. Furthermore, applying a quality cut that retains only the best 50% of photometric redshifts results in a more complex scaling. The number counts begin to diverge around $i_{\rm AB} \sim 21$, with the difference increasing toward fainter magnitudes.

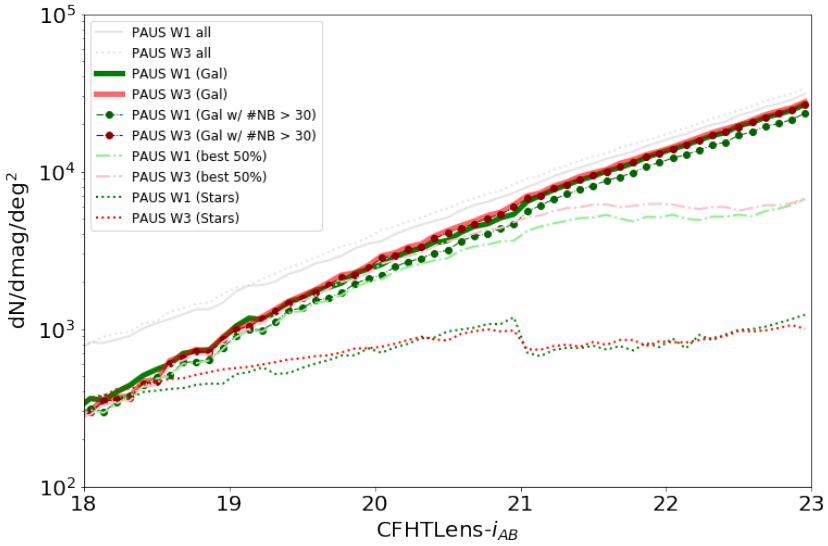


Figure 2.7: The object number counts as a function of $i-$band magnitude with different quality cuts. This plot is similar to Fig. 2 of Manzoni et al. (2024), but here PAUS W1 and W3 fields are shown separately. Red colours represent the W1 field, while green colours represent the W3 field. The number counts here include all objects (grey lines), objects classified as galaxies (solid lines), galaxies with photometric redshifts measured using more than 30 NB filters (lines with filled circles), galaxies with photometric redshifts in the best 50% of the sample (dashed lines), and object classified as stars (dotted lines).

2.6 The Effective Survey Area

In this section, we describe the method used to measure the solid angle (effective area) covered by the PAUS fields, and how this method can be generalised to account for various selection cuts applied to the data.

We adopt a Monte Carlo (MC) approach to estimate the area subtended by irregularly shaped observational footprints. In our implementation, a large number

of random points is uniformly distributed over a two-dimensional spherical surface encompassing the survey footprint. The area covered by the survey is then estimated by counting the fraction of these points that lie sufficiently close to actual galaxy positions in the dataset:

$$A_{\rm survey} = \frac{N_{\rm hit}}{N_{\rm total}} A_{\rm random}, \tag{2.4}$$

where $A_{\rm survey}$ is the estimated survey, $N_{\rm hit}$ is the number of random points within a distance threshold d_p of any galaxy position in the selected sample, $N_{\rm total}$ is the total number of random points, and $A_{\rm random}$ is the total area spanning by the random sample.

To validate our area calculation, we applied the method to a patch of random points that included masks resembling those in the actual observations. These masks consisted of two main types: **(a) circular regions** produced by saturated pixels around bright stars, and **(b) rectangular regions** associated with filter completeness. We applied these masks to the random points to reproduce a footprint similar to that of the PAUS data, as shown in Fig. 2.8. This setup provides a close analogue to the real dataset shown in Fig. 2.11. Since the mask dimensions are known (e.g., the diameters of the circular masks and the widths and heights of the rectangular masks), we were also able to calculate the covered area analytically for comparison.

The matching threshold distance d_p is defined using the mean nearest-neighbour separation of the galaxies in the full sample, which we find to be approximately $d_{p,\rm mNN} \sim 0.00245$ degrees. We also tested the impact of the choice of d_p on the estimated areas by varying its value while keeping the total numbers of random points fixed at $N_{\rm total} = 10^6$. As shown in Fig. 2.9, the area estimates remain accurate to within 1% provided that d_p lies between $0.8 \times d_{p,\rm mNN}$ and $1.8 \times d_{p,\rm mNN}$. Based on this, we adopted $d_p = 0.00245$ degrees for all calculations throughout this work. This ensures a consistent and adaptive distance scale for estimating the area

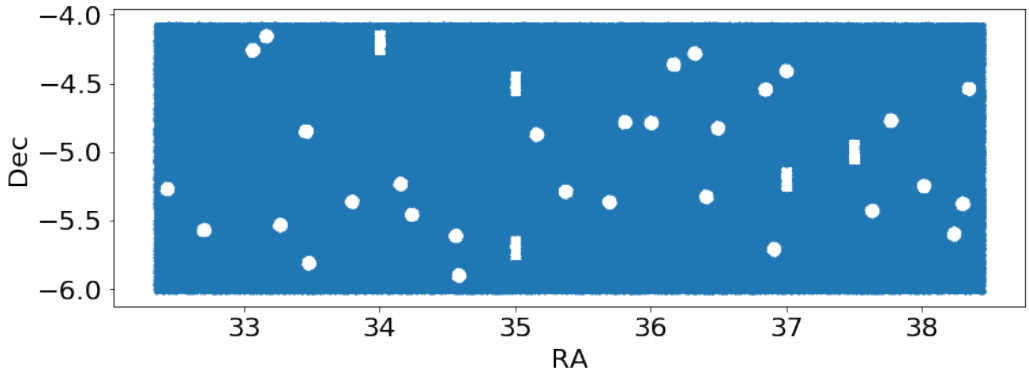


Figure 2.8: Random points (blue dots) with overlaid masked regions used to validate the Monte Carlo approach for measuring the effective survey area. The two types of masks are circular (from saturated stars) and rectangular (from filter coverage). This setup mimics the masking "patterns" in PAUS observations.

across different sample.

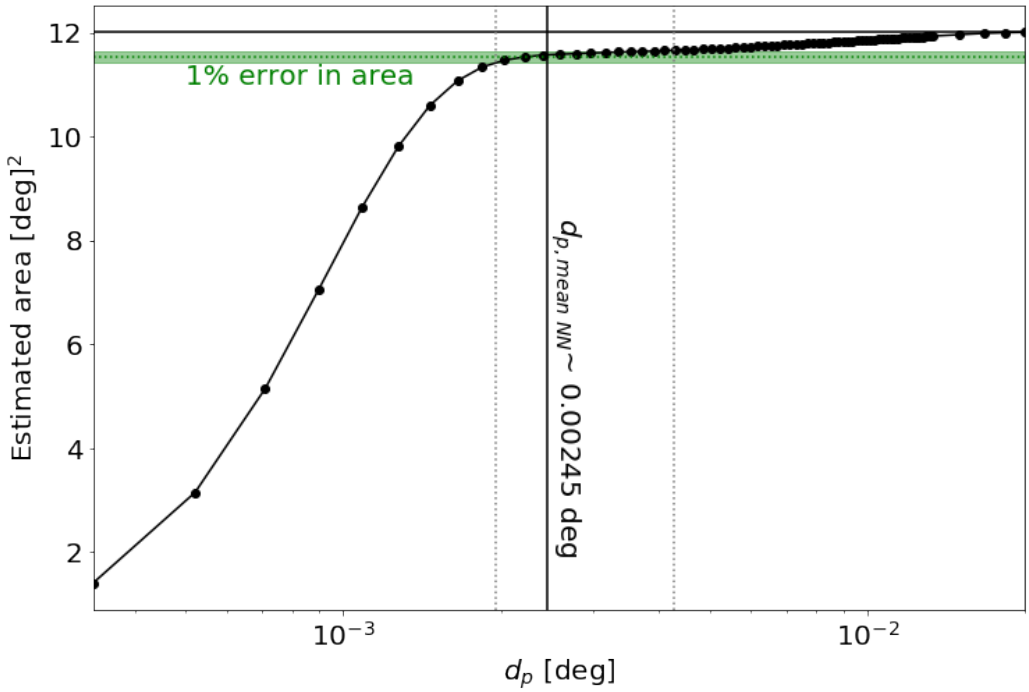


Figure 2.9: Estimated survey area as a function of the choice of d_p with $N_{\rm total} = 10^6$. Black dots with the solid curve show the estimated areas. The green dotted horizontal line marks the exact area covered by the random points in Fig. 2.8, with the ±1% error margin indicated by the shaded green band. The vertical solid black line denotes the mean nearest-neighbour separation from the PAUS data, while the vertical grey dotted lines indicate the range of d_p values that yield area estimates within 1% accuracy. The horizontal solid black line corresponds to the area covered by the random points without masking.

We carried out a similar test to assess the impact of the total number of

random points $N_{\rm total}$ on the area estimation, this time fixing $d_p = 0.00245$ degrees. To evaluate the stability of the results, we repeated the calculation 100 times and examined the fluctuations in the estimated area. Fig. 2.10 shows the estimated area as a function of $N_{\rm total}$. We find that even with as few as 4×10^3 points, the area can be estimated to within 1% accuracy, albeit with noticeable scatter. For robustness, we adopted $N_{\rm total} = 10^6$, which substantially reduces fluctuations. Since the area estimation is only performed once per field and per type of selection cut (see Table 2.3), this choice does not significantly affect computational efficiency.

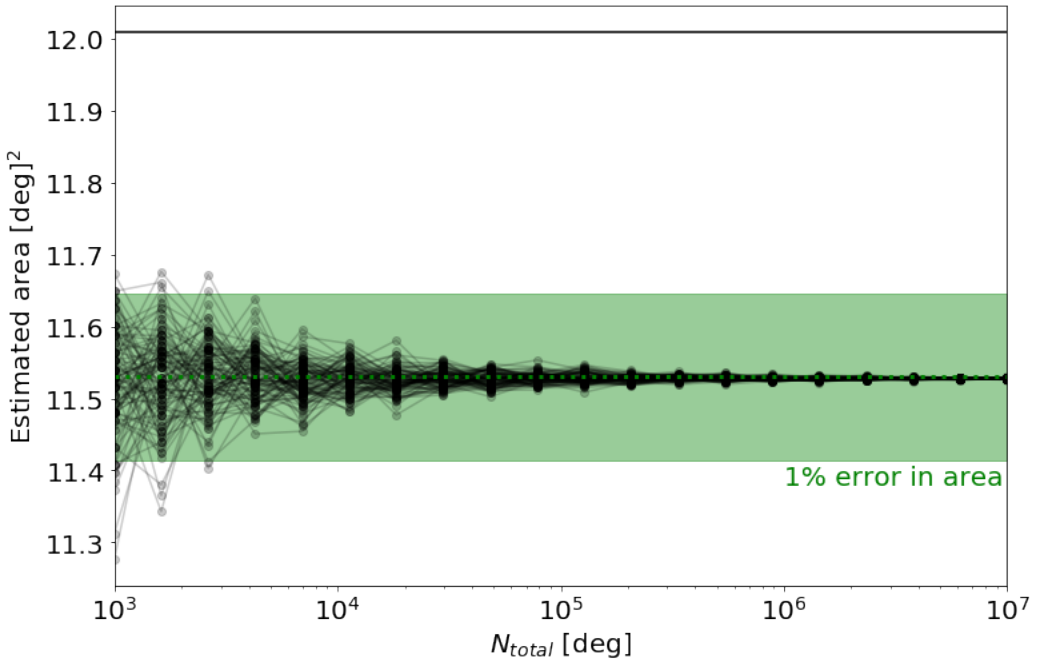


Figure 2.10: Estimated survey area as a function of $N_{\rm total}$ with $d_p = 0.00245$ degrees. Black dots with the solid curve show the estimated areas (repeated 100 times). The green dotted horizontal line marks the exact area covered by the random points in Fig. 2.8, with the $\pm 1\%$ error margin indicated by the shaded green band. The horizontal solid black line corresponds to the area covered by the random points without masking.

Fig. 2.11 illustrates the principle behind the effective area calculation using the MC technique to handle irregular mask shapes. In each panel, the grey points represent the uniformly distributed random points across the survey footprint. The black points correspond to galaxy positions from the selected sample. The red points make the subset of random points that lie within the matching threshold distance (d_p) of any galaxy–these are counted as "hits" in the area estimation. The

ratio of red to grey points directly informs the fractional area covered by the galaxy sample relative to the total sampling region.

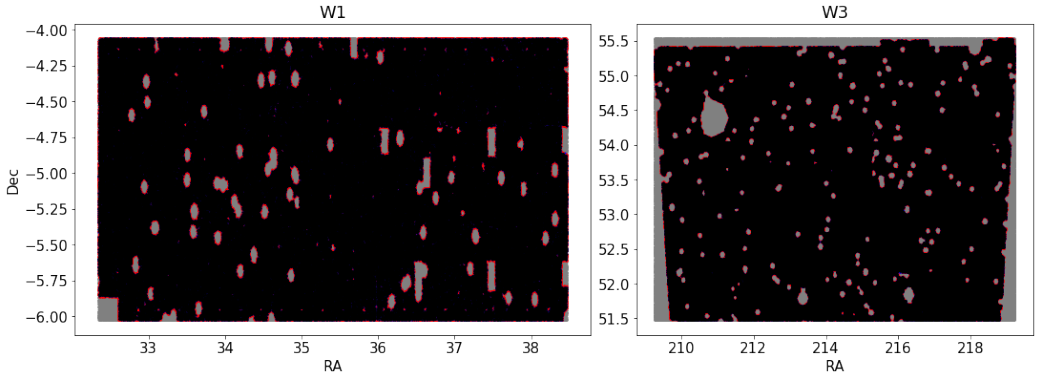


Figure 2.11: The area measurement for PAUS W1 (Production ID 1044) and W3 (Production ID 1045) fields)

This method is flexible and generalisable–it can be applied to any subsets of galaxies defined by a given set of selection criteria. For example, one may wish to estimate the area after applying cuts such as selecting only object classified as galaxies, requiring that redshift be computed using more than 30 NB filters, or selecting the top 50% of galaxies based on photometric redshift quality.

Fig. 2.12 illustrates the sky distributions of galaxies in the W1 and W3 fields for the full galaxy sample and for galaxies with at least 30 NB used for photo-z measurement. Blue points represent the full galaxy sample, while black points show the more restricted subsets. These clearly demonstrate how selection cuts reduce the effective footprints.

To validate the robustness of the area correction, Fig. 2.13 presents the area-normalised number counts for the W1 and W3 fields under various selection scenarios. After applying the appropriate area weighting, the galaxy number counts remain consistent across fields, confirming that the method correctly accounts for changes in sky coverage due to selection effects.

The measured effective areas for different cuts are summarised in Table. 2.3. These values are used throughout this work to ensure proper normalisation in the galaxy luminosity function.

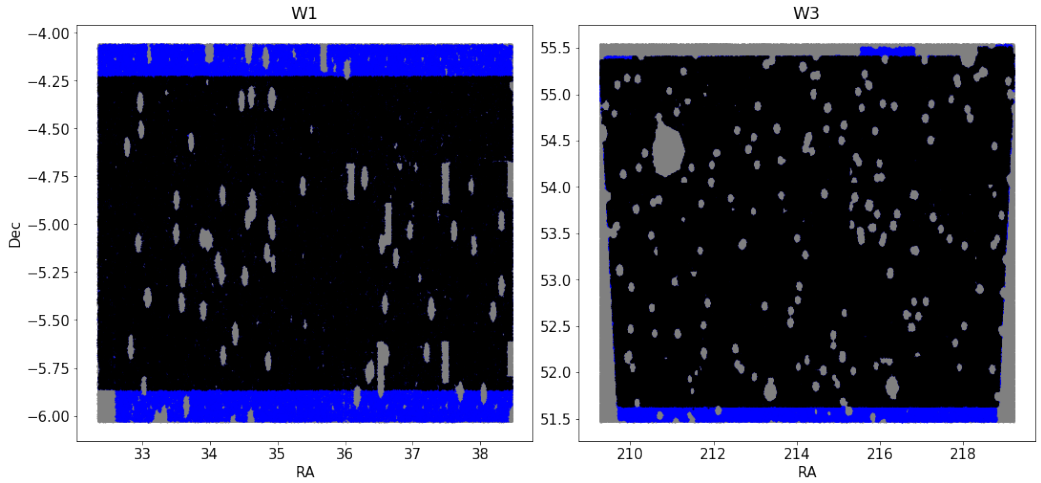


Figure 2.12: The area measurement for PAUS W1 (Production ID 1044) and W3 (Production ID 1045) fields)

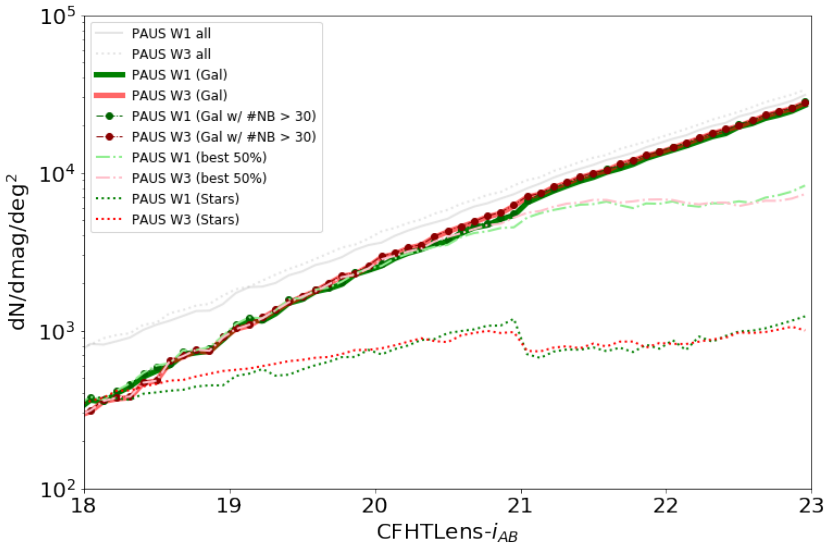


Figure 2.13: The object number counts as a function of $i-$band magnitude with different quality cuts. This plot is similar to Fig. 2 of Manzoni et al. (2024), but here shows the PAUS W1 and W3 fields separately. Red colours represent the W1 field, mean while green colours represent the W3 field. The number counts here show all objects (grey lines), objects classified as galaxies (solid lines), objects classified as galaxies with photometric redshifts measured using more than 30 NB filters (lines with filled circles), galaxies with photometric redshifts better than 50% of the sample, and object classified as stars (dotted lines).

2.7 Summary

In this chapter, we presented an overview of the Physics of the Accelerating Universe Survey (PAUS) and its key role in enabling precise photometric redshift

Selection Cuts	W1 $[deg^2]$	W3 $[deg^2]$
Galaxies only	11.178	20.532
Galaxies with > 30 NB	9.334	19.552
Galaxies in best50%	9.032	18.901

Table 2.3: The effective survey-covered area after applying different selection cuts, including sample with galaxies only, galaxies with more than 30 NB used to measure the photo-z, and galaxies in the best 50% photo-z quality sample

measurements for large-scale cosmological studies. We described the design and instrumentation of PAUCam, highlighting its unique narrow-band filter system and wide-field capabilities. By targeting fields with extensive multi-wavelength coverage, such as COSMOS, CFHTLS W1 and W3, and GAMA G09, PAUS ensures robust calibration and facilitates the integration of ancillary data.

We discussed the advantages of using NB photometry for redshift estimation, with the `BCNz2` code providing high-precision, probabilistic redshift measurements. Performance metrics including scatter, bias, and outlier fraction confirm that PAUS achieve sub-percent redshift precision for bright galaxies and maintains good consistency with spectroscopic redshifts across a wide magnitude range.

Furthermore, we described the methodology used to measure the effective survey area, including how the MC technique can be generalised to account for various selection effects. Accurate area estimation is essential for number density calculations and luminosity function measurements. The approach was validated by comparing area-weighted number counts across different quality cuts in the W1 and W3 fields.

Chapter 3

Galaxy Formation Model

Galaxy formation is one of the central topics in contemporary astrophysics, bridging observational astronomy and theoretical cosmology. Understanding how galaxies form and evolve within the Universe involves studying processes that operate across a wide range of spatial scales. Currently, the most widely accepted framework for galaxy formation is based on hierarchical clustering within the Cold Dark Matter (CDM) cosmological paradigm (e.g. Springel et al. 2006). According to this hierarchical model, cosmic structures grow through gravitational instability, beginning with initial small-scale density fluctuations. These small-scale structures originate from quantum fluctuations during the inflationary epoch and leave observable imprints as anisotropies in the Cosmic Microwave Background (CMB) radiation (Guth 1981; Mukhanov and Chibisov 1981;Planck Collaboration et al. 2020). Over cosmic time, these structures progressively merge to build larger systems, including galaxies, galaxy groups, and galaxy clusters.

Within this hierarchical scenario, galaxies form inside dark matter halos, gravitational wells that dominate their formation and evolution (White and Rees 1978). As halos merge and grow, they accrete gas from the Intergalatic Medium (IGM). In massive halos, this gas typically heats via shock-heating, forming hot gas atmospheres at roughly the virial temperature before cooling and fuelling star formation (White and Rees 1978; Cole et al. 1994; Bower et al. 2006). However, at higher

redshifts and lower masses, gas can remain cool and dense, efficiently cooling more quickly than it heats, thus accreting onto galaxies without ever approaching the halo virial temperature–a process known as "cold mode" accretion (Kereš et al. 2005).

Modelling galaxy formation poses significant theoretical challenges due to complex, interconnected astrophysical processes involved such as gas cooling, star formation, supernova (SN) and active galactic nucleus (AGN) feedback, chemical enrichment, and galaxy mergers (Lacey et al. 2016). To address these challenges, two complementary approaches have emerged: hydrodynamical simulations and semi-analytical models.

Hydrodynamical simulations numerically solve baryonic physics alongside dark matter dynamics to predict galaxy properties, providing detailed insights into complex astrophysical processes at the expense of significant computational resources (Guo et al. 2016; Ayromlou et al. 2021). Due to their computational complexity, these simulations are typically limited in terms of exploring extensive parameter spaces or rapidly comparing with observational data in a fine-grained model calibration. Hydrodynamical simulations are also limited by resolution resulting in being unable to directly compute unresolved astrophysical processes (e.g., star formation, supernova feedback, AGN feedback, and black hole accretion), which require the implementation of "subgrid" model prescription of these physical processes (Kugel et al. 2023). Semi-analytic models, in contrast, employ computationally efficient analytic prescriptions of astrophysical processes within dark matter halo merger histories extracted from cosmological N-body simulations, enabling extensive exploration of parameter space and rapid comparisons with observational constraints (Baugh et al. 2019; Gonzalez-Perez et al. 2014; Lacey et al. 2016; Griffin et al. 2019). Recent comparative studies have shown that despite their simplified treatment of baryonic processes, semi-analytic models can accurately reproduce many global properties observed in more computationally intensive hydrodynamical simulations, highlighting their complementary strengths (Neistein et al. 2012; Guo

et al. 2016; Ayromlou et al. 2021).

Among semi-analytic models, GALFORM, developed at Durham University, stands out as a sophisticated, widely adopted model (Cole et al. 2000; Bower et al. 2006; Lacey et al. 2016). GALFORM systematically incorporates critical mechanisms such as star formation laws based on molecular gas content, variations in stellar Initial Mass Function (IMF)s, and detailed treatments of AGN feedback. AGN feedback, in particular, plays a crucial role by suppressing gas cooling in massive halos, thereby addressing previously unresolved discrepancies like the overproduction of luminous galaxies at low redshift (Bower et al. 2006; Griffin et al. 2019).

This chapter focuses on GALFORM outlining its development, key physical mechanisms, and application in the interpretation of observational galaxy data. By comparing model predictions with observation constraints, astrophysicists continuously refine physical processes within GALFORM striving for a comprehensive understanding of galaxy formation and evolution.

3.1 Why Semi-Analytical Models?

Semi-analytical models offer a robust framework for understanding galaxy formation and evolution, providing significant computational advantages over fully numerical hydrodynamical simulations. Due to the inherently complex and multi-scale nature of galaxy formation, modelling all relevant physical processes explicitly through numerical hydrodynamics is computationally intensive and often prohibitively expensive for exploring a wide range of physical parameters or cosmological scenarios. Furthermore, many astrophysical processes, such as star formation, supernova feedback, and AGN feedback, occur on scales too small to be directly resolved in cosmological simulations, requiring the implementation of uncertain and parametrized “subgrid” models. These subgrid models introduce realistic galaxy populations (Kugel et al. 2023). Semi-analytic models, in contrast, employ simplified analytic or phenomenological prescriptions for these processes, providing

computational efficiency and considerable flexibility in adjusting assumptions or parameters. This flexibility enables extensive exploration of astrophysical scenarios, systematic comparison with observations, and rapid identification of the astrophysical mechanisms most important for shaping galaxy properties.

Semi-analytic models circumvent these limitations by using analytic approximations and empirical prescriptions to describe key astrophysical processes. These analytic prescriptions are embedded within cosmological dark matter halo merger trees derived from large-scale N-body simulations, enabling efficient exploration of extensive parameter spaces and rapid comparison with observational data. An example of a comparison between the outputs of models can be different semi-analytical models can be found in Contreras et al. (2013), who shows how different semi-analytic models populate dark matter halos and predict galaxy clustering using a common underlying N-body simulation. These comparison shows that some model predictions are robust to different implementations of the galaxy formation physics (e.g. such as the number of galaxies ranked by their stellar mass).

Moreover, semi-analytic models offer an inherently flexible approach, making it easier to interpret the physical assumptions underlying galaxy formation outcomes. This flexibility helps researchers isolate and investigate the impact of individual processes, thus identifying the key drivers of observable galaxy properties and guiding further refinement of theoretical models (White and Frenk 1991; Merson et al. 2013).

Additionally, semi-analytic models facilitate the construction of mock galaxy catalogues, which play a critical role in preparing for and interpreting galaxy surveys. Because of their computational efficiency, semi-analytic models can be implemented in large-volume N-body simulations, enabling the generation of statistically representative mock galaxy populations that reflect the survey geometry, selection functions, and cosmic variance (Contreras et al. 2015). These mock catalogues allow researchers to quantify observational systematics, optimise observational strategies, and validate data analysis pipelines for cosmological studies (Merson et al. 2013;

Stothert et al. 2018).

Given these advantages, semi-analytic models remain an essential and complementary approach alongside hydrodynamical simulations. They allow astrophysicists to test galaxy formation theories systematically and refine their understanding through direct comparisons with observational data across cosmic time.

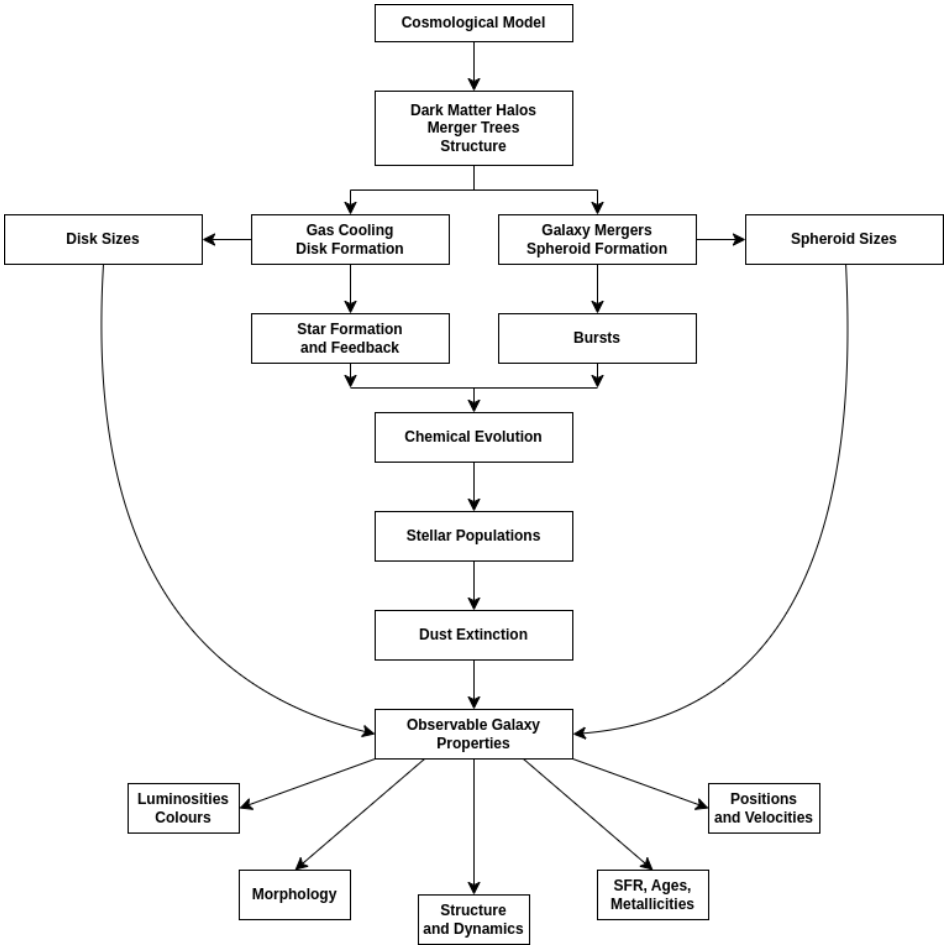


Figure 3.1: A schematic overview of GALFORM. Reproduced from Cole et al. (2000).

3.2 GALFORM

GALFORM is a semi-analytic galaxy formation model developed to address complex astrophysical processes systemically within a hierarchical clustering framework. At its core, GALFORM uses cosmological N-body simulations to construct merger trees that represent the hierarchical growth and merging of dark matter halos. These

merger trees serve as the backbone upon which analytic prescriptions describing galaxy formation processes are implemented.

Fig 3.1. illustrates the overall framework of GALFORM in a comprehensive schematic (adapted from fig 1. in Cole et al. (2000)), highlighting how each astrophysical process connects with the hierarchical growth of cosmic structures. This schematic provides a clear visualization of how the analytic prescriptions and the hierarchical merger trees interact to predict observable galaxy properties. Key components of the GALFORM framework include (i) the formation and merging of dark matter halos; (ii) the shock-heating and radiative cooling of gas inside dark matter halos, which leads to the formation of galactic disks; (iii) star formation in galaxy disks and starbursts; (iv) feedback from supernovae, from AGN and from photo-ionization of the IGM; (v) galaxy mergers driven by dynamical friction and disk instabilities in galaxy disks; (vi) calculation of the sizes of disks and spheroids; and (vii) chemical enrichment of stars and gas.

The detailed descriptions of the physical processes modelled in GALFORM are represented in Cole et al. (2000) and Lacey et al. (2016). Here I summarise these processes for completeness.

3.2.1 Formation and merging of dark matter halos

In the version GALFORM used in this work, halo merger histories are extracted directly from cosmological N-body simulations. Halo-finding algorithms (such as Fiends-of-Friends (Davis et al., 1985) and SUBFIND (Springel et al., 2001)) are applied to identify halos and subhalos at each simulation snapshot, and these are then linked across time to construct hierarchical merger trees (see for example Jiang et al. 2014). These simulation-based merger trees serve as the foundation upon which the galaxy formation physics in GALFORM is modelled.

These merger trees track the entire merger history of each dark matter halo, detailing precisely when halos form, merge, and accrete matter. GALFORM populates

these halos with baryonic matter and applies analytic prescriptions for astrophysical processes, thus linking halo assembly directly to observable galaxy properties.

3.2.2 Halo finding in simulations

In modern implementations of the `GALFORM` model, the merger histories of DM halos are extracted directly from cosmological N-body simulations. To build these merger trees, gravitationally bound structures–halos and subhalos–must first be identified in the simulation outputs. This is accomplished using halo finding algorithms followed by merger tree construction tools.

The initial identification of halos typically begins with a Friends-of-Friends (`FoF`) algorithm (Davis et al. 1985), which groups particles based on spatial proximity using a percolation algorithm and fixed linking length which is a specified fraction of the mean interparticle separation. However, `FoF` groups can artificially connect structures through tenuous bridges, leading to over-merging. To resolve this, substructure finders such as `SUBFIND` (Springel et al. 2001) are employed. `SUBFIND` decomposes `FoF` halos into hierarchically nested, self-bound subhalos by identifying local overdensities and performing iterative unbinding procedures. The most massive of these subhalos is designated the central subhalo, with others classified as satellites.

Once subhalos are identified at each snapshot, a merger trees algorithm is used to link subhalos across time. In this work, we use the `D-TREES` algorithm developed by Jiang et al. (2014), which robustly tracks the most bound cores of subhalos to construct coherent evolutionary links, even across snapshots where a subhalo may be temporarily disrupted or not well identified. This procedure produces a subhalo merger tree which is then post-processed by the `Dhalo` algorithm to build physically motivated halo merger trees. The `Dhalo` algorithm ensures that halo masses grow monotonically with time, and that subhalos having passed through a more massive host retain their identity and association with that host, reflecting the expected

behaviour of baryons during such interactions.

A study by Gómez et al. (2022) investigated the impact of different halo finders and tree builders on galaxy formation predictions using `GALFORM`. They compared four major combinations of halo finders and merger tree builders: HBT (Han et al. 2012), ROCKSTAR (Behroozi et al. 2013a) - CONSISTENTTREE (Behroozi et al. 2013b), SUBFIND (Springel et al. 2001) - DTREES (Jiang et al. 2014), VELOCICRAPTOR (Elahi et al. 2011) - DTREES (Jiang et al. 2014). Each method varies in how it defines halo boundaries, distinguishes between central and satellite structures, and handles particle unbinding. Despite these differences, the final galaxy populations predicted by `GALFORM` were found to be remarkably robust to the halo finder and tree builder used.

Here, we adopt the `SUBFIND-DTREES-Dhalo` combination, where halos and subhalos are found using `FOF` and `SUBFIND`, their evolutionary links are built using `D-TREES`, and the final merger trees are constructed via the `Dhalo` algorithm. This method is consistent with the one used to generate the lightcone mock catalogue analysed in later chapters (See Chapter §4 for lightcone mock catalogue construction).

`GALFORM` generally assumes halos have density profiles approximating Navarro-Frenk-White (NFW) profiles (Navarro et al. 1997), described by:

$$\frac{\rho_{DM}(r)}{\rho_{\rm crit}} = \frac{\delta_{\rm c}}{(r/r_{\rm s})(1+r/r_{\rm s})^2}, \tag{3.1}$$

where $\rho_{\rm crit} = 3H^2/8\pi G$, $\delta_{\rm c}$ is a characteristic density, $r_{\rm s}$ is the scale radius. $r_{\rm s}$ is related to the virial radius by the concentration, $r_{\rm s} = r_{\rm vir}/c_{\rm NFW}$ (see Navarro et al. (1997) for the analytical prescription of $c_{\rm NFW}$ and discussion about $\delta_{\rm c}$). This assumption affects several aspects of the model; the gravitational potential defined by the halo profile influences the hydrostatic equilibrium and cooling rate of hot gas within halos, and also impacts the sizes of the disk and bulge components formed from the cooling gas. While the gas is allowed to develop its own density profile,

the underlying dark matter potential–shaped by the NFW assumption–remains a key determinant in both thermal and structural evolution.

3.2.3 Shock-heating and radiative cooling of gas

In GALFORM, the treatment of gas cooling closely follows the framework introduced by White and Frenk (1991). When baryonic gas falls into a dark matter halo, it is assumed to be shock-heated approximately to the halo virial temperature. This hot gas forms a quasi-static, spherically symmetric atmosphere surrounding the central region of the halo. The hot gas then radiatively cools and sinks towards the halo centre, forming a rotationally supported disk where star formation subsequently occurs. Note that the final stages of cooling by excitation of the vibrational and rotational modes of molecules like H_2 are not explicitly modelled, but are assumed to operate on a shorter timescale than the radiative cooling which takes the gas to 10000K.

The cooling rate depends on the gas density, temperature, and chemical composition (metallicity). GALFORM calculates the cooling radius, defined as the radius within which gas can cool on a timescale shorter than the age of the halo. Gas within this radius is assumed to cool and accrete onto the central galaxy on a free-fall timescale, while gas outside this radius remains hot and in equilibrium with the gravitational potential of the halo. The cooling time $t_{\rm cool}$ at a radius r within the halo is given by

$$t_{\rm cool} = \frac{3}{2}\frac{k_B}{\mu m_H}\frac{T_{\rm vir}}{\rho_{\rm gas}(r)\Lambda(T_{\rm vir}, Z)}, \tag{3.2}$$

where μ is the mean molecular weight of the gas, m_H is the mass of hydrogen, k_B is Boltzmann's constant, $T_{\rm vir}$ is the halo virial temperature, $\rho_{\rm gas}(r)$ is the gas density at radius r, and $\Lambda(T_{\rm vir}, Z)$ is the cooling function dependent on gas temperature and metallicity. In GALFORM, the gas density profile within the halo is typically assumed to follow a singular isothermal sphere or a modified form based on more recent numerical simulations.

However, recent hydrodynamical simulations (e.g. Kereš et al. 2005) have shown that this classical *hot-mode* cooling picture does not always hold, particularly at high redshift. In denser environments and low-mass halos, the cooling times can be much shorter than the free-fall time–the time it takes for gas to collapse to the centre under gravity. In these regimes, the free-fall time, rather than the cooling time, becomes the limiting factor for gas accretion. Moreover, a significant fraction of gas may never be shock-heated to the virial temperature. Instead, it accretes directly along cold, dense filaments in what is referred to as *cold-mode* accretion. This process is inherently aspherical and more efficient at delivering gas to the central galaxy in the early Universe, playing a crucial role in the early buildup of galactic baryons.

While `GALFORM` primarily models hot-mode accretion through the cooling flow paradigm, cold-mode accretion can be approximately captured by treating halos in which the cooling radius exceeds the virial radius as experiencing rapid cooling. In such cases, the infalling gas is assumed to accrete on a timescale comparable to the free-fall time, rather than being shock-heated.

This combined treatment of cooling and free-fall timescales forms a more complete picture of gas accretion and sets the stage for modelling star formation history of galaxies in the `GALFORM` semi-analytic framework.

3.2.4 Galaxy mergers

Galaxy mergers are a fundamental process in the hierarchical formation of structure, driving the transformation of galaxy morphology, triggering bursts of star formation, and contributing to the growth of stellar mass and spheroids. In `GALFORM` mergers are modelled by tracing the dynamical evolution of satellite galaxies after their host dark matter halos merge, following prescriptions based on gravitational interactions and dynamical friction (Cole et al., 1994; Lacey et al., 2016).

When two dark matter halos merge, the smaller halo becomes a satellite within

the larger halo. Subsequently, satellite galaxies lose orbital energy through dynamical friction–a gravitationally induced drag force exerted by dark matter particles in the host halo–leading eventually to a merger with the central galaxy. The original model of Cole et al. (1994), calculated a merger timescale $\tau_{\rm merge}$ using an analytic prescription derived from the dynamical friction timescale introduced by Chandrasekhar (1943). In the newer model, GALFORM automatically includes the effects of tidal stripping of the dark matter subhalo hosting the smaller galaxy on the dynamical friction rate. This is done by replacing the Chandrasekhar formula with a modified expression obtained from cosmological N-body or hydrodynamical simulations (Jiang et al. 2008). The merger timescale is expressed as follows:

$$\tau_{\rm merge} = \frac{f(\epsilon)}{2C}\frac{M_{\rm pri}}{M_{\rm sat}}\frac{1}{\ln\left(1+M_{\rm pri}/M_{\rm sat}\right)}\left(\frac{r_{\rm circ}}{r_{\rm vir}}\right)^{(1/2)}\tau_{\rm dyn,\ halo}, \tag{3.3}$$

where ϵ is the circularity of the satellite orbit, $M_{\rm pri}$ is the mass of the central galaxy and its dark matter halo, $M_{\rm sat}$ is the mass of the satellite system, $r_{\rm circ}$ is the radius of the circular orbit. The modified parameters are the constant $C = 0.43$ and $f(\epsilon) = 0.90\epsilon^{0.47} + 60$, obtained by Jiang et al. (2008).

GALFORM distinguishes between two types of merger based on the mass ratio of merging galaxies:

- **Major mergers**, which involve galaxies of comparable mass, disrupt galaxy disks and result in the formation of elliptical galaxies or spheroids. These events also trigger intense bursts of star formation, rapidly converting available gas into stars (see § 3.2.5.2).

- **Minor mergers** occur when a satellite galaxy merges with a much larger central galaxy, typically leading to less dramatic changes. Minor mergers may increase the mass of the galactic bulge, but do not trigger significant starbursts.

GALFORM explicitly models the morphological transformations resulting from these mergers, tracking the formation and growth of disks and spheroids (bulges),

along with their associated stellar populations and chemical enrichment histories over cosmic time. By coupling the merger-driven assembly of structure with the regulation of star formation in bursts and quiescent modes, GALFORM captures the diverse evolutionary pathways of galaxies.

3.2.5 Star formation in galaxy disks and starbursts

In the GALFORM framework, star formation is modelled by distinguishing two primary modes: quiescent star formation occurring in galaxy disks, and starbursts triggered by galaxy mergers or disk instabilities. In galaxy disks, star formation proceeds gradually, governed by empirical laws that relate star formation rates to the molecular gas content. In contrast, starbursts can occur on shorter timescales and can lead to temporarily elevated star formation rates, with their intensity regulated by model parameters such as the star formation efficiency and dynamical timescale. This dual-mode approach reflects observational evidence that galaxies can experience both steady and episodic star formation, depending on their morphology, environment, and evolutionary stage.

3.2.5.1 Star formation in galaxy disks

Star Formation (SF) in galaxy disks is typically modelled using empirical relations that link the surface density of star formation ($\Sigma_{\rm SF}$) to the available molecular gas surface density ($\Sigma_{\rm mol}$). GALFORM adopts an formulation inspired by the empirical star formation law by Blitz and Rosolowsky (2006):

$$\Sigma_{\rm SF} = \nu_{\rm SF}\Sigma_{\rm mol}, \tag{3.4}$$

where $\nu_{\rm SF}$ is the star formation efficiency per unit molecular gas mass. In GALFORM, $\nu_{\rm SF}$ is treated as a free parameter, calibrated against observed star formation rates in nearby galaxies, but can vary within a range suggested by the observations.

To determine the fraction of molecular gas ($f_{\rm mol}$) in galaxy disks, GALFORM utilizes a pressure-based empirical prescription, based on observations suggesting

that molecular gas formation is strongly influenced by mid-plane pressure within the galactic disk (Blitz and Rosolowsky 2006; Leroy et al. 2008). The molecular fraction is computed as:

$$f_{\rm mol} = R_{\rm mol}/(1 + R_{\rm mol}), \tag{3.5}$$

where $R_{\rm mol}$ is defined as a fraction of the molecular gas surface density and the atomic gas surface density ($\Sigma_{\rm mol}/\Sigma_{\rm atom}$).

The total Star Formation Rate (SFR) in the galactic disk is then assumed to be proportional to the mass in the molecular component only as follows:

$$\psi_{\rm disk} = \nu_{\rm SF} M_{\rm mol,\ disk} = \nu_{\rm SF} f_{\rm mol} M_{\rm cold,\ disk}. \tag{3.6}$$

The free parameter $\nu_{\rm SF}$ is chosen based on a sample of local galaxies, which Bigiel et al. (2011) find a best fitting value of $\nu_{\rm SF} = 0.43$ Gyr^{-1} with 1σ scatter of 0.24 dex. `GALFORM` allows the free parameter to be varied within the 1σ as quoted.

3.2.5.2 Star formation in starbursts

Galaxy mergers and disk instabilities can trigger episodes of bursty star formation, known as starbursts, which are typically associated with elevated star formation rates compared to the quiescent disk mode. In `GALFORM`, starbursts occur during major and minor mergers or when disk instabilities funnel cold gas into the central regions of galaxies, where it forms stars over shorter timescales. The intensity and duration of these bursts depend on model parameters, particularly those governing the burst star formation timescale and the efficiency of gas conversion into stars.

Star formation in bursts is modelled as a rapid conversion of cold gas into stars. However, the overall efficiency of this process is regulated by feedback mechanisms–such as supernovae and AGN activity, described in later sections–which can expel or reheat gas before it forms long-lived stellar remnants. The burst-mode star formation rate is expressed as:

$$\psi_{\rm burst} = \frac{M_{\rm cold,\ burst}}{\tau_{*\rm burst}} = \nu_{\rm SF,\ burst} M_{\rm cold,\ burst}, \tag{3.7}$$

where $M_{\rm cold,\ burst}$ is the cold gas mass in the starburst component, $\nu_{\rm SF,\ burst}$ is the star formation efficiency in bursts (a free parameter), and $\tau_{*\rm burst}$ is the star formation timescale.

The $\tau_{*\rm burst}$ is defined in terms of the dynamical time of the bulge component formed during the burst:

$$\tau_{*\rm burst} = \max\left[f_{\rm dyn}\tau_{\rm dyn,\ bulge}, \tau_{*\rm burst,\ min}\right], \tag{3.8}$$

where $\tau_{\rm dyn,\ bulge} = r_{\rm bulge}/V_c(r_{\rm bulge})$ is the dynamical time computed from the bulge half-mass radius and circular velocity. The parameter $f_{\rm dyn}$ sets the proportionality between star formation timescale and bulge dynamical time, while $\tau_{*\rm burst,\ min}$ sets a lower limit on the timescale. Both $f_{\rm dyn}$ and $\tau_{*\rm burst,\ min}$ are treated as free parameters in the model

This formulation ensures that for systems with long dynamical times, the burst timescale scales with $\tau_{\rm dyn,\ bulge}$, consistent with observation of starburst galaxies in the local Universe (e.g. Kennicutt 1998). However, for compact systems with short dynamical times, the star formation timescale is prevented from becoming unphysically short by the imposed floor $\tau_{*\rm burst,\ min}$.

3.2.6 Supernova and AGN feedback

Feedback processes from supernovae (SNe) and active galactic nuclei (AGN) are essential mechanisms in galaxy formation models, significantly influencing the properties of galaxies by regulating star formation rates, gas cooling, and the chemical evolution of galaxies. `GALFORM` explicitly includes these feedback mechanisms to match observational data and prevent known theoretical discrepancies, such as the over-cooling problem in massive halos.

3.2.6.1 SN feedback

Supernova (SN) feedback refers to the injection of energy and momentum from massive stars that explode as supernovae back into the Interstellar Medium (ISM). In `GALFORM` this feedback mechanism reheats cold gas from galaxy disks and can eject it either into the surrounding dark matter halo or completely beyond the halo's virial radius, depending on the halo mass and the energy available from supernova explosions.

The rate of which cold gas mass ejected ($\dot{M}_{\rm eject}$) from the galaxy due to supernova explosions is modelled to be proportional to the instantaneous star formation rate (ψ), with a mass loading factor β that follows a power-law dependence on the galaxy circular velocity V_c, expressed as:

$$\dot{M}_{\rm eject} = \beta\psi = \left(\frac{V_{\rm c}}{V_{\rm SN}}\right)^{-\gamma_{\rm SN}} \psi, \tag{3.9}$$

where $V_{\rm SN}$ and $\gamma_{\rm SN}$ are both free parameters in the model. The parameter $V_{\rm SN}$ sets the characteristic velocity scale for feedback, while $\gamma_{\rm SN}$ determines how strongly the mass loading varies with the galaxy mass. The parametrisation was suggested by early gas dynamic simulations (Navarro and White, 1993); the values of the parameters are set by forcing the model to reproduce the local luminosity function.

Gas ejected beyond the halo from the galaxy by supernova feedback is assumed to be stored in an external reservoir of mass $M_{\rm res}$. This gas is not permanently lost, but is located beyond the virial radius and is allowed to re-accrete into the halo over time. The rate at which this gas returns to the hot halo gas component, $\dot{M}_{\rm return}$, is given by as follows:

$$\dot{M}_{\rm return} = \alpha_{\rm return}\frac{M_{\rm res}}{\tau_{\rm dyn,\ halo}}, \tag{3.10}$$

where $\tau_{\rm dyn,\ halo}$ is halo dynamical time, defined as $r_{\rm vir}/V_{\rm vir}$, and $\alpha_{\rm return}$ is another free parameter that sets the efficiency of this gas recycling process.

This feedback mechanism plays a crucial role in regulating the gas content of galaxies, particularly in low-mass systems, and helps `GALFORM` match observed

galaxy stellar mass functions and star formation rates (Baugh 2006; Lacey et al. 2016).

3.2.6.2 AGN feedback

Active galactic nucleus (AGN) feedback plays a crucial role in suppressing star formation in massive galaxies and mitigating the over-cooling problem in high-mass dark matter halos (White and Frenk 1991; Benson et al. 2003; Bower et al. 2006). Without this mechanism, models tend to over-predict the number of luminous galaxies at the bright end of the luminosity function. In `GALFORM`, AGN feedback is implemented by inhibiting the cooling of hot gas in massive halos, effectively shutting down the supply of cold gas required for star formation (Bower et al. 2006; Lacey et al. 2016).

The suppression is based on two physical criteria that must be satisfied simultaneously for AGN heating to prevent gas cooling;

- (i) The cooling time of the hot halo gas at the cooling radius, $t_{\rm cool}(r_{\rm cool})$, must be longer than the free-fall time $t_{\rm tt}(r_{\rm cool})$ by a factor set by the parameter $\alpha_{\rm cool}$. Specifically,

$$t_{\rm cool}(r_{\rm cool})/t_{\rm ff}(r_{\rm cool}) > 1/\alpha_{\rm cool}, \tag{3.11}$$

  where $\alpha_{\rm cool}$ is an adjustable model parameter that determines how readily AGN feedback can switch on ($\alpha_{\rm cool}$ has a typical value of ~ 1, and this condition ensures that the gas is in a quasi-hydrostatic regime suitable for AGN heating).

- (ii) The energy output required to balance radiative cooling, $L_{\rm cool}$, must be less than a fraction $f_{\rm Edd}$ of the Eddington luminosity of the central supermassive black hole:

$$L_{\rm cool} < f_{\rm Edd}\, L_{\rm Edd}(M_{\rm BH}), \tag{3.12}$$

where $f_{\rm Edd}$ is another free parameter, and $L_{\rm Edd}(M_{\rm BH})$ is the Eddington luminosity corresponding to the black hole mass $M_{\rm BH}$. $f_{\rm Edd}$ is adopted to be 0.01 in GALFORM based on discussion in Fanidakis et al. (2011).

If both conditions are met, AGN feedback is assumed to completely suppress further gas cooling from the hot halo, effectively quenching star formation in the central galaxy. This mechanism is particularly important for reproducing the observed cut-off at the bright-end in the galaxy luminosity function and at the high-mass end in the stellar mass function.

3.2.7 Supermassive Black Hole (SMBH) Growth

Supermassive black holes are key components in the GALFORM model, as their growth is directly tied to the regulation of star formation in massive galaxies through AGN feedback, as mentioned in the previous section. The growth of SMBHs in GALFORM occurs via two main channels; accretion of cold gas during starbursts, and mergers with other black holes during galaxy coalescence events (Bower et al. 2006; Malbon et al. 2007).

The dominant growth mode is through gas accretion during starbursts, which are triggered by galaxy mergers and disk instabilities. During these events, a fraction of the available cold gas is assumed to be funnelled toward the central black hole. This is implemented in the model using a simple scaling relation in which the accreted black hole mass is taken to be a fixed fraction of the cold gas converted into stars, and can be described as:

$$\Delta M_{\rm BH} = f_{\rm BH} \Delta M_{\rm stars}, \tag{3.13}$$

where $\Delta M_{\rm BH}$ is the mass accreted onto the black hole, $\Delta M_{\rm stars}$ is the stellar mass formed during the bursts (accounting for feedback and stellar mass loss), and $f_{\rm BH}$ is the black hole accretion efficiency. This efficiency parameter is treated as a free parameter in GALFORM and it is typically to be less than 0.03, for reproducing the

observed relation between black hole mass and bulge mass in the local galaxies (see Fig. 6 in Lacey et al. (2016)).

In addition to accretion, black hole-black hole mergers contribute to SMBH growth when two galaxies merge. In such events, it is assumed that the central black holes of the progenitor galaxies coalesce efficiently and instantaneously to form a single, more massive black hole in the remnant galaxy. Although the timescales for black hole mergers in reality may be longer and involve dynamical processes such as dynamical fiction and gravitational wave emission, `GALFORM` does not explicitly model these processes and instead assumes that coalescence occurs promptly after galaxy mergers.

The final SMBH mass is assumed to be the sum of the accreted mass and the combined mass of the two progenitors. The total mass of the SMBH governs its Eddington luminosity, which is used in AGN feedback prescriptions to determine whether gas cooling can be suppressed in massive halos. As such, the evolution of SMBHs in `GALFORM` is intimately connected to the thermal state of halo gas and the quenching of star formation in high-mass galaxies.

3.2.8 Galaxy sizes

Accurately predicting galaxy sizes is a critical aspect of galaxy formation modelling. In `GALFORM`, galaxy sizes are computed by modelling the evolution and conservation of angular momentum for galaxy disks, and by applying energy conservation and the virial theorem to spheroids formed through mergers or disk instabilities (Lacey et al. 2016).

Disk sizes: Disks are assumed to form through the radiative cooling of hot halo gas, which is presumed to initially share the specific angular momentum of the dark matter halo. As the gas cools and collapses, it retains this angular momentum, settling into a rotationally supported disk. The disk radius is then calculated by solving for centrifugal equilibrium in the combined potential of the disk, spheroid,

and dark matter halo. During subsequent evolution, changes in the gravitational potential (e.g. due to stellar mass growth or halo contraction) lead to adiabatic adjustment of the disk size, but without changes in specific angular momentum (Lacey et al. 2016).

While the assumption of angular momentum conservation is a simplification, it provides a reasonable approximation and is widely adopted in semi-analytic models. However, it neglects possible losses of angular momentum due to non-axisymmetric instabilities, dynamical friction, and misalignment between gas inflow and the existing disk plane (see e.g. Lagos et al. 2015 for a generalised treatment).

Spheroid sizes: Spheroids (bulges and ellipticals) are formed in `GALFORM` through both major mergers and disk instabilities. In both cases, the half-mass radius of the resulting spheroid is calculated using energy conservation and the virial theorem. For mergers, the total energy of the system is given by the sum of the internal binding energies of the progenitor galaxies and their orbital energy prior to coalescence. The remnant size $r_{\rm remnant}$ is obtained from:

$$\frac{(M_{\rm gal,\,1}+M_{\rm gal,2})^2}{r_{\rm remnant}} = \frac{M_{\rm gal,\,1}^2}{r_{\rm gal,\,1}} + \frac{M_{\rm gal,\,2}^2}{r_{\rm gal,\,2}} + \frac{f_{\rm orbit}}{c_{\rm gal}}\frac{M_{\rm gal,\,1}M_{\rm gal,2}}{r_{\rm gal,\,1}+r_{\rm gal,\,2}}, \tag{3.14}$$

where M_i and r_i are the mass and half-mass radius of progenitor galaxies, $c_{\rm gal}$ is a structural constant (typically 0.5), and $f_{\rm orbit}$ is a free parameter that encapsulates the orbital energy contribution and is typically in the range $0 \leq f_{\rm orbit} \leq 1$ (Lacey et al. 2016).

For disk instabilities, the sizes of newly formed spheroids are computed similarly by treating the pre-existing disk and bulge as the two progenitor systems. This approach assumes that the instability event results in an immediate redistribution of mass into a new equilibrium spheroid, again based on energy conservation (Lacey et al. 2016). A fitting formula analogous to the merger case is used, including contributions from the interaction energy between the pre-existing components, described as:

$$c_{\rm bulge}\frac{(M_{\rm disk}+M_{\rm bulge})^2}{r_{\rm new}} = c_{\rm bulge}\frac{M_{\rm bulge}^2}{r_{\rm bulge}} + c_{\rm disk}\frac{M_{\rm disk}^2}{r_{\rm disk}} + f_{\rm int}\frac{M_{\rm disk}M_{\rm bulge}}{r_{\rm disk}+r_{\rm bulge}}, \tag{3.15}$$

where M_i* and r_i have the same descriptions as in Equation 3.14, the dimensionless factor $c_{\rm disk} = 0.49$ for a pure exponential disk and $c_{\rm bulge} = 0.45$ for an $r^{1/4}$ -law spheroid, and a factor $f_{\rm int} = 2.0$ representing the gravitational interaction energy of the disk and bulge.

GALFORM also accounts for the adiabatic contraction of the dark matter halo in response to the condensation of baryons into the central galaxy. This alters the halo potential and influences both disk and bulge sizes. The size calculation therefore includes baryonic self-gravity and the dynamical response of the surrounding dark matter halo (Lacey et al. 2016).

3.2.9 Chemical Enrichment

Tracking the chemical enrichment of gas and stars is a key component of galaxy formation models, as metal abundances affect gas cooling, star formation, and the interpretation of galaxy observables such as spectral energy distributions and emission lines. Although this work does not involve full spectral energy distribution (SED) modelling or investigate the impact of nebular emission lines on galaxy photometry, we include this description for completeness and to provide a coherent overview of the GALFORM framework. GALFORM models the production and distribution of heavy elements (metals) through a simplified but physically motivated framework that follows the mass and metallicity of gas and stars in different galactic components over time.

Chemical enrichment in GALFORM is driven by star formation. When stars form from cold gas in the disk or during a starburst, a fraction of their mass is returned to the interstellar medium (ISM) via stellar winds and supernova explosions. This recycled material is enriched with heavy elements synthesised in stars during their lifetimes. The model assumes instantaneous recycling, where a fixed fraction R of the stellar mass formed is immediately returned to the cold gas reservoir, enriched

*The disk and bulge masses in this formula only consider stars and cold gas.

with metals. The evolution of four different baryon components (hot gas in halos, rejected gas outside halos in the reservoir, cold gas in galaxies, and stars in galaxies) are, respectively, described as:

$$\dot{M}_{\rm hot} = -\dot{M}_{\rm acc} + \alpha_{\rm ret}\frac{M_{\rm res}}{\tau_{\rm dyn,\ halo}} \tag{3.16}$$

$$\dot{M}_{\rm cold} = \dot{M}_{\rm acc} - (1 - R - \beta)\psi \tag{3.17}$$

$$\dot{M}_{*} = (1 - R)\psi \tag{3.18}$$

$$\dot{M}_{\rm res} = \beta\psi - \alpha_{\rm ret}\frac{M_{\rm res}}{\tau_{\rm dyn,\ halo}}. \tag{3.19}$$

The mass of metals ejected into the ISM per unit stellar mass formed is characterised by the metal yield p, defined as the mass of newly formed metals produced and released by stars per unit mass locked into long-lived stellar remnants. The changes in the total mass of metals in the hot gas, in the cold gas, stars in galaxies, and the reservoir are thus, respectively, governed by;

$$\dot{M}^{Z}_{\rm hot} = -Z_{\rm hot}\dot{M}_{\rm acc}\frac{M^{Z}_{\rm res}}{\tau_{\rm dyn,\ halo}} \tag{3.20}$$

$$\dot{M}^{Z}_{\rm cold} = Z_{\rm hot}\dot{M}_{\rm acc} + [p - (1 - R + \beta)Z_{\rm cold}]\psi \tag{3.21}$$

$$\dot{M}^{Z}_{*} = (1 - R)Z_{\rm cold}\psi \tag{3.22}$$

$$\dot{M}^{Z}_{\rm res} = \beta Z_{\rm cold}\psi - \alpha_{\rm ret}\frac{M^{Z}_{\rm cold}}{\tau_{\rm dyn,\ halo}}, \tag{3.23}$$

where M_i^Z is the metal mass of each component and Z_i is defined as M_i^Z/M_i.

While the model adopts a simplified treatment assuming instantaneous recycling and constant yields (e.g. R and p^*), these prescriptions are sufficient to reproduce key observational trends in galaxy metallicities when combined with self-regulated star formation and feedback (see Fig. 22 and Fig. 23 of Lacey et al. 2016).

*These values depend on the IMF used.

3.2.10 Stellar initial mass function

The initial mass function (IMF) is a fundamental ingredient in galaxy formation models, describing the stellar masses formed in a given star formation event. The IMF influences multiple aspects of galaxy evolution, including in principle the rate supernova feedback, metal production, stellar mass-to-light ratios, and SED of galaxies.

The IMF is typically represented as a power law or a piecewise power law in stellar mass, expressed as:

$$\Phi(m) = \frac{dN}{d\ln m} \propto m^{-x}, \tag{3.24}$$

where m is stellar mass, N is the number of stars formed with the mass range $m, m + dm$, and x is the slope of the power law.

In `GALFORM` the IMF is treated as a model assumption and can be specified independently for different star formation mode. In the version of Lacey et al. (2016), `GALFORM` adopts two IMF prescriptions:

- Kennicutt (1983) IMF, which has $x = 0.4$ for $m < 1\ M_\odot$ and $x = 1.5$ for $m > 1\ M_\odot$, is used for the quiescent star formation in galaxy disks.
- A single power law with slope $x = 1$, referred to as a top-heavy IMF, is used in starbursts. The two IMF model was introduced by Baugh et al. (2005) to reproduce submillimetre galaxy counts and the local galaxy luminosity function.

Because the IMF shapes the mass returned to the interstellar medium and the yield of metals, it has downstream impacts on both the star formation history and the chemical enrichment of galaxies. Assumptions about the IMF must therefore be carefully considered when comparing model predictions with observed stellar masses, luminosities, and metallicities.

3.3 Stellar Population Synthesis

In order to predict the photometric properties of galaxies, GALFORM models the integrated light from stellar populations of varying ages and metallicities. While the model does not explicitly compute full SEDs, it uses pre-computed Stellar Population Synthesis (SPS) models to derive galaxy luminosities in a chosen set of filters. This approach enables comparisons with observational data while preserving computational and memory efficiency.

The SED at time t contains stars of different ages and metallicities, depending on the star formation history, and can be calculated as

$$L_\lambda(t) = \int_0^t dt' \int_0^\infty dZ' \Psi(t', Z') L_\lambda^{(SSP)}(t - t', Z'; \Phi), \tag{3.25}$$

where $\Psi(t', Z')dt'dZ'$ is the mass of stars which formed between t' and $t' + dt'$ and have metallicity in the range between Z' and $Z' + dZ'$, and $L_\lambda^{(\mathrm{SSP})}(t, Z; \Phi)$ is the SED of a Simple Stellar Population (SSP) of a unit mass with age t, metallicity Z, and an IMF $\Phi(m)$. $\Psi(t, Z)$ is calculated by summing over the star formation histories of all progenitor galaxies which merged to form the final galaxy. $L_\lambda^{(\mathrm{SSP})}$ is calculated from the luminosity of a single star $L_\lambda^{(\mathrm{star})}(t, Z, m)$ and it is expressed by

$$L_\lambda^{(SSP)}(t, Z; \Phi) = \int_{m_{\mathrm{lo}}}^{m_{\mathrm{up}}} L_\lambda^{(\mathrm{star})}(t, Z; \Phi)\Phi(m) d\ln m, \tag{3.26}$$

where m_{lo} and m_{up} are the lower- and upper- mass cutoff for the mass distribution of the IMF, typically around 0.1 - 100 $M_\odot$, respectively. The relation between $L_\lambda(\lambda)$ and $L_\nu(\nu)$ is described by $\lambda L_\lambda(\lambda) = \nu L_\nu(\nu)$.

Before reaching the observer, starlight is attenuated by dust through absorption and re-emission. Radiation in the UV, optical, and near-infrared (IR) is absorbed by dust and subsequently re-emitted at IR and sub-mm wavelengths. GALFORM incorporates a self-consistent model for dust absorption and emission. It assumes a

two-phase dust medium: dense clouds and a diffuse component. Stars are assumed to be embedded in the dust. The results of radiative transfer calculation are used to compute the attenuation by the dust. For a full description of the dust modelling implemented in `GALFORM`, the reader is referred to Section 3.9.2 of (Lacey et al., 2016).

The absolute magnitude M_Q is defined as the apparent magnitude that an object would have if it were 10 pc away. It is related to the $L_\nu(\nu)$ as

$$M_{\mathrm{AB},Q} = -2.5 \log_{10} \left[\frac{\int \frac{d\nu_e}{\nu_e} \frac{L_\nu(\nu_e)}{4\pi(10pc)^2} Q(\nu_e)}{\int \frac{d\nu_e}{\nu_e} g_\nu^Q(\nu_e) Q(\nu_e)} \right], \tag{3.27}$$

where $Q(\nu)$ is the transmission curve for filter Q and g_ν is the AB reference flux per unit frequency (Oke and Gunn 1983; Hogg et al. 2002).

To illustrate the predictions of `GALFORM` relevant to this thesis, Fig. 3.2 shows the evolution of the i-band galaxy luminosity function across redshifts from $z = 0.00$ to $z = 2.00$, sampled directly from `GALFORM` snapshot outputs. The redshift range matches that explored in the observational analysis presented in later chapters.

Furthermore, Fig. 3.3 presents the evolution of the i-band LF separately for red and blue galaxy populations, defined using rest-frame $(g-r)$ colours. Red galaxies are selected with $(g-r)_{\mathrm{rest}} \geq 0.4$, while blue galaxies have $(g-r)_{\mathrm{rest}} < 0.4$.

3.4 Limitations of DM-only simulations and baryonic effects on halo mass

The implementation of `GALFORM` relies on merger trees extracted from cosmological N-body simulations that model the evolution of dark matter only. These simulations are computationally efficient and well-suited for generating large-volume halo catalogues, but they neglect the effects of baryonic physics on the growth and structure of halos.

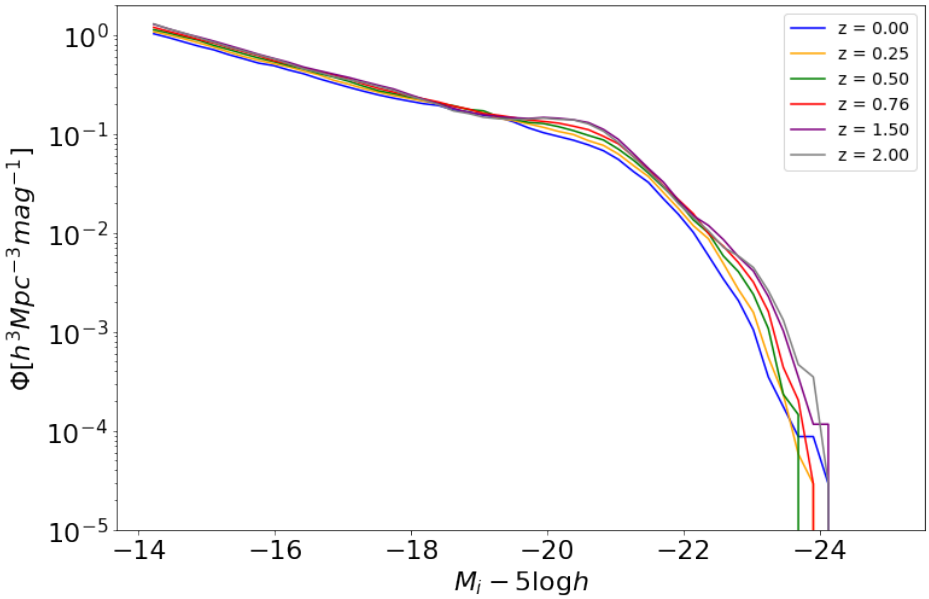


Figure 3.2: The evolution of the i-band luminosity function from `GALFORM` snapshots between redshift $z = 0$ and $z = 2$.

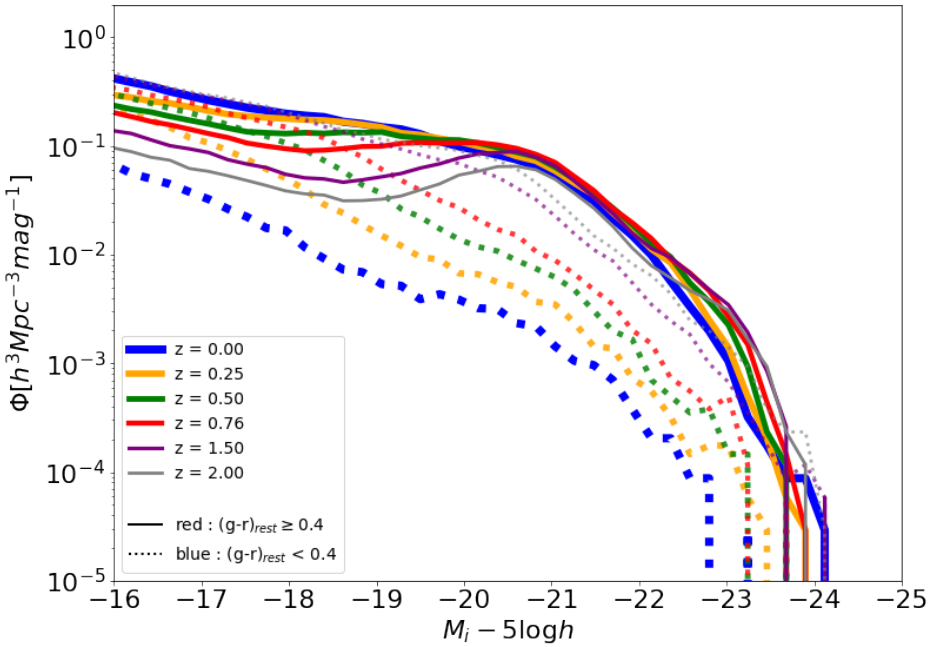


Figure 3.3: The evolution of the i-band luminosity function for red and blue galaxies from `GALFORM` snapshots between redshift $z = 0$ and $z = 2$.

In reality, baryonic processes such as gas cooling, star formation, and feedback from supernovae and AGN can alter the mass distribution within halos. For example, radiative cooling and baryon condensation can deepen the central potential

wells, increasing the total mass of halos (a process often referred to as *adiabatic contraction*). Conversely, energetic outflows from feedback processes can expel gas and cause halos to expand slightly, reducing their effective mass and concentration. Hydrodynamical simulations that include these effects have shown that baryons can lead to systematic shifts in halos masses and profiles compared to their dark matter-only counterparts (e.g. Schaller et al. 2015).

Here, GALFORM does not apply any correction to account for such baryonic effects on halo mass. The merger histories used by GALFORM are constructed directly from DM-only simulations using FoF halo finders and subhalo identification algorithm SUBFIND, as mentioned in §3.2.2. As a result, the galaxy formation processes modelled in GALFORM evolve within a framework where halo masses reflect purely gravitational collapse in the absence of baryons.

It is important to note that this approximation may introduce subtle differences in the predicted assembly histories of galaxies, particularly at small scales or high redshift. Since baryonic effects can shift halo masses and merger rates across cosmic time, they may influence the timing of gas accretion and galaxy mergers. However, this approach remains widely adopted in semi-analytic modelling due to its flexibility and computational efficiency.

3.5 Observation tests of GALFORM

A critical strength of the GALFORM semi-analytic model is its ability to produce detailed predictions that can be directly compared against observational data, providing a robust way to test and refine galaxy formation theories. Such comparisons play a pivotal role in constraining the theoretical assumptions, physical prescriptions, and free parameters within the model.

GALFORM has been extensively tested against a broad range of observational constraints, spanning various cosmic epochs and wavelengths. The observational data used for constraining GALFORM includes: i) the optical and near-IR LFs at

$z = 0$, ii) the HI mass function at $z = 0$, iii) the morphological fractions at $z = 0$, iv) the black hole-budge mass relation at $z = 0$, v) the evolution of near-IR LF between $z = 0-3$, vi) the sub-mm galaxy number counts and redshift distributions, vii) the far-IR number counts, and far-UV FS and Lyman-break galaxies (Cole et al. 2000; Bower et al. 2006; Gonzalez-Perez et al. 2014; Lacey et al. 2016; Baugh et al. 2019).

After calibration, GALFORM provides several observation predictions based on model outputs (see Fig. 3.1). In this thesis, I focus on two key tests of GALFORM: the galaxy luminosity function and the stellar mass function, as these are directly relevant to the analyses presented in the subsequent chapters.

- **Galaxy Luminosity Functions (LFs)**: GALFORM predicts galaxy luminosity functions across a range of wavelengths, from optical to infrared and sub-millimetre. The shape and evolution of the LF provide stringent constraints on star formation histories, feedback processes, and IMF assumptions. The original model by Cole et al. (2000) showed good agreement with the local optical LF. Later refinements, including the incorporation of AGN feedback (Bower et al. 2006) and the implementation of a top-heavy IMF in starbursts (Lacey et al. 2016), significantly improved the model's performance at the bright end and at the higher redshifts, where earlier versions underpredicted luminous galaxies (e.g. see Fig. 17 in Lacey et al. 2016).

- **Stellar Mass Functions (Stellar Mass Function (SMF)s)**: The stellar mass function is another fundamental observable that constrains the integrated star formation history of galaxies. GALFORM predictions of the SMF across redshifts have been compared against observationally inferred stellar mass functions, providing critical constraints on parameters related to star formation efficiency, supernova feedback, and AGN heating. For instance, the inclusion of AGN feedback was essential to reproducing the break in the

SMF at the high-mass end by preventing excessive stellar mass buildup in massive halos (Lacey et al. 2016).

3.6 Conclusion

This chapter has outlined the fundamental astrophysical processes incorporated in the GALFORM semi-analytic model of galaxy formation. The model builds on the hierarchical growth of structure in cold dark matter cosmologies and implements a suite of physically motivated prescriptions to simulate gas cooling, star formation in disks and starbursts, feedback from SNe and AGN, galaxy mergers, black hole growth, and chemical enrichment. It also includes treatments of galaxy sizes, stellar populations, and photometric properties through stellar population synthesis modelling.

GALFORM successfully reproduces many observed properties of galaxies, including the luminosity and stellar mass functions. These comparisons serve as critical observational tests and demonstrate the model's predictive power.

In the next chapters, this model will be applied to generate lightcone mock catalogues and to interpret measurements of the galaxy luminosity function and stellar mass function from the PAUS survey data.

Chapter 4

Building the Lightcone Mock Catalogue

4.1 Overview

This chapter describes the construction of a mock galaxy catalogue by Manzoni et al. (2024) to emulate observations from the Physics of the Accelerating Universe Survey (PAUS), using the methods described in Merson et al. (2013) and Stothert et al. (2018). The material is presented here for completeness in the thesis, with some of my own analysis added. The mock is based on the `GALFORM` semi-analytic model of galaxy formation, applied to the Planck Millennium (PMILL) N-body simulation (Baugh et al., 2019). Galaxies are extracted from simulation snapshots and placed within an observer's past lightcone using interpolation techniques, enabling a realistic representation of a galaxy population across cosmic time.

The primary goal of this lightcone mock is to reproduce the key observational features of the PAUS dataset, including survey geometry, photometric selection, and photometric redshift performance. The mock provides a test for validating measurements of the galaxy luminosity function, photometric redshift performance, and other statistical analyses.

The chapter is structured as follows: §4.2 describes the construction of the

lightcone, including the choice of N-body simulation, tiling of simulation boxes, and the interpolation of halo positions between snapshots. §4.3 presents how intrinsic galaxy properties are assigned in the lightcone, including the treatment of rest-frame magnitudes and physical properties. §4.4 describes the computation of observer-frame photometry. §4.5 outlines the angular and flux-based survey selection criteria used to mimic the PAUS observational footprint and depth. §4.6 discusses the inclusion of photometric and photometric redshift uncertainties using noise models and empirical redshift error distributions. §4.7 presents a validation of the mock against fundamental PAUS statistic, including redshift distributions, number counts, and colour-redshift relations. Finally, §4.8 summarises the key components and outcomes of the mock construction.

This chapter provides the basis of subsequent scientific analyses in this thesis, including luminosity function estimation and rest-frame properties inference.

4.2 Lightcone Construction

4.2.1 Choice of N-body simulation

For our PAUS lightcone mock catalogue we use the lightcone mock catalogue built by Manzoni et al. (2024) which implements the `GALFORM` semi-analytical model into the high-resolution cosmological N-body simulation, the Planck Millennium run (PMILL; Baugh et al. 2019). Manzoni et al. used the Lacey et al. (2016) version of `GALFORM`. The PMILL simulation follows $5040^3 \simeq 1.28 \times 10^{11}$ dark matter particles in a comoving box of side 542.16 h^{-1}Mpc, with a particle mass of 1.06×10^8 $h^{-1}\mathrm{M}_\odot$. The high mass resolution of PMILL–an order of magnitude better than earlier Millennium-class runs–allows us to resolve halos hosting galaxies down to the faint limits probed by PAUS over a wide range in redshift, and to include better resolved merger histories to allow more accurate predictions of galaxy properties.

The PMILL simulation outputs 271 snapshots (compared to the 60 typically

available) between $z = 127$ and $z = 0$, equally spaced at an average interval $\Delta a \sim 0.0037$ in scale factor. Such fine time sampling is crucial to locate halos on the observer's past lightcone with minimal interpolation error in both position and velocity (Stothert et al. 2018).

4.2.2 Defining the observer's past lightcone geometry

Because a single box of length $L_{\rm box} = 542.16\ h^{-1}$Mpc only reaches out to $z \sim 0.19$, the simulation boxes are tiled using periodic boundary replications to cover the full PAUS redshift range (up to $z_{\rm max} = 2.0$). The number of box replicas, $n_{\rm rep}$ per axis required to cover a sufficient volume to contain the survey is given by

$$n_{\rm rep} = \left\lfloor \frac{r(z_{\rm max})}{L_{\rm box}} \right\rfloor + 1, \tag{4.1}$$

where $r(z_{\rm max})$ is the comoving radial distance at the maximum redshift of the survey we want to build the mock and $\lfloor x \rfloor$ means that the value of x is rounded down to the closest integer. In total, there is a grid of $(n_{\rm rep} + 1)^3$ box copies. This number of replicated boxes is typical of what is required when constructing a full-sky mock catalogue. However, this value can be reduced for smaller solid angles.

In the case of non full-sky mock, the line of sight direction chosen for the centre of the survey should not be chosen to align with one of the axes of the box arrangement to avoid repeating patterns (Merson et al. 2013).

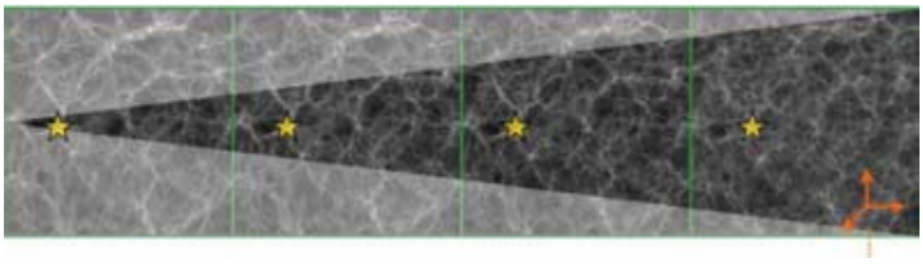


Figure 4.1: The tiling process (with the repetition emphasized by the yellow star). Image taken from Blaizot et al. (2005).

4.2.3 Interpolating halo positions between snapshots

Due to the discrete nature of PMILL simulation, halo catalogues are available only at specific simulation snapshots as described earlier in the previous section. Halos, however, do not necessarily enter the observer's past lightcone exactly at these snapshot epochs. To achieve precise positioning of halos along the observer's past lightcone, it is necessary to interpolate their positions and velocities between simulation snapshots. Here, we follow closely the interpolation method detailed in Merson et al. (2013).

The interpolation process consists of several steps:

1. **Identification of Bounding Snapshots**: We first identify the pair of consecutive simulation snapshots between which a halo crosses the observer's past lightcone. This is done by checking the halo's position relative to the observer at successive snapshots and determining between which two snapshots the crossing occurs.

2. **Interpolating Halo Positions and Velocities**: Having identified the relevant snapshots, we apply a cubic polynomial interpolation independently to each Cartesian coordinate of the halo's position. The interpolation uses both the positions and velocities of the halo at the two snapshots. This approach, following Merson et al. (2013), ensures a smooth and realistic trajectory for the halo between snapshots and allows an accurate estimate of the halo's position and velocity at the exact epoch of crossing the observer's past lightcone.

3. **Handling Satellite Galaxies**: Satellite galaxies within halos have their own orbital motions, meaning their positions do not follow that of their host halo centre in a simple way. To accurately capture satellite trajectories, a simplified two-dimensional polar interpolation around the halo centre is employed, separately interpolating the radial and angular components of the satellite's orbit. Fig. 2 of Merson et al. (2013) shows that this approach

retains realistic orbital motion and avoids artificial distortions or clustering artefacts. It is worth noting that Smith et al. (2022) reported unrealistically large velocities being predicted using a cubic interpolation, leading them to use a linear interpolation scheme in their DESI lightcone mock catalogue. For consistency with previous work of Manzoni et al. (2024), we use Merson et al.'s interpolation scheme for our lightcone mock catalogue.

The linear interpolation technique described above naturally results in a continuous distribution of halos along the observer's past lightcone. By accurately placing each halo at the precise redshift at which its light would reach us today, this method effectively assembles continuous redshift shells without discrete jumps or gaps between snapshots. Such continuity ensures a realistic spatial distribution of galaxies and preserves the accuracy of clustering measurements and photometric redshift distributions required for PAUS analyses.

4.3 Intrinsic Galaxy Properties in the Lightcone

4.3.1 Physical properties

Once galaxies are positioned on the observer's past lightcone, they must be assigned appropriate intrinsic properties–such as stellar mass, star formation rate (SFR), and metallicity. In principle, one could attempt to interpolate these properties between snapshots, similar to the approach used for halo positions. However, galaxy properties often evolve non-linearly due to physical processes like starbursts, mergers, and AGN feedback, which can occur on timescales shorter than the snapshot intervals (see Lacey et al. 2016 for a discussion of the additional time steps or substeps used between the simulation snapshots to improve the time resolution with which some processes are modelled, such as star formation).

Following the approach outlined in Merson et al. (2013), we adopt a pragmatic solution that avoids potentially misleading interpolation. For each galaxy placed

on the lightcone at redshift z, we assign it the galaxy properties from the GALFORM snapshot immediately preceding that redshift. That is, we use the output from the higher-redshift snapshot whose epoch lies just before the galaxy's lightcone crossing time.

This approach ensures that we do not attempt to interpolate through periods of rapid or complex evolution, such as major mergers, which can significantly alter a galaxy's mass and the star formation history between snapshots. Although this introduces a stepwise approximation in the redshift evolution of galaxy properties, it is a conservative method that avoids introducing artificial structure or double-counting progenitors during merger events.

The resulting lightcone catalogue thus contains galaxies with positions interpolated smoothly between snapshots, while their physical properties are fixed at the nearest earlier snapshot. Otherwise, it would require the full calculation using GALFORM for each galaxy at the epoch at which it crosses the past lightcone. Hence, this method provides a practical balance between accuracy and computational efficiency.

4.3.2 Rest-frame SEDs and magnitudes

The one exception to the treatment described above for galaxy properties on the lightcone is galaxy magnitudes. For historical reasons that were relevant when GALFORM was devised 30 years ago, the model does not store a spectral energy distribution for each galaxy. Instead, the spectra read in from the population synthesis model are converted into mass-to-light ratios for a set of filters specified prior to run time. These values are combined to return the mass-to-light ratio that would have been obtained from a composite spectral energy distribution, combining all of the episodes of star formation, taking into account the look-back times and metallicity of the star formation. Note that this emission is purely stellar emission and does not contain the contribution from nebular emission lines, which are calculated

separately (Baugh et al., 2022).

These filters can be in the rest or observed frame of the galaxy, which means that we know the exact k-correction for each model galaxy at the simulation snapshots, a fact that we exploit below to build a model to estimate the k-correction.

4.4 Computation of Observer-Frame Photometry

The observed flux of a galaxy at redshift z is calculated by first deriving its luminosity distance, $d_L(z)$, based on its comoving position in a flat Universe:

$$d_L(z) = r_c(z)(1+z), \tag{4.2}$$

where $r_c(z)$ is the comoving radial distance. The emitted luminosity per unit frequency, $L_\nu(\nu_e)$ is related to the observed flux per unit frequency $S_\nu(\nu_o)$ via:

$$S_\nu(\nu_o) = (1+z)\frac{L_\nu(\nu_e)}{4\pi d_L^2(z)}, \tag{4.3}$$

with the relation between emitted and observed frequency being $\nu_e = \nu_o(1+z)$, due to redshifting.

The observer-frame apparent magnitude in a given bandpass R is then computed using the standard AB magnitude definition:

$$m_{\mathrm{AB},R} = -2.5\log_{10}\left[\frac{\int \frac{d\nu_o}{\nu_o} S_\nu(\nu_o)R(\nu_o)}{S_{\nu,o}\int \frac{d\nu_o}{\nu_o}R(\nu_o)}\right], \tag{4.4}$$

where $R(\nu_o)$ is the transmission curve of the filter R and S_{ν_o} is the AB reference flux per unit frequency.

The observer-frame absolute magnitude of a galaxy can be calculated using

$$M_{\mathrm{AB},R} = -2.5\log_{10}\left[\frac{\int \frac{d\nu_e}{\nu_e} L_\nu(\nu_e)R(\frac{\nu_e}{1+z})}{4\pi(10\mathrm{pc})^2 S_{\nu_o}\int \frac{d\nu_o}{\nu_o}R(\frac{\nu_e}{1+z})}\right]. \tag{4.5}$$

We now can calculate the observer-frame apparent magnitude of a galaxy using

$$m_{\mathrm{AB,\ R}} = M_{\mathrm{AB},R} + 5\log_{10}\left(\frac{d_L(z)}{\mathrm{Mpc/h}}\right) + 25 - 2.5\log_{10}(1+z). \tag{4.6}$$

4.5 Survey Selection

To produce a mock catalogue that accurately represents the PAUS observations, we apply a set of angular and flux-based selection criteria to the galaxies in the lightcone. These selections mimic the survey geometry, photometric depth, and quality cuts used in the actual PAUS data reduction pipeline, and are crucial for enabling a fair comparison between the mock and the real data.

- **Angular Mask and Survey Footprint**: The central line-of-sight vector of the observer's lightcone is looking down the $\hat{Z}'(X, Y, Z)$ axis as shown in Fig. 4.2. The field of view of the lightcone is governed by the angle θ'_r. Galaxies with position vector $\vec{r}'(X', Y', Z')$ resulting in $\theta' > \theta'_r$ are excluded from the observer's lightcone. The solid angle or survey area, Ω, is related to the opening angle, θ'_r, as

$$\Omega = 2\pi[1 - \cos(\theta'_r)]. \tag{4.7}$$

  Table 2.1 shows that the PAUS as made up from the W1, W3, and G09 fields covers $\Omega_{\rm total} \sim 50$ deg^2 of the sky in total. The lightcone mock built by Manzoni et al. 2024, however, is designed to cover 100 deg^2 (equivalent to $\theta_{\rm lightcone} \sim 5.66$ deg).

  Fig. 4.3 shows the side view of the lightcone mock. This lightcone is big enough to cover the solid angle of the PAUS survey area. One can reduce the field-of-view of the mock to match any individual field in post processing. Fig. 4.4 shows the cross section of the lightcone at different redshift ranges, representing the galaxy positions as seen on the sky.

- **Flux limit selection**: As mentioned in §2, PAUS is a flux limited survey. To match the depth of lightcone mock with that of PAUS, we set the magnitude limit to have a limiting depth of $i_{\rm AB} = 23$. Manzoni et al. (2024) shows that

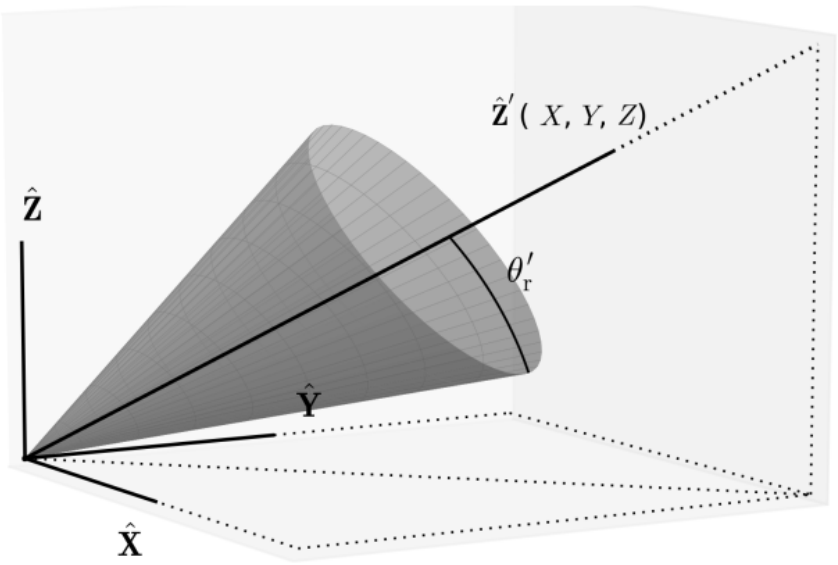


Figure 4.2: Illustration of lightcone geometry showing the central light-of-sight vector alignment $\hat{Z}'$ and the field-of-view angle θ'_r. Image taken from Merson et al. (2013).

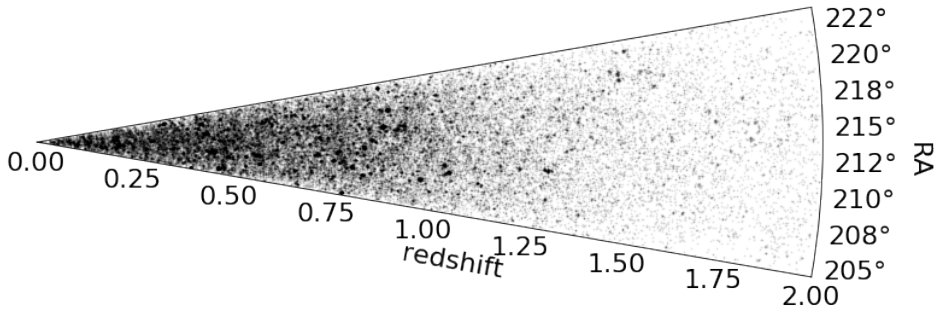


Figure 4.3: RA vs spectroscopic redshift for galaxies in the GALFORM lightcone mock catalogue. Each galaxy is plotted with the same symbol size.

the redshift distribution of the lightcone matches that of PAUS, as shown in Fig. 4.7.

4.6 Flux Errors and Photometric Redshift Errors

We also consider the effect of errors in the galaxy photometry and photometric redshift estimation on the recovered luminosity function, by perturbing these quantities in the Manzoni et al. mock catalogue.

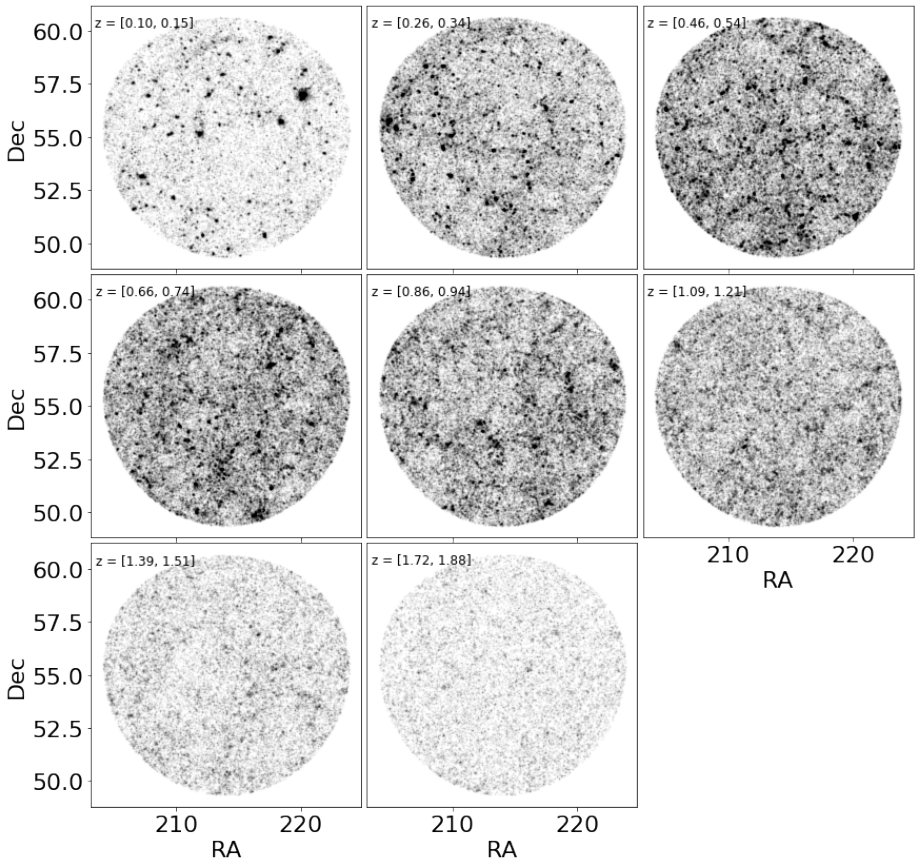


Figure 4.4: The projected 2D positions of lightcone mock galaxies at different redshift bins, shown at the top left of each panel. Each panel only show RA and DEC positions of galaxies and each galaxy has the same symbol size.

4.6.1 Flux errors

Photometry errors are assumed to have a Gaussian distribution in magnitude. The perturbed magnitude in the band labelled by j, m_j^{obs} is obtained by adding a Gaussian distributed quantity, x, which has zero mean and a variance of σ_j to the true magnitude predicted by `GALFORM`, m_j^{true}:

$$m_j^{obs} = m_j^{true} + x. \tag{4.8}$$

The variance of Gaussian is related to the signal-to-noise ratio in band j, $(S/N)_j$ by:

$$\sigma_j^2 = \frac{2.5}{\ln 10} \frac{1}{(S/N)_j}. \tag{4.9}$$

The signal-to-noise ratio is set to be 5 at the magnitude limit in a given band. The broad- and narrow-band magnitude limits for the PAUS mock are given in Manzoni et al. (2024). This model assumes that all galaxies are treated as point sources. To ensure completeness at $i_{\rm AB} = 23$ after applying photometric uncertainties, we start from a deeper sample limited at $i_{\rm AB} = 24$.

4.6.2 Photometric redshift errors

The photometric redshift code `BCNZ2` (Eriksen et al., 2019) was run using a random sample of 44,725 galaxies from the Manzoni et al. catalogue, to reduce the computational overhead. This exercise gives the distribution of estimated photometric redshifts as a function of the true redshift, including outliers. We sample this distribution, in narrow, running bins of the true redshift, to estimate the photometric redshift for all galaxies in the mock. Using this approach, we have tested that we can recover the scatter and outlier fraction obtained when running the photometric redshift code directly as shown in Fig. 4.5. Note that in this exercise we are not using the emission lines predicted by `GALFORM` (see Baugh et al. 2022), so the performance of the photometric redshift estimator is somewhat poorer than it would be for the real data, thus giving a conservative estimate of the errors.

To apply similar photometric redshift errors to the rest of the lightcone mock galaxies, we use the acceptance-rejection sampling method to reproduce the photometric redshift error distribution function. We first separate the 44,725 galaxies into 200 bins, with each bin containing an equal number of galaxies. The cumulative distribution of the photometric redshift errors of each bin is shown in Fig. 4.6. The photometric redshift error is then selected randomly using the acceptance-rejection technique based on the Cumulative Distribution Function (CDF) of the photometric redshift errors of the redshift bin that the galaxy falls into.

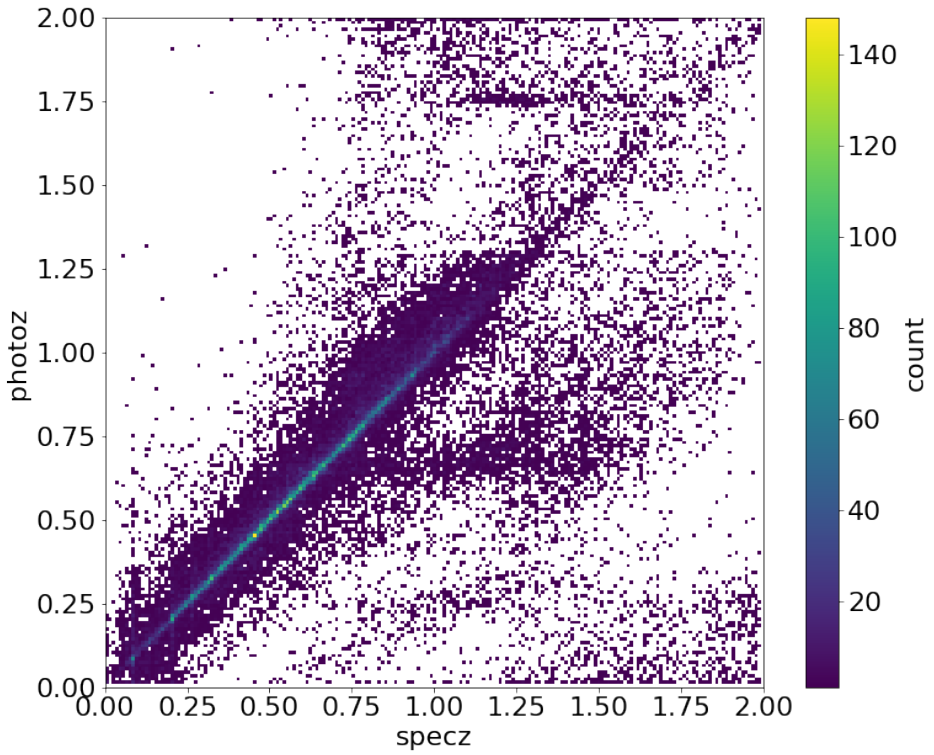


Figure 4.5: The photometric redshifts obtained using the `BCNz` code applied to a subsample of the lightcone mock compared to the true redshifts from the mock.

Our approach of sampling the error distribution rather than trying to model it using e.g. a Gaussian, is vindicated by the errors above $z \sim 1$. At these redshifts, the fraction of outliers increases substantially, to the extent that it is hard to define the scatter using the central 68 percent of galaxies (because an increasing fraction of the central 68 percentile range of estimated redshifts corresponds to outliers). In these high redshift bins, the photometric redshift errors can lead to a distortion in the number of galaxies in the redshift bins used to estimate the luminosity function.

4.7 Validation Against PAUS Statistics

The validation of the lightcone mock catalogue is essential to demonstrate that it reproduces key statistical properties of the observed galaxy populations in the PAUS. In this section, we follow the methodology and results presented in Manzoni et al. (2024), who carried out a detailed comparison between the lightcone mock

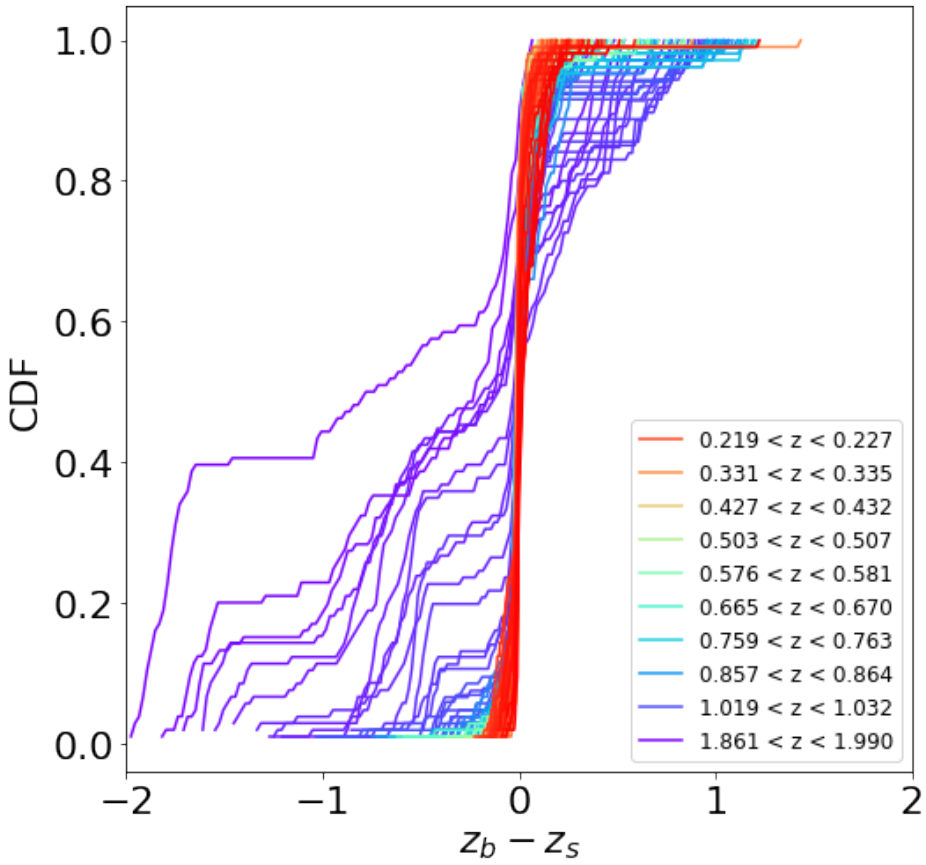


Figure 4.6: The cumulative distribution diagram of the photometric errors obtained from the `BCNz` code on the 44,725 mock galaxies. The x-axis shows the errors between the photometric redshift z_b and the spectroscopic redshift z_s.

and PAUS data. Their work serves as the foundation for evaluating the consistency between the simulated and observed galaxy distributions in redshift, colour, and photometric redshift performance.

4.7.1 Redshift distribution and number counts

Manzoni et al. (2024) compared the redshift distribution, $N(z)$, and magnitude number counts of mock galaxies to those observed in PAUS, applying the same flux limit. They showed that the mock reproduces the overall shape and amplitude of the $N(z)$ distribution across W1 and W3 fields. Similarly, the i-band number counts in the mock agreed well with the observed counts.

In addition to the results from Manzoni et al. (2024), we extend the redshift

distribution out to $z = 2$ and apply the apparent magnitude limit to $i_{\rm AB} = 23$ (compared to $i_{\rm AB} = 22.5$ in their work). Fig. 4.7 shows the redshift distribution of the lightcone galaxies compared to that of the data from PAUS W1 and W3 fields. We also show the effect of the photometric redshift uncertainties on the redshift distribution by applying a similar photo-z error similar to result obtained from `BCNz` code.

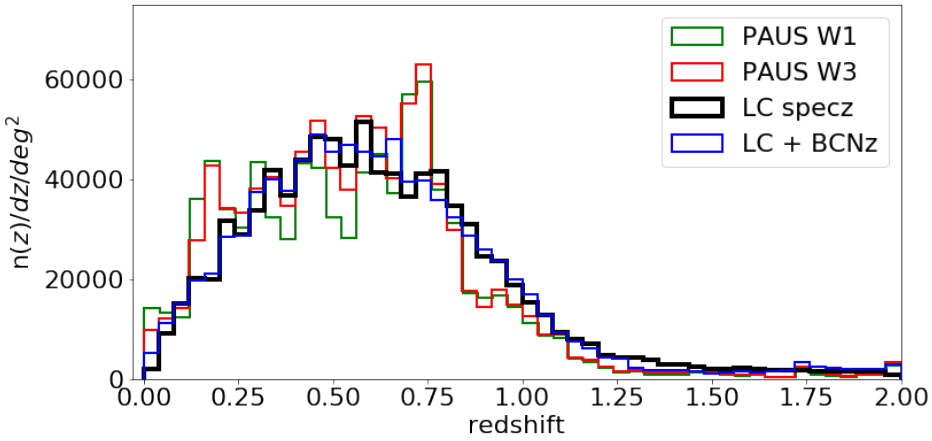


Figure 4.7: The redshift distribution from the lightcone mock catalogue compared to that from PAUS W1 and W3 data. For the mock we show the distribution with the true cosmological redshifts (black line) and also including the typical photometric redshift errors (blue curve.)

Fig. 4.8 shows the i-band galaxy number counts of the lightcone mock compared to the observational data from PAUS W1 and W3. In addition to the galaxy number counts in PAUS data, we also show the number counts of all PAUS objects, PAUS galaxies with the best 50% photometric redshift quality, and object classified as stars, labelled using suffix as "all", "best50%", and "stars", respectively.

These comparisons demonstrate that the lightcone construction and survey selection applied to the mock provide a faithful representation of the PAUS galaxy sample, at least at the level of global distributions. This work will carry out a further validation between the mock and the observed data by comparing the luminosity functions.

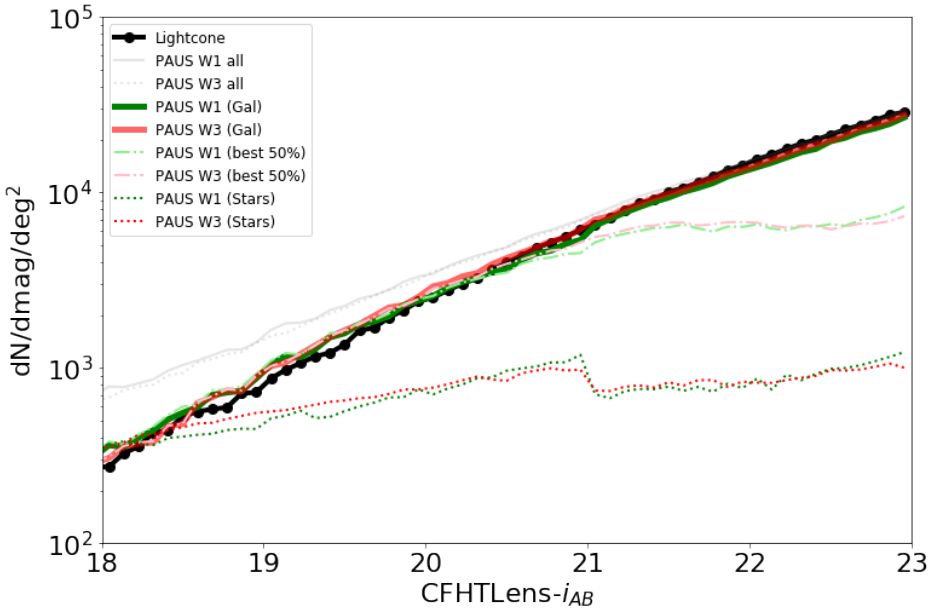


Figure 4.8: The i-band number counts of the `GALFORM` lightcone mock compared to that of the observational data from PAUS W1 and W3 fields with different cuts. The dashed lines show the galaxy counts after rejecting the 50 per cent of objects which have the poorest quality photometric redshifts; an increasing fraction of galaxies have poor photometric redshifts moving faint wards. The dotted lines show the counts of objects classified as stars; the methodology used to do this was changed by the CFHTLS team at $i_{\rm AB} = 21$.

4.7.2 Colour-redshift relations

A another key validation test performed by Manzoni et al. (2024) involved examining the colour–redshift evolution in the mock and comparing it with the PAUS data. They analysed broad-band $(g - r)$ colour, including the separation of red and blue populations.

The results showed a good agreement between mock and data in both the median colours and the spread of colour distribution as a function of redshift after considering survey uncertainties. The redshift evolution of galaxy colours is well reproduced, confirming that the model captures the basic trends in galaxy properties over the considered cosmic time.

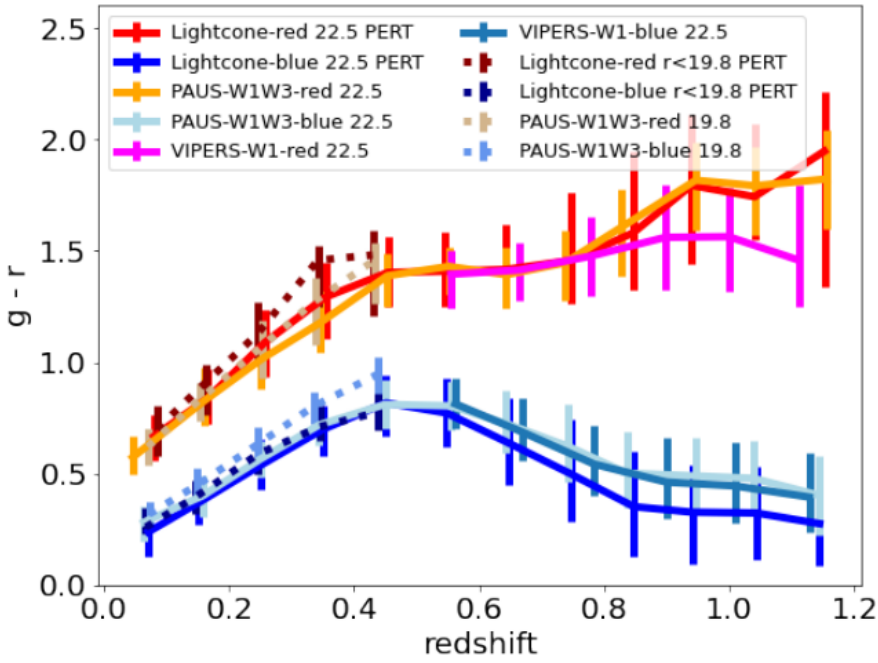


Figure 4.9: The medians of the observed (g-r) colour as a function of redshift for red and blue populations in the lightcone mock and observational data. Figure taken from Manzoni et al. (2024).

4.8 Conclusions

In this chapter, we have described the construction of a lightcone mock catalogue tailored to the Physics of the Accelerating Universe Survey (PAUS) by Manzoni et al. (2024), using galaxies generated from the `GALFORM` semi-analytic model and positioned within the observer's past lightcone constructed from the Planck Millennium N-body simulation. This mock serves as a synthetic counterpart to the real PAUS data, enabling us to model survey selection effects, test analysis pipelines, and interpret observational measurements.

We began by outlining the process of constructing the lightcone, including the replication and tiling of the simulation volume to cover the full PAUS redshift range and sky area. Halo positions were interpolated between snapshots to precisely determine their locations on the lightcone, ensuring a continuous spatial distribution of galaxies and accurate clustering properties.

Galaxy properties such as stellar mass, star formation rate, and metallicity were assigned from the `GALFORM` snapshot immediately preceding the lightcone crossing, avoiding unreliable interpolation across rapidly evolving periods. Rest-frame SEDs were constructed from these properties using mass-to-light ratios generated internally by `GALFORM`, allowing us to compute observer-frame magnitudes through redshifting and filter convolution.

Photometric and redshift errors were applied to mimic realistic observational conditions. We modelled photometric uncertainties with Gaussian perturbations and introduced photometric redshift scatter using a data-driven sampling approach based on the `BCNz2` estimator. This allowed us to incorporate realistic outliers and redshift-dependent error distributions.

Finally, we validated the mock against key PAUS observables, including redshift distributions, galaxy number counts, and colour–redshift trends. These comparisons, largely based on the work of Manzoni et al. (2024), showed good agreements with the PAUS data, supporting the utility of the mock for scientific analyses. The mock reproduces both the global statistics and redshift evolution of galaxy colours and magnitudes, forming a reliable foundation for further work, such as luminosity function estimation and environmental studies.

Chapter 5

The Use of Machine Learning for Predicting the Rest-Frame Absolute Magnitudes of Galaxies

Machine Learning (ML) techniques have become powerful tools in modern astronomy, as datasets have grown in size and fast computers have become more widely available. They allow us to analyse large datasets and discover complex patterns that are often difficult to capture using traditional methods. In particular, supervised ML algorithms are widely used to predict physical properties of galaxies based on observational data (e.g., Bonfield et al. 2010; Acquaviva 2016; Domínguez Sánchez et al. 2018; Pasquet et al. 2019; Chu et al. 2024; Daza-Perilla et al. 2025). One key application is the estimation of rest-frame absolute magnitudes of galaxies from their observed photometric properties, without relying on model-dependent spectral energy distribution (SED) fitting. In this chapter, we explore the use of Random Forest Regression (RFR), a supervised learning method, to predict the rest-frame absolute magnitudes of galaxies.

For completeness, before introducing the machine learning technique used in this study–Random Forest–and describing its basic building block, the Decision Tree, we first explain the concept of k-correction and the conventional methods

typically used to compute it. The chapter is organised as follows: §5.1 discusses key aspects of k-correction and its role in measuring the intrinsic brightness of galaxies. §5.2 outlines the working of a Decision Tree and how it can lead to overfitting. §5.3 describes how the Random Forest Regression model addresses the overfitting problem. We also detail the procedure for splitting the sample into training, testing, and validation sets, as well as the fine-tuning of model hyperparameters using cross validation. Finally, §5.4 presents the k-corrections and rest-frame i-band magnitudes predicted using the machine learning approach.

5.1 k-correction

The luminosity of a galaxy is an intrinsic property that is proportional to the number of stars in the galaxy, their ages, and metallicities (see the review by Conroy 2013). In galaxy surveys that span a wide redshift range, a given observed filter samples progressively shorter rest-frame wavelengths of a galaxy's spectral energy distribution (SED) as redshift increases. To simplify comparisons across redshifts, it is common practice to evaluate galaxy luminosities at a fixed rest-frame wavelength. In this study, we select galaxy samples in an observed band, the CFHTLS-i band (as measured in PAUS), and estimate the luminosity function at a fixed reference redshift of $z = 0$, also in the i-band.

The adjustment between the observed i-band magnitude of a galaxy at redshift z and the equivalent i-band magnitude in the $z = 0$ rest frame is known as the k-correction. A full derivation of the k-correction can be found in Hogg et al. (2002). Here, we show the final result, given by:

$$K_{\rm QR} = -2.5 \log_{10} \left[\frac{\int \frac{d\nu_o}{\nu_o} S_\nu(\nu_o) R(\nu_o)}{S_{\nu,o} \int \frac{d\nu_o}{\nu_o} R(\nu_o)} \frac{S_{\nu,e} \int \frac{d\nu_e}{\nu_e} Q(\nu_e)}{\int \frac{d\nu_e}{\nu_e} S_\nu(\frac{\nu_e}{1+z}) Q(\nu_e)} \right], \tag{5.1}$$

where S_ν is the galaxy SED, R and Q are the transmission functions of the observed and rest-frame filters, respectively, and the integrals are over frequency space in the observer and emitter frames.

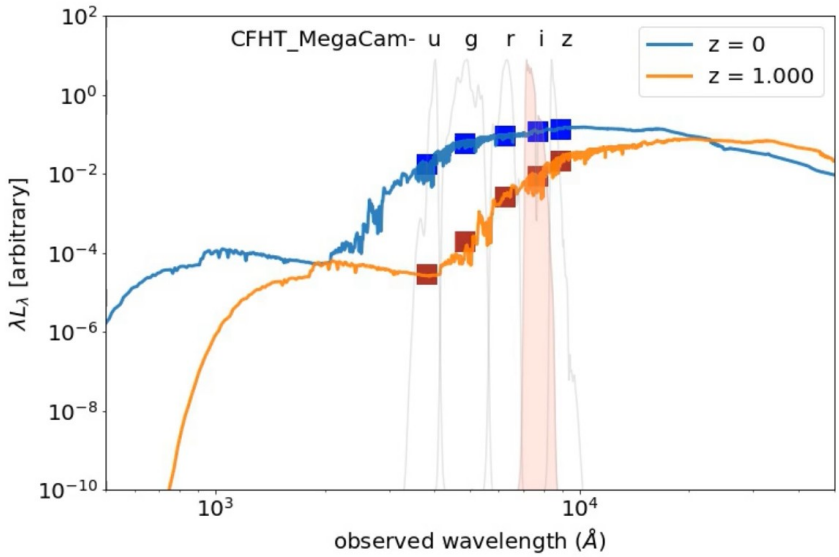


Figure 5.1: K-correction of a model SED at $z = 1$ (orange). The SED is plotted in the observed wavelength frame. The same model galaxy is shown at redshift $z-0$ in blue. The vertical difference between each pair of symbols in the same filter indicates the k-correction.

Fig. 5.1 illustrates how the luminosity of a galaxy can be k-corrected when both the shape of its spectral energy distribution (SED) and its redshift are known. The k-correction quantifies the difference between the observed and rest-frame fluxes in a given filter, and depends sensitively on the SED shape. In practice, a galaxy's SED can be estimated by fitting empirical templates to the observed photometry or by generating synthetic SEDs based on assumed star formation histories (and, more rarely, chemical enrichment histories; e.g., Mitchell et al. 2013). However, constructing an accurate SED for every single galaxy in a large survey like PAUS can be computationally intensive, especially given its depth and wide area coverage.

`GALFORM` predicts the Star Formation History (SFH) and chemical evolution of the disk and bulge components of model galaxies. It also includes a dust attenuation model (see Lacey et al. 2016 for details). For each galaxy, `GALFORM` provides both rest-frame and observed-frame magnitudes, allowing the "*exact*" k-correction to be computed at the redshift corresponding to its lightcone crossing.

Fig. 5.2 shows the distribution of k-corrections predicted by `GALFORM` as a function of redshift for all galaxies in the PAUS lightcone mock catalogue down

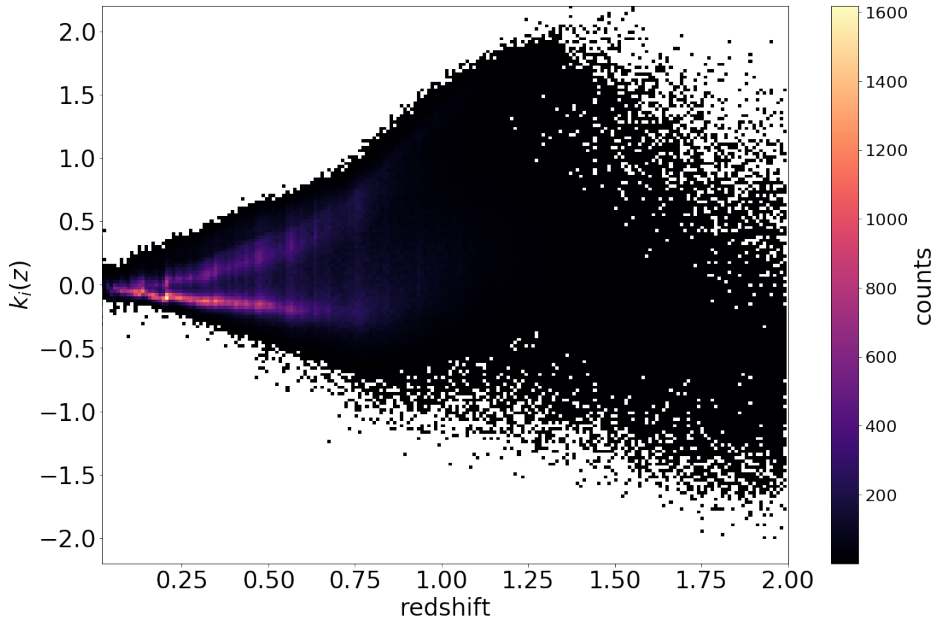


Figure 5.2: The "true" k-correction values for galaxies with $i_{AB} \leq 23.0$ in the `GALFORM` lightcone mock catalogue, plotted as a function of redshift. The shading indicates the number of galaxies per pixel.

to $i_{\rm AB} = 23.0$. The distribution exhibits a clear bimodality, with two distinct populations corresponding to red and blue galaxies (see Fig. 5 of Manzoni et al. 2024 for observed $(g - r)$ colour vs. redshift). The negative branch of the k-corrections corresponds to blue galaxies.

A simple model for the k-correction would be to fit a parametric function of redshift to the corrections predicted for the bulk of the red and blue galaxies, tracing the ridges in the galaxy distribution in Fig. 5.2. Although the peaks in the k-corrections are relatively well defined, there is considerable scatter around the ridges. As a result, such a simplified model would lead to significant errors in the k-correction, and in turn in the rest frame i-band absolute magnitude. These magnitude errors can affect the shape of the estimated luminosity function, particularly at the bright end where the function changes rapidly with increasing luminosity.

To reduce this problem, the k-correction is often computed using multiple colour bins (e.g. McNaught-Roberts et al. 2014). This strategy reduces the offset between a galaxy's true k-correction and the nearest parametric fit by tailoring

the correction to galaxies with similar colours. However, this method does not account for the evolution of stellar populations across the redshift interval over which the k-correction is applied–such as the formation of new stars or the fading of existing ones (i.e., the luminosity evolution correction). Although this approach has the practical advantage of producing a $k(z)$ curve, it is not guaranteed that a given galaxy would remain in the same colour bin at all redshifts, potentially undermining the accuracy of the correction.

We adopt a supervised machine learning technique–Random Forest Regression (RFR: Breiman 2001)–to address the k-correction problem. Although we do not test alternative algorithms here, RFR has been widely applied in astronomy alongside methods such as neural networks and gradient boosting. Compared to these, RFR requires relatively little hyperparameter fine-tuning and can be trained efficiently without GPU acceleration. While gradient boosting methods often achieve higher performance, they demand more careful optimisation (Carliles et al., 2010; Baron, 2019) . For these reasons, we select RFR as a practical and robust choice, while noting that exploring alternative models could further improve the k-correction framework.

For our calculation, we train the ML model using the `GALFORM` lightcone mock catalogue, where both observed and rest-frame properties of galaxies are known. The input features consist of 5 broadband photometric magnitudes along with redshift, and the model is trained to predict the k-correction values of galaxies.

It is worth noting that an evolutionary correction is sometimes applied to the rest-frame luminosities to account for the changes in the stellar population between the redshift of lightcone crossing and $z = 0$ due to ongoing star formation in the galaxy, as well as the ageing of the existing stellar population. We do not attempt to make any such correction. Instead, differences in the luminosity function between redshifts reflect this evolution, as well as changes in number density at a given luminosity due to galaxy mergers.

5.2 Decision Tree

A decision tree is a simple yet powerful machine learning algorithm that can be used for both classification and regression tasks. It works by recursively splitting the data into subsets based on feature values, aiming to create groups that are as homogeneous as possible with respect to the target variable. In the case of regression, the decision tree predicts a continuous value by learning a set of decision rules from the input features (Quinlan 1986).

The tree starts at a root node, where the entire dataset is considered. At each internal node, the data is split based on the value of one of the features. The choice of feature and split value is determined by minimising a loss function, such as the Cross-Entropy (i.e. log-loss) for classification problems or the mean squared error (Mean Squared Error (MSE)) for regression problems. These loss functions determine how well the model can perform by comparing the predictions and the true value. This splitting process continues recursively until a stopping condition is reached, such as a maximum depth (i.e. a maximum number of splits made) or a minimum number of samples is reached in a leaf node. The log-loss is defined as:

$$\text{Log-Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} log(p_{i,k}), \tag{5.2}$$

where N is the number of the samples, K is the number of classes, $y_{i,k}$ is set to be 1 if sample i belongs to class k, else 0 (this method of assigning values to classes is called "one-hot encoding"), $p_{i,k}$ is the predicted probability that sample i belongs to class k with $\sum_{k=1}^{K} p_{i,k} = 1$ and $0 < p_{i,k} < 1$. However, this loss function is primarily used for classification problems, as discussed above. For regression tasks, The MSE is more appropriate, as it directly measures the deviation between continuous model outputs and the true values. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{N}(y_i - p_i)^2, \tag{5.3}$$

where N is the number of samples, y_i is the true value of the sample i, and p_i is the prediction.

As an example, the mock catalogue from Manzoni et al. (2024) shows that galaxies can be classified into red or blue populations based on their observed $(g-r)$ colour, based on the colour cut given by Manzoni et al's Equation 2. Fig. 5.3 displays the distribution of $(g-r)$ colour as a function of redshift, with the red and blue populations highlighted as red and blue points, respectively. The location of the cut is based on a valley in the density of points in the colour-redshift plane.

A decision tree can be trained to classify galaxies into these populations using 3 input parameters (or "features"): redshift, $(g-r)$, and Δ. The auxiliary parameter Δ is introduced to simplify the structure of the decision tree and improve its visualisation. It is defined as the difference between $(g-r)$ and the RHS of Equation 2 of Manzoni et al. (2024), defined as:

$$\Delta = (g-r) - 1.7 \times \text{redshift} - 0.35. \tag{5.4}$$

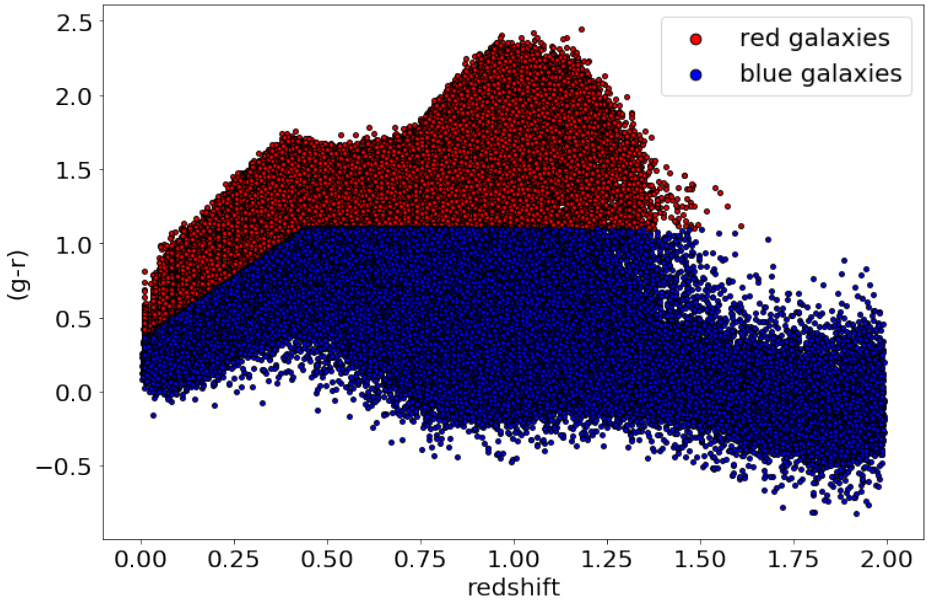


Figure 5.3: Distribution of observed galaxy $(g-r)$ colour as a function of redshift for the `GALFORM` lightcone mock catalogue. Galaxies are classified as red or blue according to the colour cut in Manzoni et al. (2024), and are shown in red and blue accordingly.

The example decision tree trained on 713,622 galaxies from the `GALFORM` lightcone mock catalogue is shown in Fig. 5.4. There are 473,727 blue galaxies and 239,895 red galaxies to begin with. At the root node (node #0), galaxies with ob-

served $(g-r)$ greater than 1.1 are classified as red galaxies while the other fail to satisfy the condition are further to classified in the next node. After the first node, the model is able to extract $201,660$ red galaxies from the sample. The remaining $511,962$ galaxies are then split at node #1 using the Δ parameter. This final split results in a clear separation between red and blue galaxies at the leaf nodes (Node #2, #3, and #4).

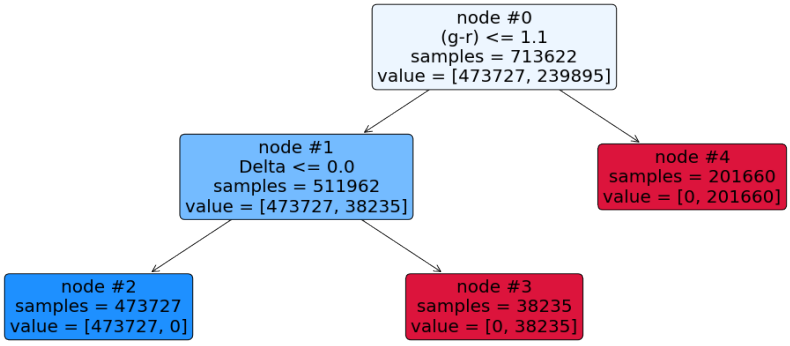


Figure 5.4: Example of a decision tree trained to classify galaxies into red and blue populations using redshift, $(g-r)$ colour, and the auxiliary parameter Δ (see text). The training set contains 713,622 galaxies from the `GALFORM` lightcone mock catalogue. The sample consists of 473,727 blue and 239,895 red galaxies, represented as "value" in the lowest line of each node in the chart. The tree structure shows how the sample is split at each node to achieve final classification at the leaf nodes.

Decision trees are fast to train and easy to interpret, as shown in this example. However, a single decision tree tends to overfit the training data, especially if the tree is allowed to grow very deep. This means that while it may perform well on the training set, its performance on unseen data can be poor. To overcome this limitation, ensemble methods such as Random Forests have been developed.

5.3 Random Forest Regression

Although decision trees are powerful and easy to interpret, they tend to suffer from overfitting. This overfitting leads to high variance, where the model fits the noise in the training data instead of capturing the underlying patterns. One popular

method to overcome this limitation is the Random Forest algorithm (Breiman 2001).

A random forest is an ensemble learning method that builds a large number of individual decision trees during training and combines their outputs to improve predictive performance. For regression tasks, the final prediction of a random forest is the average of the predictions from all the individual trees. This ensemble approach significantly reduces the risk of overfitting and improves the model's ability to generalise to unseen data.

Random forest introduce randomness in two ways:

- **Bootstrap sampling**: Each decision tree is trained on a random subset of the training data, drawn with replacement.
- **Random feature selection**: At each split within a tree, a random subset of features is considered for splitting rather than evaluating all features. This introduces diversity among the trees and further reduces correlations between them.

The general structure of a random forest is illustrated in Fig. 5.5. Multiple decision trees are trained independently, and their predictions are averaged to obtain the final output. This process helps to stabilise predictions and leads to higher overall accuracy.

In Astronomy, random forests have been successfully applied to a variety of problems, such as predicting galaxy metallicities (Acquaviva 2016), estimating redshifts and stellar masses (Mucesh et al. 2021), estimating star formation rates (Bonjean et al. 2019), and classifying galaxy morphologies (Fontirroig et al. 2024). Their ability to handle large datasets, noisy inputs, and complex, non-linear relationships makes them highly suitable for astrophysical applications. Moreover, in general random forests have comparable performance to other machine learning methods, such as neural networks (Carliles et al. 2010)

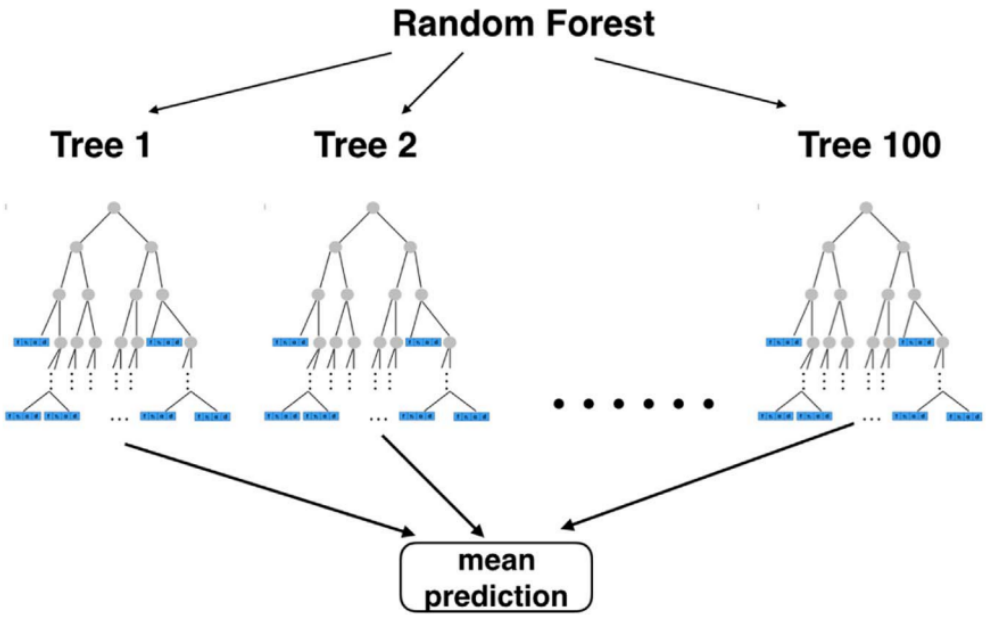


Figure 5.5: An illustration of Random Forest Regression that is formed of 100 trees. The output of the Random Forest is averaged over the predictions of each individual tree. Image taken from Nedjati-Gilani et al. (2017).

Here, we extend the use of random forest to estimate the k-corrected rest-frame absolute magnitudes of galaxies based on their observed photometric properties. Instead of relying on traditional model-dependent SED fitting methods, the random forest learns the mapping between observed-frame magnitudes, redshifts, and the target rest-frame absolute magnitudes directly from the data, which in this case is the output of a galaxy formation model.

We train the random forest to learn the mapping between observed-frame magnitudes (and redshifts) and the rest-frame absolute magnitude. The input features include observed magnitudes in CFHT$-ugriz$ bands and a photometric redshift estimate. The target value is the $k-$correction computed for a subset of galaxies from the lightcone mock catalogue. After training, the random forest model can predict the rest-frame absolute magnitudes for new galaxies directly from their observed photometry and redshift, without the need for explicit SED fitting, which requires a set of model assumptions.

Here, the method we chose is the Random Forest Regression (RFR) algorithm

from the publicly available package `RandomForestRegressor` from Scikit-Learn (Pedregosa et al., 2011), as this algorithm has been widely tested in astronomy (Baron, 2019; Hernandez Vivanco et al., 2020; Hoffman et al., 2021).

In the following subsections, we describe in detail the preparation of the training and testing data sets, the procedure for tuning the random forest model, and the method used to validate the model performance.

5.3.1 Training and testing data set

The training and testing datasets for the RFR model are constructed using galaxies generated by the `GALFORM` semi-analytic model. In this study, we use the lightcone outputs from Manzoni et al. (2024). For computational reasons, 250 sub-boxes (out of 1024) are selected to create the dataset for training the machine learning model. This means that the overall count of galaxies is about one quarter of what it should be. By using the output of a galaxy formation model, we do not need to make assumptions about the form of the star formation history (the model calculates this) or about the chemical evolution of the interstellar medium or the dust extinction (again, these are computed in the model).

The observational properties extracted for each galaxy include the observed magnitudes in the CFHT MegaCam u, g, r, i, and z bands, as well as the observed redshift, $z_{\rm obs}$. Additionally, the intrinsic i-band absolute magnitude is retrieved to compute the target variable.

The target variable for the machine learning model is the $k-$correction for the $i-$band, denoted as $k_{\rm i}(z)$, which is defined as the difference between the observed apparent magnitude, $m_{\rm i}$, and the rest-frame absolute magnitude ($M_{\rm i}$) after correcting for known distance effects (i.e. taking into account the luminosity distance for the assumed cosmology):

$$k_{\rm i}(z) = m_{\rm i} - M_{\rm i} - 5\log_{10}\left(\frac{d_{\rm L}}{{\rm Mpc}/h}\right) - 25, \tag{5.5}$$

where $d_{\rm L}$ is the luminosity distance, obtained by interpolating a pre-computed distance-redshift table to save time. The `GALFORM` model predicts the k-correction exactly at the simulation output redshifts, as the i-band magnitude can be requested in the rest and observed frames. The observer frame at the redshift of lightcone crossing is interpolated between the values at the snapshots either side of this redshift (see Merson et al. 2013 for a discussion of this point). The distribution of $k_{\rm i}(z)$ as a function of redshift for the galaxies in the lightcone is shown in Fig. 5.2.

The input features used for training consist of 6 parameters:

- Observed redshift, z_{obs}
- 5 apparent magnitudes, CFHT-$ugriz$.

To ensure the quality of the training set, galaxies with redshift $z_{obs} < 0.001$ are excluded from the analysis, some of these objects could be misclassified as stars. Additionally, only galaxies with apparent i-band magnitude brighter than the PAUS flux limit ($i_{AB} = 23.0$) are selected for training.

The dataset is randomly divided into a training set and a testing set, with 80% of the galaxies used for training and the remaining 20% reserved for testing. This division allows for the evaluation of the model's performance on unseen data, ensuring that the model does not simply memorise the training examples.

The detailed tuning of the random forest model parameters and performance evaluation are discussed in the following sections.

5.3.2 Tuning the model

Before training the final RFR model, it is important to optimise its hyperparameters to achieve the best predictive performance (whilst also balancing against a high computational cost and avoiding overfitting). Each hyperparameter controls a different aspect of how the model builds its ensemble of decision trees. The definition of each hyperparameter in the RFR model is described below:

- `n_estimators`: The number of decision trees in the forest. For example, Fig. 5.5 shows a random forest with `n_estimators` = 100. Increasing this generally improves performance, but with diminishing return for very large values.

- `max_features`: The maximum number of features considered when looking for best split at each node. The choices of `max_features` includes "$\mathtt{n}_f$" (all features, or inputs, which is $\mathtt{n}_f = 6$ in our case), "`sqrt`" (square root of the total features), and "`log2`" (log base 2 of the total features). The value of "`sqrt`" and "`log2`" after rounding down to the closest integer is the same in our case. Therefore the final hyperparameter choices are "$\mathtt{n}_f$" and "`sqrt`". Using fewer features increases randomness, helping to decorrelate the trees.

- `max_depth`: The maximum depth of each individual tree. The maximum depth can be set to any arbitrary integer with bigger number for deeper trees. The tree can also be set to grow indefinitely using `max_depth =` "`None`". This allows a tree to grow until its leaf nodes reach `min_samples_leaf`. Limiting the depth prevents tress from growing too complex and overfitting to noise.

- `min_simples_split`: The minimum number of samples required to split in an internal node. Increasing this value can make the model more conservative.

- `min_simples_leaf`: The minimum number of samples required to be at a leaf node. Setting this to a large value can smooth predictions and reduce model variance.

- `min_weight_fraction_leaf`: The minimum weighted fraction of the input samples required to be at a leaf node. In this study, it is fixed at 0.0.

- `max_leaf_nodes`: The maximum number of leaf nodes. Setting this limits the complexity of the trees; in this work, it is also fixed to "`None`" to allow unlimited growth.

The original hyperparameter grid considered in this work is shown in Table. 5.1.

RFR Hyperparameters	Range Tested	Best Values
`n_estimators`	[1, 2, 3, 5, 6, 8 ,10, 12, 14, 16, 18, 20, 25, 30, 40, 50, 75, 100, 125, 150]	40
`max_features`	[n_f, 'sqrt']	n_f
`max_depth`	["None", 1, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 50]	15
`min_samples_split`	[2, 5, 10]	2
`min_samples_leaf`	[1, 2, 4]	1
`min_weight_frac_leaf`		0.0
`max_leaf_node`		"None"

Table 5.1: The hyperparameters for the RFR. The first block of rows shows the tuned hyperparameters, the range of values or options tested (second column) and the best values adopted (third column). The lower block of parameters were not varied. The hyperparameter names are explained in the text.

Rather than exhaustively searching all possible combinations, we use a random search approach with the `RandomizedSearchCV` module (Bergstra and Bengio 2012) from `scikit-learn`. This method randomly samples hyperparameter combinations, offering a more efficient way to explore a wide parameter space. If an exhaustive grid search had been performed, the total number of possible combinations would have been 5,040 (each across 7 folds), making random sampling a much more computationally efficient alternative. The performance metric used during tuning is the mean squared error (MSE).

A total of 200 random hyperparameter combinations are sampled and evaluated. For each combination, a 7-fold cross-validation is performed, where the training set is split into 7 parts (see §5.3.3 for details). The model is trained on 6 parts and validated on the remaining part, rotating through all possible splits.

The combination of hyperparameters that yields the lowest-cross-validation error is selected as the best model. In total, the Randomized Search involved training and evaluating the model 1,400 times, corresponding to 200 random hyperparameter combinations and 7-fold cross-validation for each combination. The optimised random forest model is then used for further predictions and evaluations.

As a result from the hyperparameter tuning using `RandomizedSearchCV`, Fig. 5.6

shows the performance metric (i.e. MSE in this case) against the number of trees (`n_estimator`) and Fig. 5.7 shows the performance metric plotted against maximum depth (`max_depth`). Here we show the performance metrics for both the training (blue dots) and testing sets (orange dots). We find that the performance of the random forest model changes dramatically at the beginning when the number of trees and the value of maximum depth increase, but after certain values the performance does not improve much in terms of accuracy. Despite `RandomizedSearchCV` suggesting that 150 trees and maximum depth of 50 is the best combination (i.e. providing the lowest MSE value), it was found to produce signs of overfitting. To avoid the overfitting problem, we use Fig. 5.6 and Fig. 5.7 to manually choose the acceptable combination of hyperparameters. The final set of hyperparameters used in our RFR model consists of `n_estimator` = 40 and `max_depth` = 15 as the model parameters along with `max_features` = n_f, `min_samples_split` = 2, and `min_samples_leaf` = 1, as shown in the last column of Table 5.1.

5.3.3 Cross validation

Cross-validation is a model validation technique used to assess how a machine learning model generalises to an independent dataset. It involves partitioning the original training data into a set of smaller subsets, known as "folds." The model is trained on a combination of these folds and validated on the remaining fold. This process is repeated several times so that each fold is used once as the validation set.

The main advantage of cross-validation is that it allows the model's performance to be evaluated more reliably than using a simple train-test split. It helps detect overfitting by ensuring that model performs consistently across different subsets of the data.

Fig. 5.8 shows an example of 5-fold cross-validation. The training set is divided

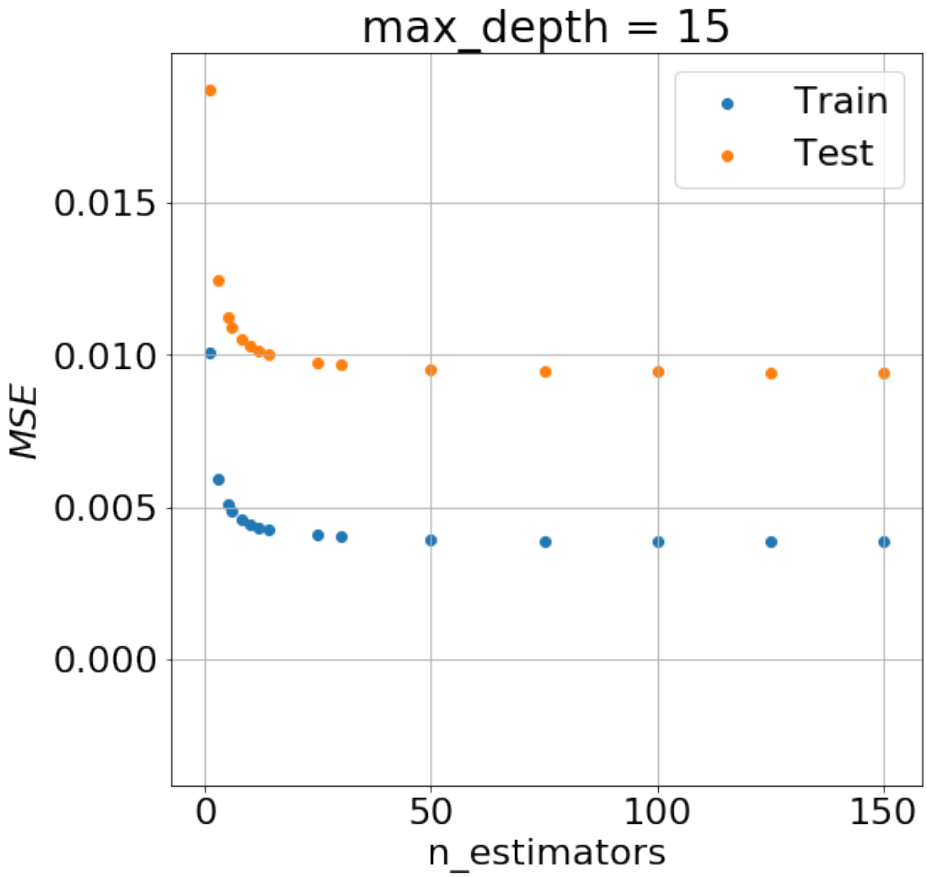


Figure 5.6: The model performance metric during the hyperparameter tuning process using `RandomizedSearchCV`. We use MSE as the performance metric and plot it against the number of trees (`n_estimators`). The model evaluations during training and testing are shown as orange and blue dots, respectively. The optimal number trees is chosen to be 40.

into 5 equal parts. For each training iteration, 4 parts are used for training the Random Forest model (or any machine learning algorithm in general), while the remaining part is used for validation. This process is repeated 5 times, rotating the validation set each time. The model's final performance is then averaged over all 5 runs. In this work, we use 7-fold cross-validation as described in the previous section.

Cross-validation is employed both during the hyperparameters tuning stage and for evaluating the predictive accuracy of the final model. By using 7-fold cross-validation, we reduce the variability associated with a single train-test split and obtain a more robust estimate of the model's generalisation ability.

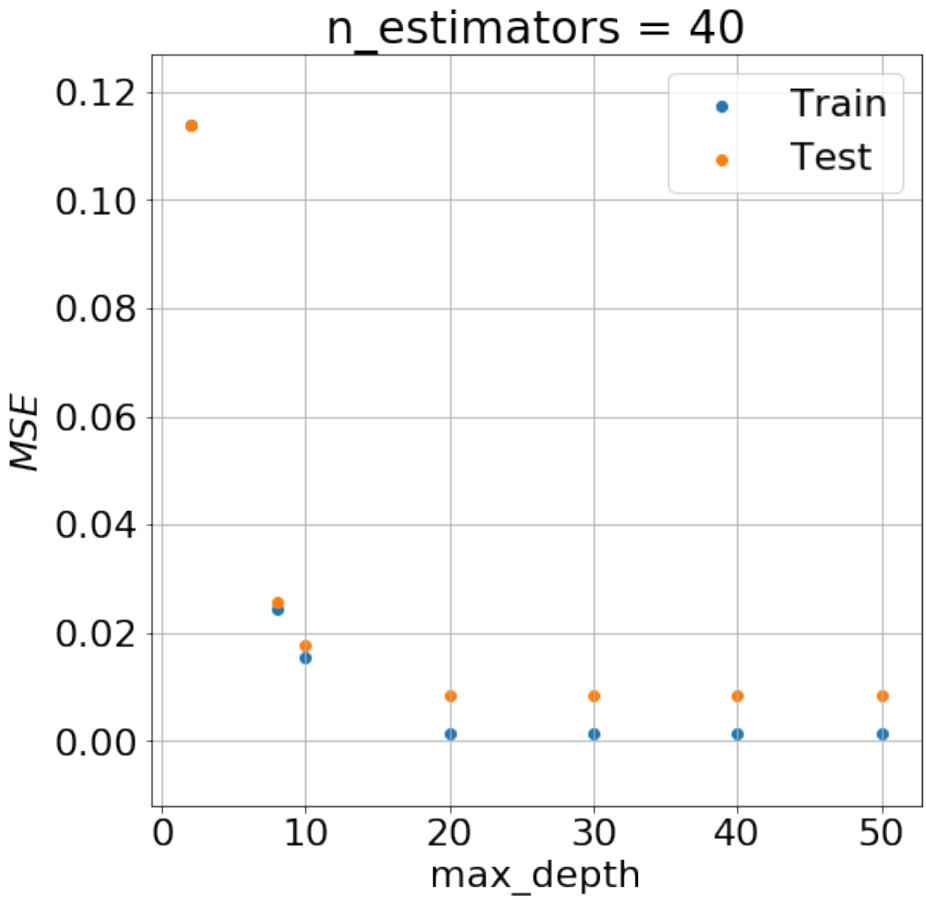


Figure 5.7: The same description as Fig. 5.6, but showing MSE as a function of maximum depth instead. Here `max_depth` is chosen to be 15.

5.4 Machine Learning-based k-correction

5.4.1 k-corrections and estimation of the rest-frame magnitude

Rather than using a single observed colour to predict the galaxy k-correction, we use all of the available photometry in the *ugriz* filters. We use machine learning to find the relation between observable properties–*ugriz* photometry with the redshift–and the k-correction predicted by `GALFORM`. The RFR is computationally efficient and its performance is not overly sensitive to the choice of hyperparameter values. Nevertheless, to find the best RFR model for our problem, we selected the hyperparameter values using the random grid search technique combined with k-fold cross-validation to avoid overfitting.

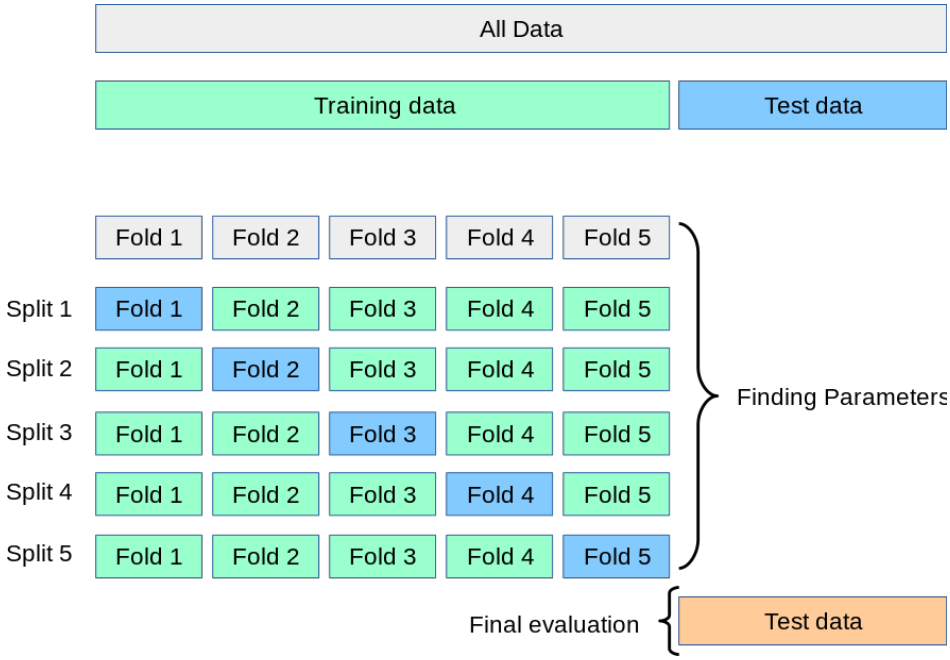


Figure 5.8: The schematic chart explaining the sample split into train and test datasets. An example of 5-fold cross-validation is performed for finding hyperparameters, before final evaluation on the test data. Chart taken from https://scikit-learn.org/stable/modules/cross_validation.html.

5.4.2 The prediction error of the rest-frame absolute magnitude using the Random-Forest-Regression k-correction

The RFR machine learning prediction of the k-correction is compared to the exact answer from `GALFORM` in Fig. 5.9. The error is presented as the difference between the predicted and true absolute magnitudes. A magnitude difference of zero would indicate that the machine learning method reproduces the `GALFORM` answer exactly. The results are shown for different redshift slices, as indicated by the legend. For $z \leq 1$, there is no bias in the predicted absolute magnitude, just a small scatter which reaches ≈ 0.05 mag for the brightest galaxies. At $z > 1$ there is a bias in the recovered magnitude, with the median differing from zero by up to ± 0.2 magnitude. The scatter is also larger, ~ 0.1 magnitudes. This sign of the bias means that the brightest galaxies are assigned an estimated absolute magnitude that is too faint, while the faintest galaxies in each redshift slice are brighter than they should be.

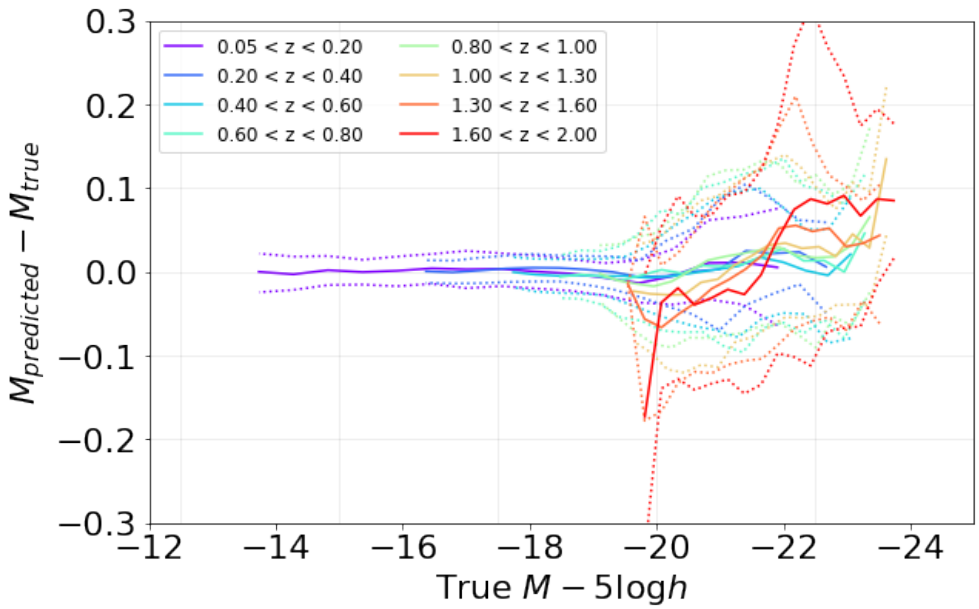


Figure 5.9: The performance of the RFR machine learning estimate of the k-correction, expressed in terms of the difference between the predicted and true absolute magnitude for each galaxy. The true magnitude is predicted using the exact k-correction from `GALFORM`. The solid lines show the median difference between the true and predicted magnitudes. The dotted lines show the $16^{\text{th}} - 84^{\text{th}}$ percentile interval, a centralised version of the 1σ scatter which is not affected by outliers. Different colours indicate different redshift ranges, as shown by the legend.

Chapter 6

The galaxy luminosity function

In this chapter, we present our main results on galaxy luminosity function (LF) using data from the Physics of the Accelerating Universe Survey (PAUS). The LF is a fundamental statistical tool for understanding the distribution of galaxy luminosity and their evolution over cosmic time. Our analysis leverages the high-resolution photometric redshifts provided by PAUS to estimate rest-frame luminosities and derive the LF across a wide redshift range.

We begin in § 6.1 by describing the methodology used to estimate the LF via the $1/V_{\rm max}$ technique. In § 6.2, we validate our approach using the `GALFORM` lightcone mock, illustrating key selection effects and the role of photometric uncertainties. § 6.3 presents the estimated i-band LF from the PAUS W1 and W3 fields, including a detailed analysis of observational uncertainties in § 6.3.1. In § 6.4, we compare our measurements with previous LF estimates from the literature. § 6.6 explores the LFs of red and blue galaxy populations, and § 6.5 investigates the redshift evolution of the LF. We assess sample completeness and the impact of photometric redshift quality in § 6.8. In § 6.9, we extend our analysis to the u, g, r, and z bands to examine wavelength-dependent trends. Finally, we summarise the key findings of this chapter in § 6.10.

6.1 The V_{max} Methodology

Before introducing the method used to measure the luminosity function, it is useful to first outline the general behaviour of galaxy populations. The galaxy luminosity function (LF)–the number density of galaxies as a function of their luminosity–is commonly described by the Schechter function (Schechter, 1976):

$$\phi(L)dL = \phi^* \left(\frac{L}{L_*}\right)^{\alpha} \exp\left(-\frac{L}{L_*}\right) d\left(\frac{L}{L_*}\right), \tag{6.1}$$

where $\phi(L)dL$ is the number density of galaxies with luminosities between L and $L + dL$, L_* is the characteristic luminosity marking the transition between the faint-end power-low slope and the bright-end exponential cut-off, ϕ_* is the number density normalisation constant, and α is the faint-end slope. Fig 3.2 shows the overall shape of the galaxy luminosity function predicted by `GALFORM` at different redshifts. At the faint end, the power-law term $(L/L_*)^{\alpha}$ dominates; the slope α determines how steeply the LF rises at faint luminosities. For bright galaxies, the exponential term $\exp(-L/L_*)$ dominates; reflecting the rarity of very bright galaxies.

In this work, we estimate the galaxy luminosity functions using the $V_{\rm max}$ methodology (Schmidt, 1968). In a flux-limited sample, the number of sources observed varies strongly with luminosity (Driver and Phillipps, 1996). At the faint end, few galaxies are detected because they are only brighter than the apparent magnitude of the selection limit at low redshifts. Thus, although many faint galaxies are numerous per unit volume, they are only visible over relatively small comoving volumes compared to brighter galaxies. Conversely, while intrinsically luminous galaxies can be seen over larger redshift ranges, they are rarer, with their number density falling off exponentially beyond a characteristic luminosity L_*, as discussed in Equation 6.1. As a result, the observed number of galaxies therefore peaks around L_*.

The V_{max} approach corrects the observed number of objects to provide an estimate of the true, underlying number density of galaxies, by accounting for the different volumes over which galaxies of varying luminosities can be observed. The luminosity function $\phi(M)$, defined as the number density of galaxies per unit absolute magnitude, is estimated by performing a weighted summation over all galaxies in a given absolute magnitude bin:

$$\phi(M)\Delta M = \sum_i \frac{1}{V_{\mathrm{max},i}}, \tag{6.2}$$

where $V_{\mathrm{max},i}$ is the maximum comoving volume over which the i-th galaxy could have been observed, and ΔM is the width of the magnitude bin.

The value of V_{max} for each galaxy is determined by the maximum redshift, z_{max} at which the galaxy would still be included in the sample. Starting at the redshift of observation, z, we can imagine increasing the redshift gradually, adjusting the luminosity distance and k-correction accordingly until we reach the maximum redshift, z_{max}, at which the galaxy is selected in the sample. At this redshift, the apparent magnitude of the galaxy is equal to the limit that defines the catalogue, m_{limit}:

$$m(z_{\mathrm{max}}) = m_{\mathrm{limit}} = M + 5\log_{10}\left(\frac{d_{\mathrm{L}}(z_{\mathrm{max}})}{h^{-1}\mathrm{Mpc}}\right) + k(z_{\mathrm{max}}) + 25, \tag{6.3}$$

where d_{L} is the luminosity distance. An intrinsically faint galaxy therefore has a smaller z_{max} than an intrinsically bright galaxy, and hence contributes a larger $1/V_{\mathrm{max}}$ weight to the LF.

PAUS galaxies are visible over a wide range in redshift (Manzoni et al., 2024). Hence, we can divide the volume covered into a series of thin redshift shells to isolate evolution in the luminosity function. Due to using thin shells in redshift, most galaxies are visible over the full redshift interval of the shell, $z_1 < z <_2$, so $z_{\mathrm{max}} > z_2$. This means that for most galaxies in each shell, $V_{\mathrm{max}} = V_{\mathrm{shell}}$. It is only for the faintest galaxies that $z_{\mathrm{max}} < z_2$, and these galaxies are only visible over part of the shell. Note that we know the value of the k-correction at the

redshift of observation, but we do not know the functional dependence on redshift, we make the approximation that the k-correction is constant when perturbing the redshift to find $z_{\rm max}$. Effectively this means $k(z_{\rm max}) = k(z)$, where z is the redshift at which the galaxy crosses the observer's past lightcone. Again, this assumption only affects a small number of faint galaxies in each redshift slice.

6.2 Testing the Estimation of the i-Band Luminosity Function using the GALFORM Lightcone Mock

Before applying our methodology to observational data, we first validate our approach using a mock catalogue constructed from the GALFORM semi-analytic model of galaxy formation. This provides a controlled setting to illustrate key features of our analysis and to understand how observational selection effects impact the estimated luminosity function.

To begin, we focus on the estimated luminosity function in a thin redshift slice at $z \sim 1$ to illustrate some features of our analysis. Fig. 6.1 shows various estimates of the luminosity function from the lightcone mock compared to the original prediction from GALFORM. The target or "true" luminosity function is the GALFORM prediction from the simulation box, which is simply a histogram of all the galaxies in the simulation volume, binned in luminosity, without any consideration of whether or not the galaxy is bright enough to meet the selection limit in the observed i-band. Hence, for this prediction, the number of galaxies keeps rising as the luminosity bin gets fainter (eventually, if we move to a faint enough bin, the GALFORM luminosity function will turn over due to mass resolution effects in the N-body simulation in which GALFORM has been implemented). The single curve labelled GALFORM in Fig. 6.1 is actually a weighted combination of the snapshot predictions that fall within the redshift shell, with the weight being the redshift distribution of galaxies at each redshift.

After taking into account the magnitude limit, $i_{\rm AB} = 23.0$, we obtain the

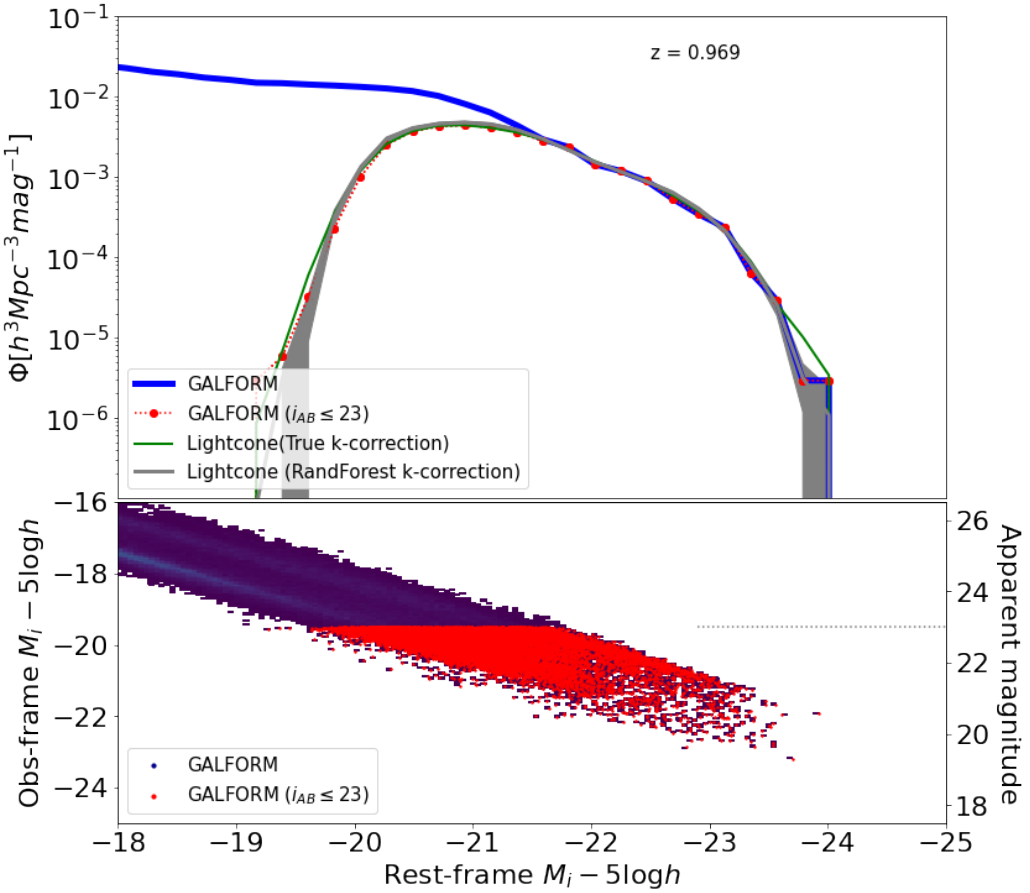


Figure 6.1: The impact of selection effects on the estimated luminosity function, shown at $z \sim 1$ for illustration. In the upper panel the blue curve shows the GALFORM LF without *any* selection effects. The red points show the LF after applying the observed i-band limit of $i_{\rm AB} = 23$. This curve is weighted version (using $\mathrm{d}N/\mathrm{d}z$) of the snapshot LFs over the redshift interval of the shell. The lower panel shows the observed and rest i-band absolute magnitudes of the model galaxies; points coloured red pass the sample selection. The green curve shows the LF estimated from the lightcone mock, assuming the exact k-correction predicted by GALFORM. The grey shading shows the LF recovered using the k-correction obtained using the random forest and the *ugriz* photometry.

red symbols in the upper panel of Fig. 6.1. Rather than showing a sharp cut in the LF in the rest frame i-band, there is a gradual reduction in the number of galaxies found as we move to fainter magnitudes. The reason for this is that the sample selection is in the observed i-band, whereas we plot the estimated LF in the rest i-band. This point is illustrated further in the lower panel of Fig. 6.1 which shows the observed frame i-band absolute magnitude (i.e. the absolute magnitude in Equation 6.3, but without applying the k-correction) plotted against the rest frame i, down to apparent magnitudes much fainter than $i_{\rm AB} = 23$. The overall distribution of galaxies in this plane is plotted in blue, with the (intrinsically) red and blue populations showing as light blue density enhancements. The galaxies which meet the observed $i_{\rm AB} = 23$ selection are coloured red. Galaxies with red colours make it into the sample over a wider range of rest-frame i magnitude than blue galaxies.

The estimated luminosity function over the magnitude range of the turn over is incomplete and is driven by the underlying luminosity function and the colour distribution of galaxies. Hence, this is still useful information with which to constrain galaxy formation models, if the same selection effects can be applied to the model galaxies, as is the case with our mock catalogue.

In the lower redshift bins, the observed and rest frame i-bands are closer together in redshift and the turnover at faint magnitudes in the recovered LF is narrower.

6.3 i-Band Luminosity Function: Estimate from PAUS

We compare the galaxy luminosity function estimated from the `GALFORM` lightcone mock catalogue of Manzoni et al. (2024) with that obtained from observational data in the PAUS W1 and W3 fields (Navarro-Gironés et al. 2024), using the $1/V_{\rm max}$ methodology. The `GALFORM` mock catalogue serves as a baseline to test our lumin-

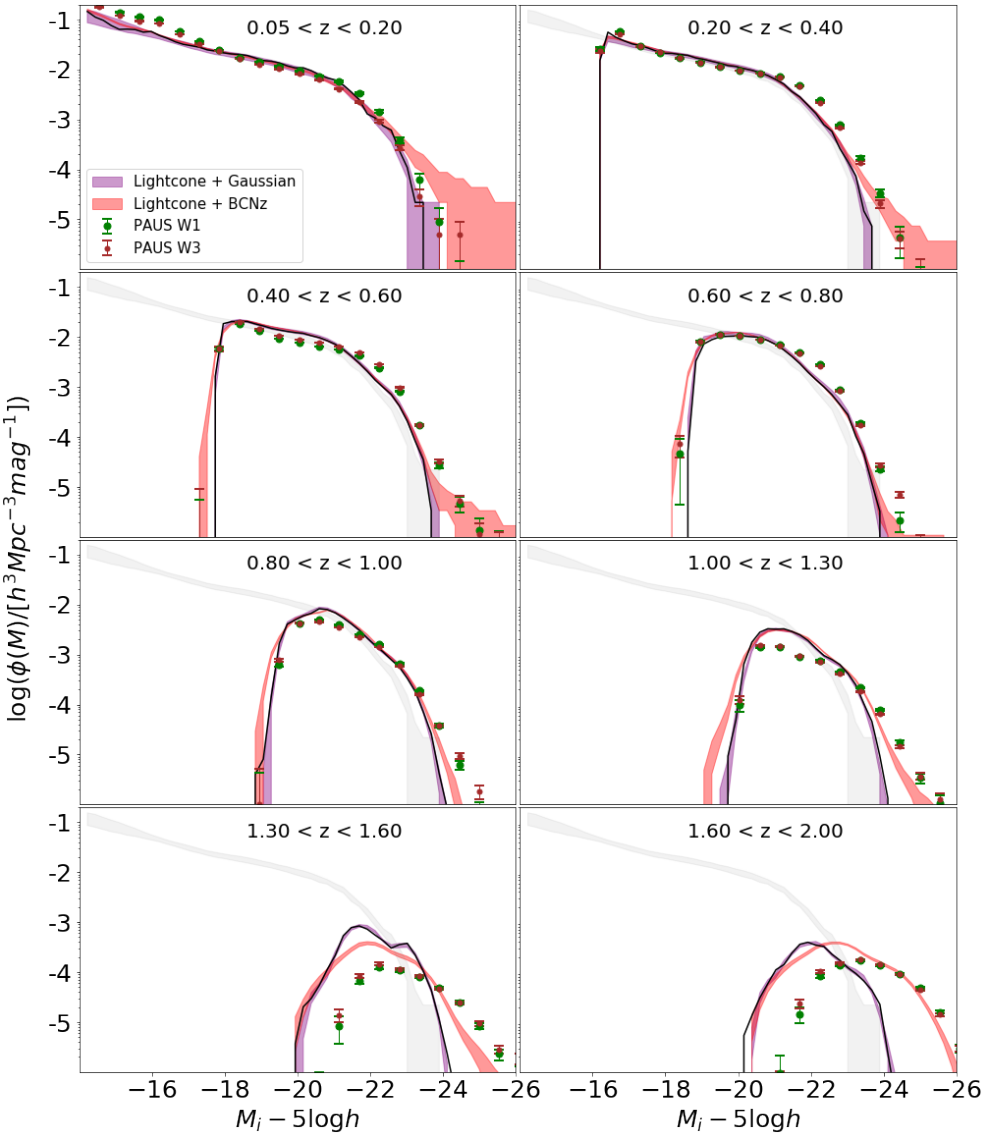


Figure 6.2: The rest-frame $i-$band luminosity function from PAUS data in the W1 field (green dots) and PAUS W3 field (red dots) compared to that from the `GALFORM` lightcone mock catalogue with the Gaussian-like photometric redshift uncertainties (purple shaded region) and the `BCNz`-like errors (pink shaded region) between redshift $z = 0.05$ and $z = 2.00$. The black solid line represents the "true" luminosity function from the lightcone mock catalogue (i.e. no photometric and photo-z errors).

osity function estimates and to reveal the impact of uncertainties and systematic effects introduced by the survey selection and methodology. These uncertainties–arising from the photometric errors, Large Scale Structure (LSS) sampling variance, and photometric redshift (photo-z) errors–are incorporated into the mock catalogue following the approach described in §4.6, and can be selectively switched on and off to examine their effects on the recovered luminosity function.

In this analysis, we consider two treatments of photometric redshift errors. In the first case, we apply errors obtained from the `BCNz2` code, which includes the presence of outliers in the photometric redshifts. In the second case, we mimic the central distribution of photo-z errors using Gaussian-like perturbations based on the $\sigma_{68}(\Delta z)$ values reported by Alarcon et al. 2021. We deliberately ignore outliers in this case. This dual approach enables us to explore the impact of low-quality photometric redshift measurements (or outliers) on the recovered luminosity function.

Fig. 6.2 presents the rest-frame CFHTLS i-band luminosity function in redshift slices over the redshift range $0.05 < z < 2.00$. In each redshift bin, the black solid line represents the "true" luminosity function from the lightcone mock catalogue–this is, the LF including the i-band selection but without any photometric or photo-z errors. The purple shaded region shows the mock luminosity function when Gaussian-like photo-z errors are applied, while the pink shaded region corresponds to the effect of the more realistic `BCNz2`-like photometric redshift uncertainties, which include redshift outliers. Observational data are shown using green filled diamonds for the PAUS W1 field and red solid circles for the PAUS W3.

The comparison in Fig. 6.2 demonstrates how photo-z measurement errors affect the shape of the estimated luminosity function. With Gaussian-like errors and no formal outliers (i.e. no redshift errors that lie beyond the wings of a Gaussian distribution), the errors in redshift lead to errors in luminosity but these are small compared to the size of the luminosity bins used to plot the luminosity function. Hence, the overall luminosity function fluctuates without a significant change in

shape (purple shading). In contrast, when the more realistic photometric redshift errors are applied, including outliers, the shape of the luminosity function is markedly different, particularly at the bright end, where the break becomes less pronounced. This occurs because a large error in redshift can lead to a large change in luminosity which places a galaxy in a different luminosity bin; this can lead to an appreciable change in the shape of the luminosity function at the bright end, where the variation of number density with increasing luminosity is rapid. At the faint end, the number density of galaxies varies more slowly with luminosity, so errors in luminosity have less impact on the shape of the luminosity function. The error bars are plotted using the combination of LSS and photometric+photo-z errors (see §6.3.1).

Up to $z \sim 1$ (equivalent to a lookback time of ~ 8.5 Gyr, more than 60 per cent of the age of the Universe), the luminosity function predicted from the `GALFORM` mock catalogue agrees reasonably well with the observational data from PAUS W1 and W3. At higher redshifts, the introduction of realistic `BCNz2`-like photo-z errors improves the match at the bright end. However, discrepancies at the faint end become apparent from $z \sim 1.0$ onward, growing larger at higher redshifts.

We do not attempt to fit a parametric form to the measured luminosity function, as our goal is to directly compare the observed LF with the mock catalogue that shares the same selection effects. For instance, fitting a single Schechter function would require discarding magnitude bins near the turnover at the faint end, which contain valuable information for testing galaxy formation models.

6.3.1 The errors in the luminosity function due to large scale structures, photometric and photometric redshift errors.

The uncertainties in the estimated luminosity function are calculated as follows: (1) large-scale structure or sampling variance errors are estimated using the Jackknife method (following Norberg et al. 2009) and (2) photometry and photometric

redshift errors are derived via a Monte Carlo (MC) approach.

We begin with the Jackknife resampling. Fig. 6.3 shows the layout of the Jackknife regions used for estimating the large scale structure (LSS) errors in the PAUS W1 field (left) and W3 field (right). In this work we use $N_{\text{regions}} = 64$ for both fields.

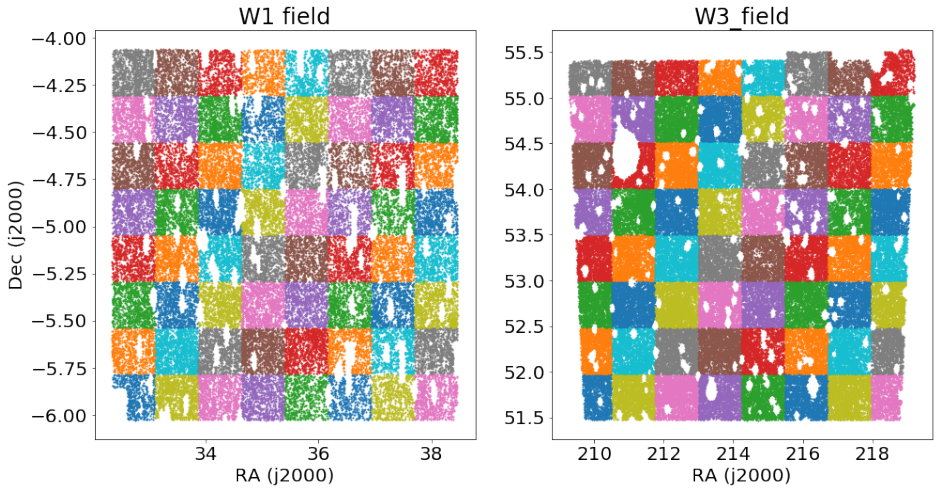


Figure 6.3: The Jackknife regions used for estimating the large scale structures (LSS) or sample variance errors in the PAUS W1 and W3 fields. Each field contains 64 regions.

The effect of using different numbers of Jackknife regions is shown in Fig. 6.4. The plot illustrates how the fractional error (standard deviation by mean) evolves with magnitude bin for both the W1 (solid lines) and W3 (dotted lines) fields.

In parallel, we estimate the contribution of the photometric and photo-z uncertainties using a Monte Carlo method. In each MC realisation, galaxy fluxes and photo-z are perturbed according to their observational errors. By repeating this process for many iterations, we quantify the scatter in the recovered luminosity function due to measurement uncertainties.

The results of the MC runs are shown in Fig. 6.5, where the scatter in LF estimates across iterations is visualised for both W1 and W3. To determine an appropriate number of iterations, we monitor the convergence of the standard deviation in each magnitude bin. As shown in Fig. 6.6, the error estimates stabilise after roughly 300 iterations. Although this would be sufficient for convergence, we

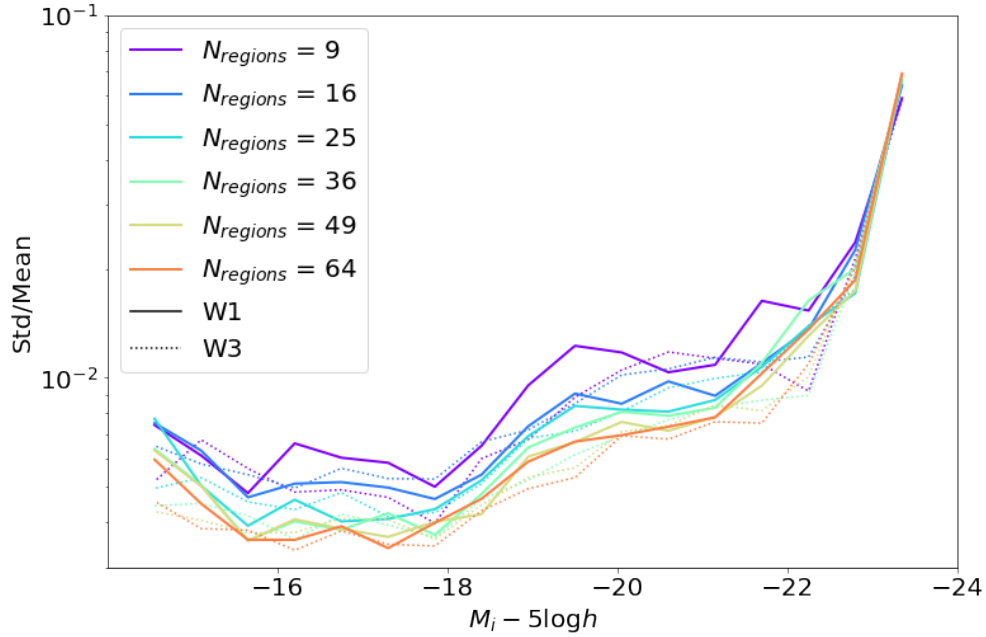


Figure 6.4: The Large-Scale Structure sample variance error using the Jackknife method with different number of regions. The solid lines represent the estimates for PAUS W1, while the dotted lines represent that from PAUS W3

conservatively use 500 iterations in all subsequent analysis.

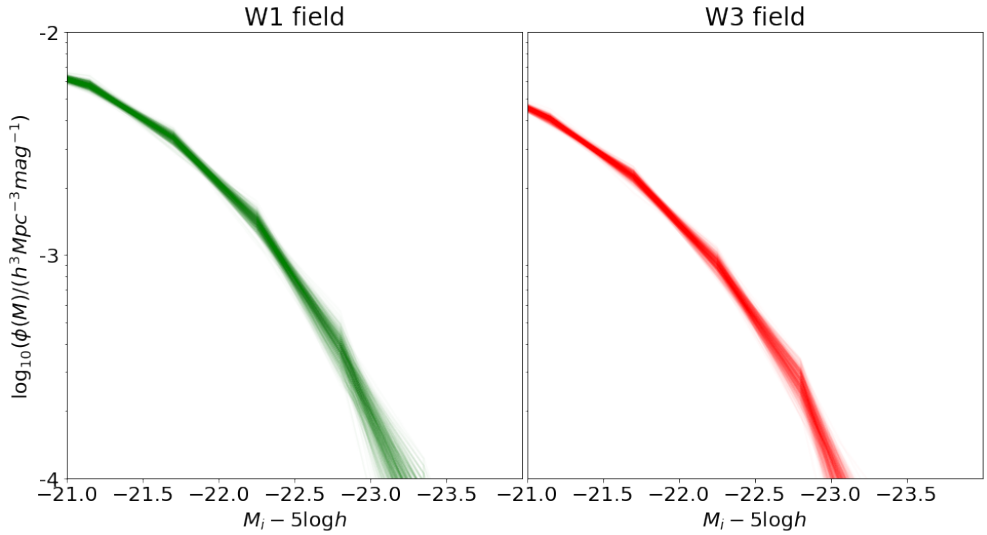


Figure 6.5: The scatter of the luminosity function after applying the photometric and photometric redshift at different Monte Carlo iterations. Note that both plots are cropped around the knee of the luminosity function for better visualisation. All 500 MC iterations are plotted in both panels.

The total errors in the luminosity function for each magnitude bin is then

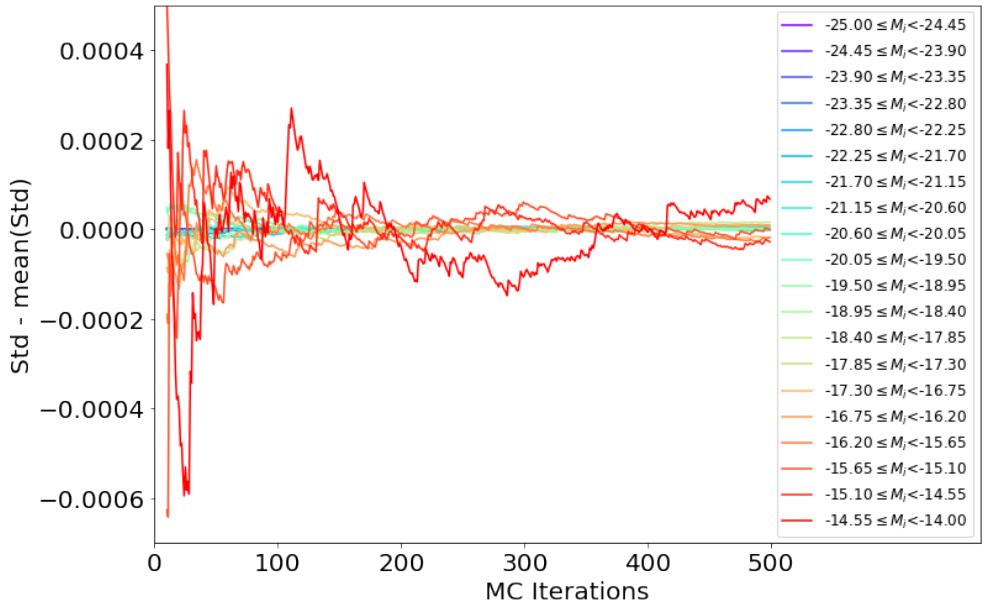


Figure 6.6: The change in the luminosity function errors estimated from the photometric and photometric redshift errors using different number of Monte Carlo iterations. The y-axis shows the value of Std. after i-th iteration (i.e. this means scatters of all estimated between 0^{th} and i^{th}) subtracted by the mean of Std. after 500 iterations. Colours represent the magnitude bins used in the LF estimates, as labelled in the legend.

calculated by combining the Jackknife and Monte Carlo components in quadrature:

$$\sigma_i = \sqrt{\sigma^2_{\text{Jackknife,i}} + \sigma^2_{MC,i}}, \tag{6.4}$$

where $\sigma_{\text{Jackknife},i}$ is the error from the Jackknife error estimation and $\sigma_{\text{MC},i}$ is the photometric and photo-z uncertainties.

Fig. 6.7 presents a comparison of the fractional errors in the luminosity function between PAUS data and the `GALFORM` mock catalogue. In all redshift bins, the photometric and photo-z errors (plotted as squares) dominate over the Jackknife errors (plotted as lines), by roughly an order of magnitude. This confirms that the main source of uncertainty in the LF measured in this work arises from observational measurement errors.

We have also investigated the contributions of photometric and photo-z errors to the luminosity function (LF) estimation separately. Fig. 6.8 presents the fractional errors, analogous to Fig. 6.7, but distinguishing between photometric and photo-z errors. The results indicate that the photo-z error is roughly an order of

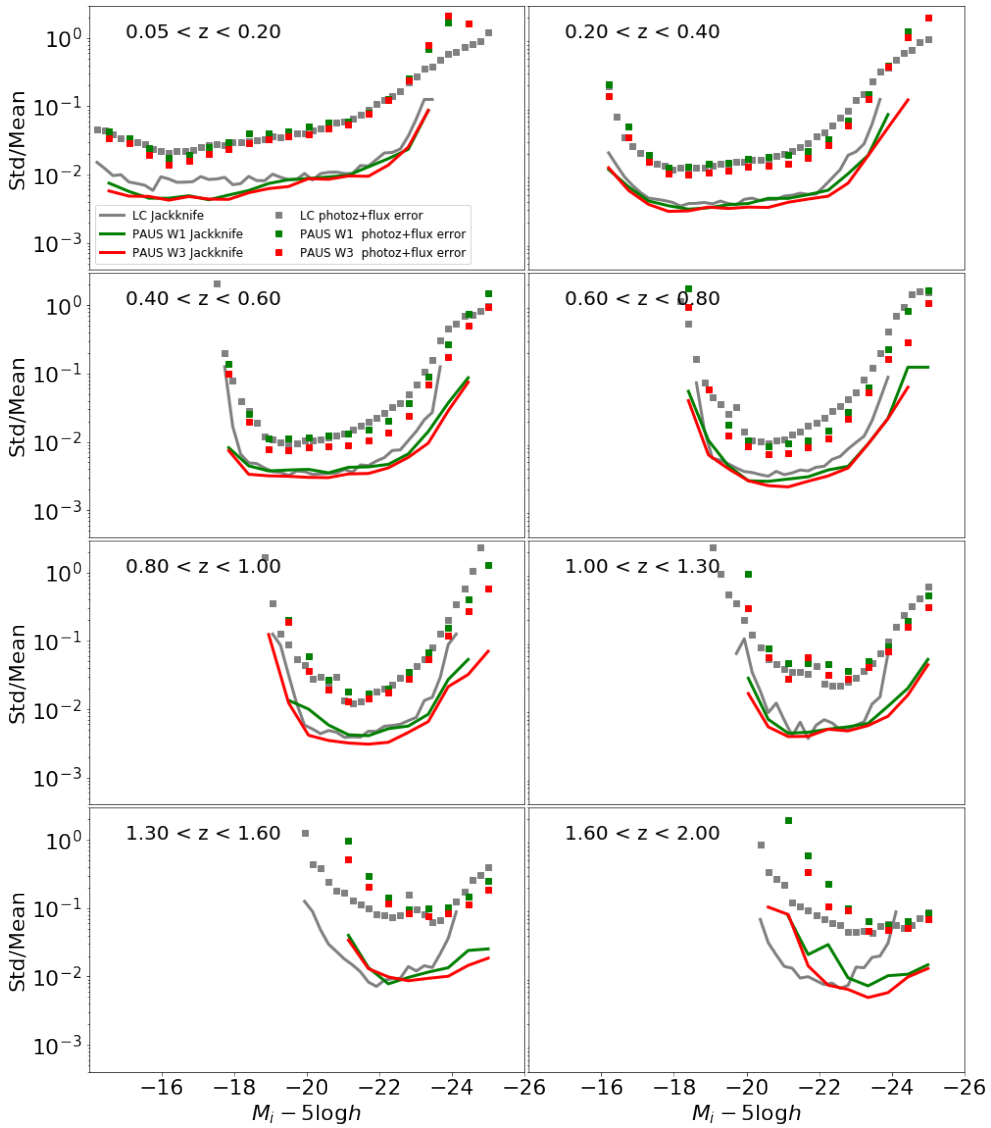


Figure 6.7: The errors in the $i-$band luminosity function estimate from PAUS data in the W1 field (green) and PAUS W3 field (red) compared to that from the `GALFORM` lightcone mock catalogue (grey) between redshift $z = 0.05$ and $z = 2.00$. The Jackknife errors are plotted as solid lines. Meanwhile, the errors from photometric and photometric redshift errors are shown as squares.

magnitude larger than the photometric error, while the photometric error contributes to the LF estimate at a level comparable to that of cosmic variance. This demonstrates that photo-z uncertainty is the dominant source of error in our LF estimation.

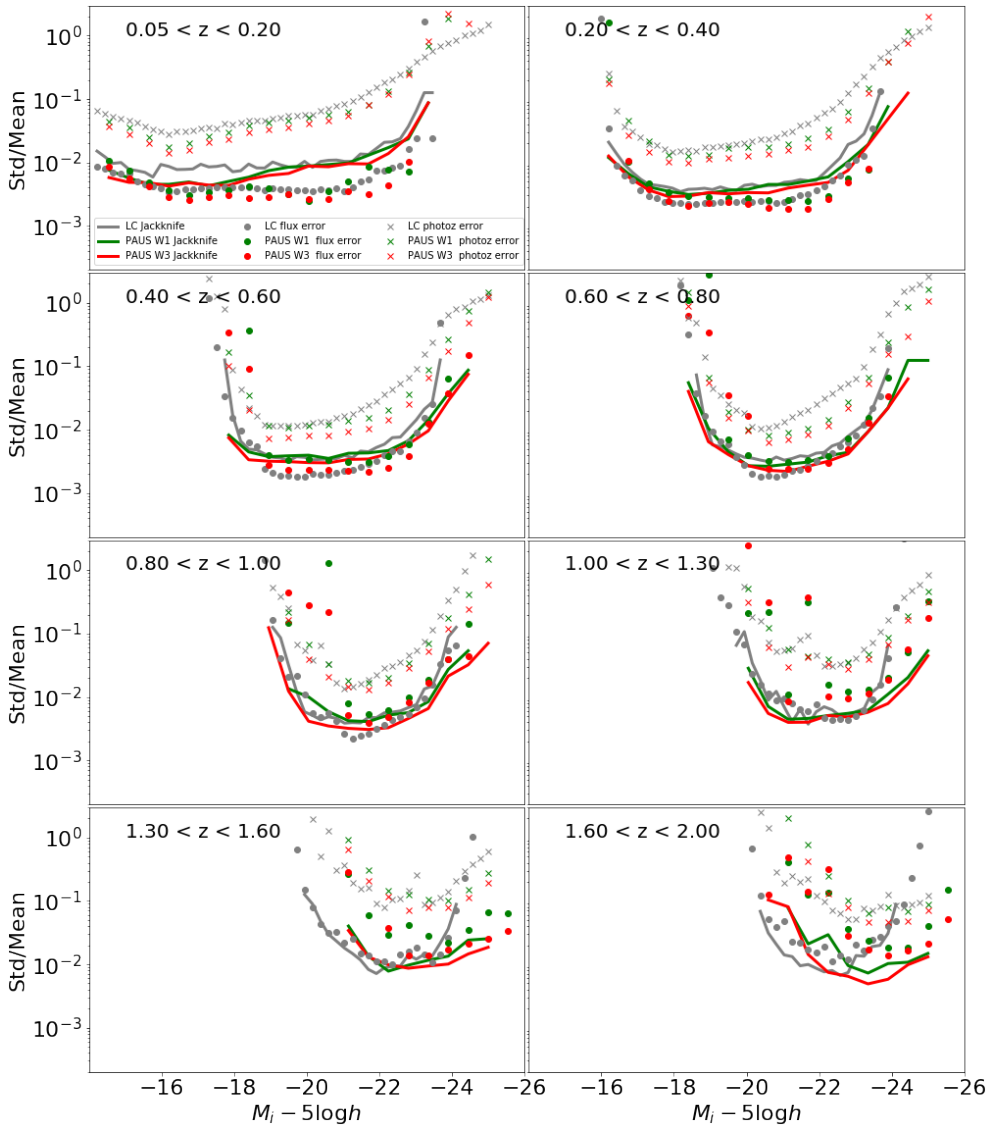


Figure 6.8: The errors in the $i-$band luminosity function estimate from PAUS data in the W1 field (green) and PAUS W3 field (red) compared to that from the `GALFORM` lightcone mock catalogue (grey) between redshift $z = 0.05$ and $z = 2.00$. The Jackknife errors are plotted as solid lines. The errors from photometric fluxes and photometric redshift errors are shown as circles and crosses, respectively.

6.4 Comparison with Previous Estimates of the Luminosity Function

Here we compare our estimate of the i-band luminosity function from PAUS with results from previous surveys. The redshift bins in Fig. 6.9 use intervals to match

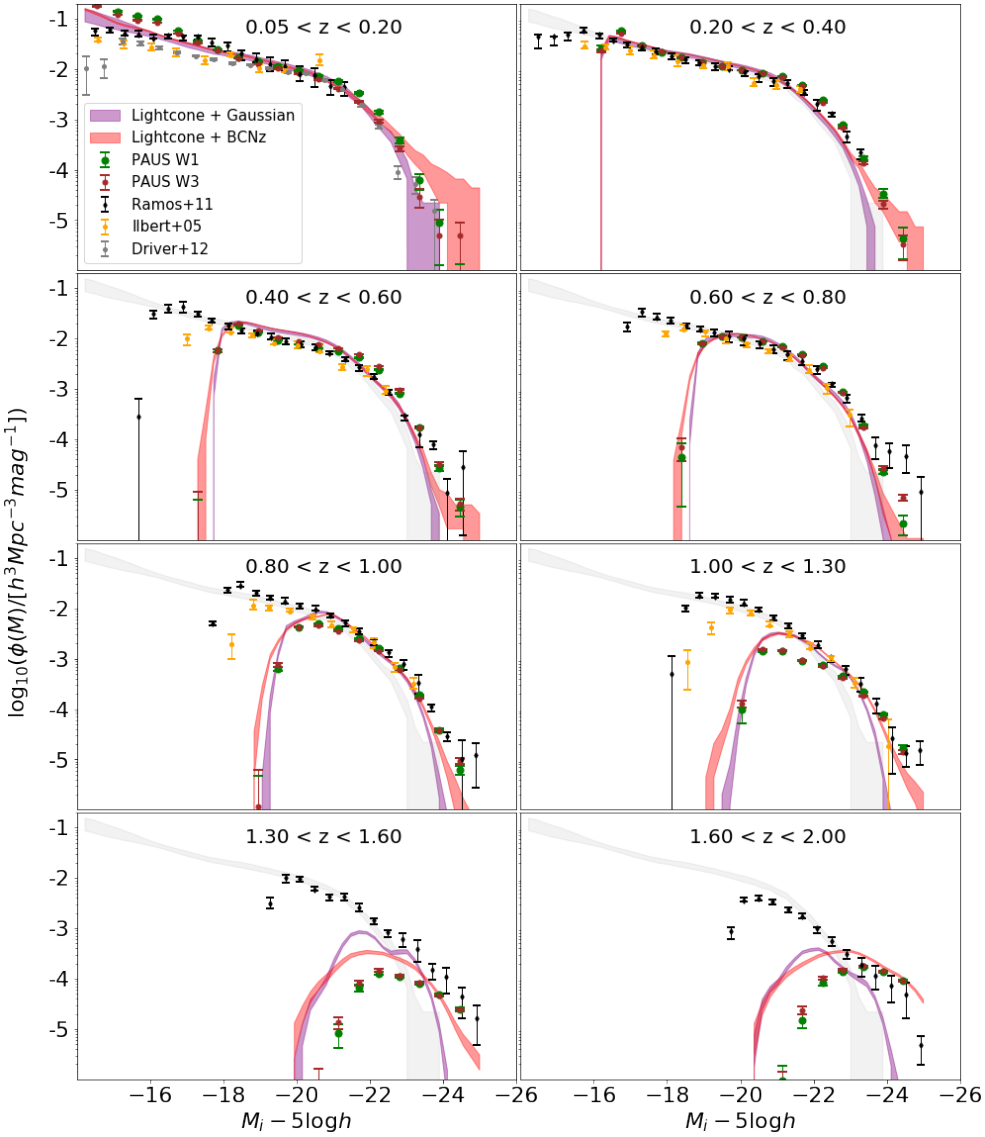


Figure 6.9: The rest-frame $i-$band luminosity function from PAUS data in the W1 field (green dots) and PAUS W3 field (red dots) compared to that from the GALFORM lightcone mock catalogue with the Gaussian-like photometric redshift uncertainties (purple shaded region) and the BCNz-like errors (pink shaded region) between redshift $z = 0.05$ and $z = 2.00$. Other observational estimates of the i-band LF are plotted, as indicated by the key in the top-left panel; see text for a discussion of the comparison with these estimates.

those used in the Ramos et al. (2011) estimates discussed below.

Ilbert et al. (2005) measured spectroscopic redshifts for 11,000 galaxies to $i_{\rm AB} = 24.0$ in the VIMOS VLT Deep Survey. The turnover in the Ilbert et al. LF estimates should appear about 1 magnitude deeper than in the estimates from PAUS. However, the Ilbert et al. estimate is from a relatively small solid angle and so does not extend to as bright a magnitude as the PAUS LF estimates. Also, Ilbert et al. impose a bright magnitude cut of $i_{\rm AB} = 17.5$.

Ramos et al. (2011) measured the LF from the CFHTLS deep fields, covering in total just under 3 square degrees to depths close to $i_{\rm AB} = 26$. These authors use photometric redshifts, derived from the broad band photometry of CFHTLS. Hence the scatter and outlier fraction for their photometric outliers are expected to be larger than for those in PAUS. The Ramos et al. estimates extend to the faintest rest-frame i-band absolute magnitude of the various estimates, as expected from their deeper apparent magnitude limit. At the bright end, the estimates from Ramos et al. are affected by sample variance and the errors in the photometric redshifts.

Finally, Driver et al. (2012) used the Galaxy And Mass Assembly (GAMA) survey to measure the LF in many bands. Driver et al. use spectroscopic redshifts. Their sample is selected in the r band to depths of $r = 19.4$ and $r = 19.8$ depending on the field. We show a low redshift estimate from Driver et al. (top left panel).

The differences in the LFs from the literature and our new estimates can be readily understood at the faint end as being due to the different i-band cuts applied, as outlined above. The differences at the bright end mainly reflect the smaller fields used in Ramos et al. and Ilbert et al., which limit how bright their estimate can reach. At intermediate magnitudes there is reasonable agreement between the different estimates.

6.5 Evolution of the Galaxy Luminosity Function

We now focus on the redshift evolution of the luminosity function derived from PAUS observation. Fig. 6.10 presents the LFs in multiple redshift bins, overplotted in a single panel for each of the two fields (W1: left panel and W3: right panel). Here use curves, with the thickness of each curve reflecting the 1σ uncertainties, to describe the luminosity function.

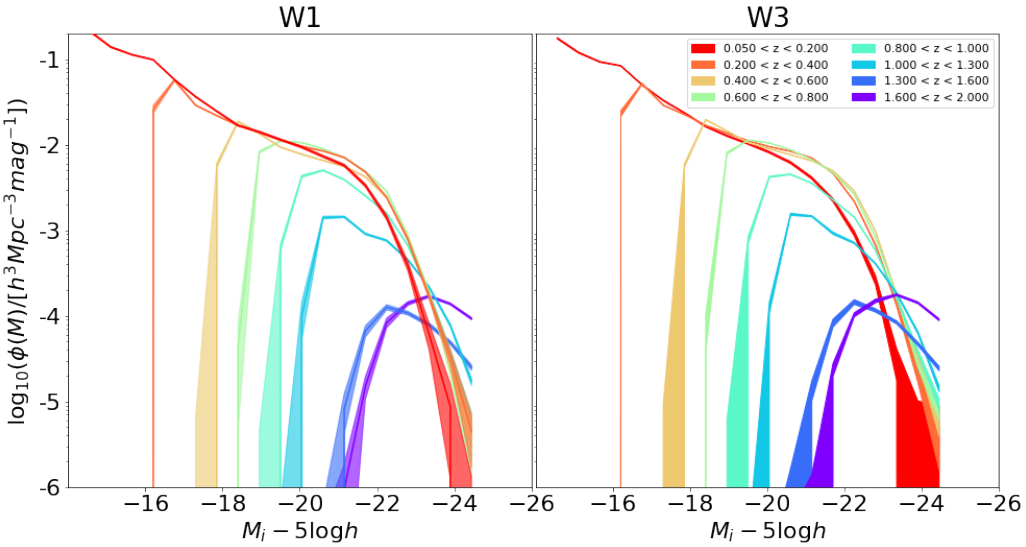


Figure 6.10: The galaxy luminosity function estimated from the PAUS W1 (left) and W3 (right) fields, focusing on the evolution with redshift. The LFs from the different redshift slices are plotted on a single panel for each field, as indicated by the legend in the right panel. Red colours correspond to low redshift and blue colours to high redshift.

The dominant effect with increasing redshift is the progressive shift of the faint-end turnover to brighter magnitudes. This is primarily a selection effect, as the observed i-band selection limit corresponds to increasing brighter rest-frame luminosities at higher redshifts. The turnover also becomes less sharp at high redshift due to the growing offset between observed and rest-frame i-band. Aside from this, the faint-end slope of the LFs shows relatively little evolution, remaining approximately constant across redshifts. There is modest brightening around the characteristic L_*, near $z \sim 0.5$, while at higher redshifts the break at the bright end becomes less distinct. This is likely due to contamination of photo-z outliers.

6.6 The Luminosity Function of Red and Blue Galaxies

To further understand the origin of the evolution observed in the overall luminosity function, we follow Lilly et al. (1995) and examine the LF separately for red and blue galaxy populations. Whereas Lilly et al used the rest-frame colour to label galaxies as red or blue, here we follow Manzoni et al. (2024) and use the observed $(g - r)$ colour. In practice this means that the dividing line between red and blue galaxies is a function of redshift, rather than a constant as would be the case for a rest-frame colour.

Fig. 6.11 shows the i-band LF for red and blue galaxies in redshift slices from $z = 0.05$ to $z = 2.00$. Observational data are shown for both PAUS W1 and W3 fields. We also plot the corresponding predictions from the PAUS mock, with a realistic "BCNz" version that includes them. It is important to note that the redshift error model is based on an i-band selected sample and does not account for colour dependence in redshift error distributions. The "BCNz" version typically produces a modest excess of bright galaxies due to the inclusion of outliers.

Qualitatively, the observed and predicted LFs of red and blue galaxies agree. The turnover at faint magnitudes is arguably less well reproduced for red galaxies than for blue galaxies, with this difference becoming more pronounced at higher redshifts.

6.7 Evolution of the Luminosity Function of Red and Blue Galaxies

To explore the evolution by galaxy colour, Fig. 6.12 and Fig. 6.13 show the redshift evolution of the red and blue galaxy luminosity functions separately, for the W1 and

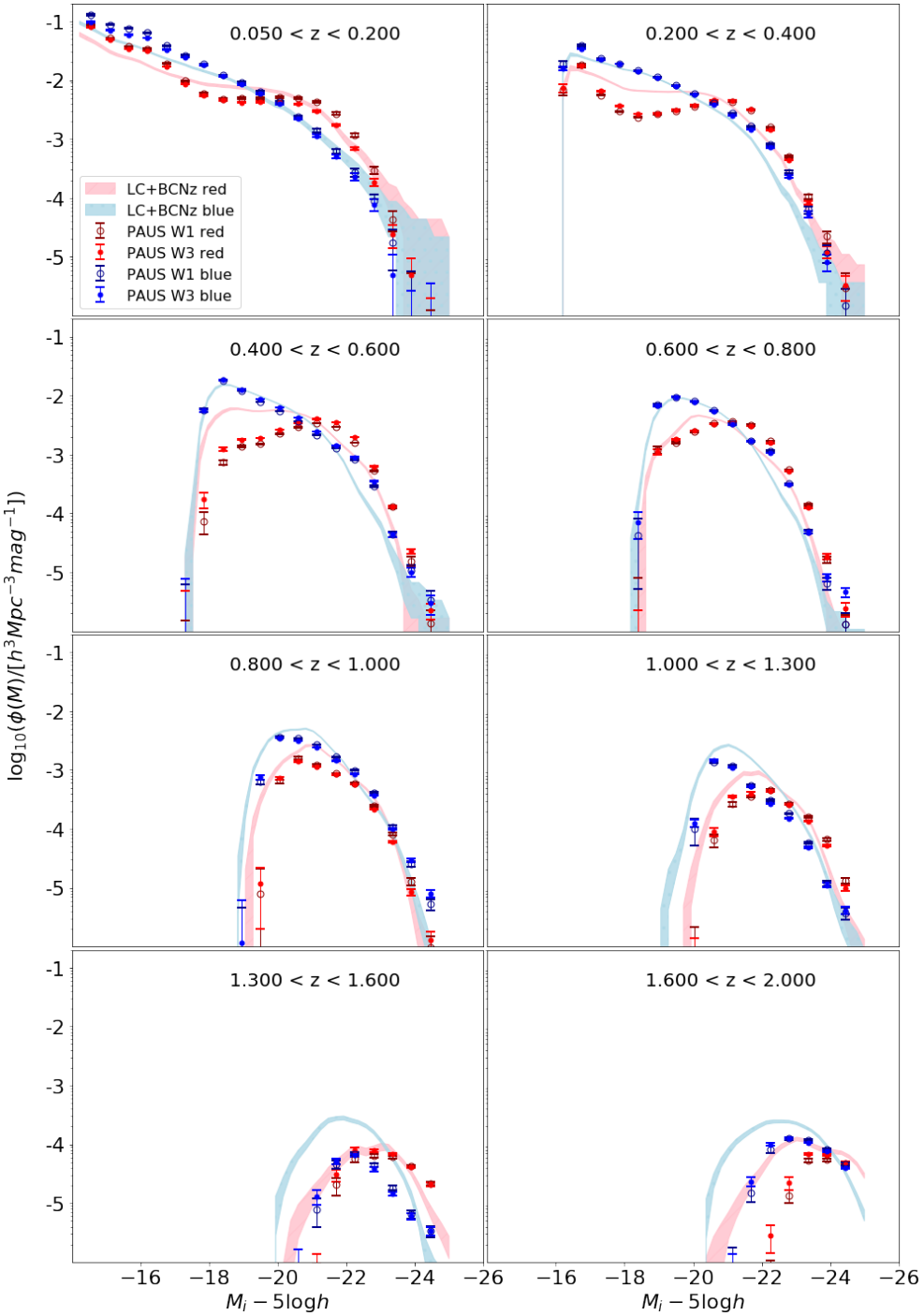


Figure 6.11: The rest-frame i-band luminosity function for red and blue galaxies between redshift $z = 0.05$ and $z = 2.00$. Red and blue galaxy populations are defined based on observed $(g-r)$ colour. Solid points show estimates from PAUS W1 and W3 fields; shaded regions show predictions from the GALFORM mock with photometric redshift errors.

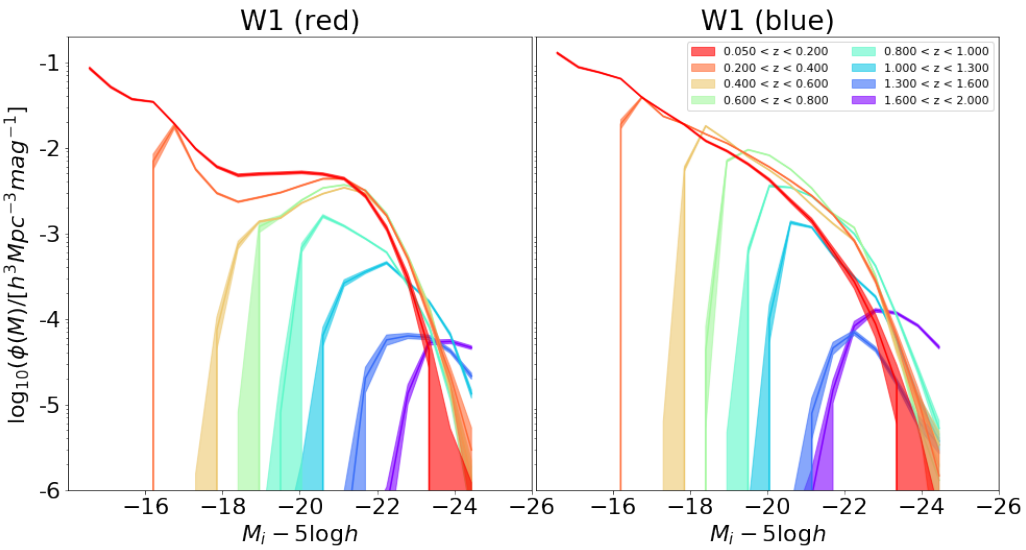


Figure 6.12: The $i-$band galaxy luminosity function for rest-frame red (left panel) and blue (right panel) galaxies for multiple redshift bins ranging from $z = 0.05$ to $z = 2$, as indicated by the legend. Here we focus on the PAUS W1 field. The observational estimates are shown by curves rather than symbols with the width of the curve indicating the $1 - \sigma$ error on the measurement.

W3 fields, respectively. Again, we use curve thickness to represent the measurement uncertainties.

In both fields, the faint end of the blue LF retains an approximate power-law shape and shows little evolution at low redshift. In contrast, the red LF shows stronger evolution at the faint end, with the slope becoming shallower over time. At intermediate and high redshift, both red and blue LFs exhibit a shift in the position of L_* with a stronger shift for blue galaxies. At highest redshifts, the sharpness of the LF break reduces, particularly for red galaxies, consistent with the increasing effect of photo-z errors and sample incompleteness.

Overall, these results demonstrate that the evolution of the galaxy luminosity function from PAUS is driven by a combination of intrinsic population changes (especially for red galaxies), photometric depth limitations, and redshift error effects. These measurements also reproduce the expected trend that blue galaxies dominate the faint end, while red galaxies contribute significantly to the bright end, particularly at low redshift.

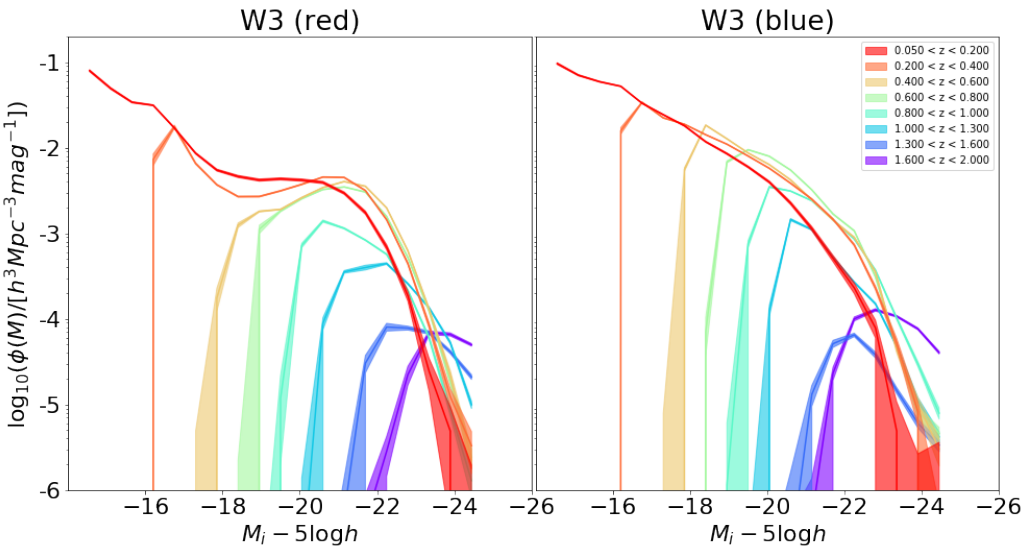


Figure 6.13: The $i-$band galaxy luminosity function for rest-frame red (left panel) and blue (right panel) galaxies for multiple redshift bins ranging from $z = 0.05$ to $z = 2$, as indicated by the legend. Here we focus on the PAUS W3 field. The observational estimates are shown by curves rather than symbols with the width of the curve indicating the $1 - \sigma$ error on the measurement.

6.8 The Completeness and redshift quality

We now return to the overall galaxy luminosity function and assess the completeness and systematic effects that may influence our estimates

One immediate factor to consider is any incompleteness arising from the requirements imposed on the photometric redshift estimation. A galaxy must be imaged in a sufficient number of narrow-band filters–at least 30 out of 40–to obtain a reliable photometric redshift. During the earlier stages of the PAUS survey, when the mosaic of camera pointings was still being built up, different regions of the sky had varying levels of filter coverage due to partial exposure with specific trays. This raised concerns about potential incompleteness in the redshift sample.

This effect was evident in Fig. 2 of Manzoni et al. (2024), where the number counts of galaxies with more than 30 narrow-band detections were approximately 90% of the full galaxy sample. Initially, this implied that the LF normalisation might need to be adjusted upward by a factor of 1/0.9. However, further inspection

shows that this reduction was primarily due to a different in the **effective survey area**–not a deficiency in the galaxy sample itself. As shown in Fig. 2.12 and confirmed by the number count comparison in Fig. 2.13, the surface density of galaxies with more than 30 NB detections agrees with that of the full sample once the correct areas are used to normalise the counts. Hence, the normalisation of the luminosity function is based on the correct effective area, and no further correction (e.g. the previous assumption of 1/0.9 factor) for incompleteness is required.

Next we investigate the distribution and mean value of $V/V_{\rm max}$ in each redshift shell. If the sample is a fair sample of the Universe, the mean value of $\langle V/V_{\rm max}\rangle = 0.5 \pm 1/\sqrt{12N}$ where N is the total number of galaxies in the sample (Peacock, 1999). This error is for a uniform random distribution of points. For $N = 20,000$, the random uncertainty is only $\sim 2 \times 10^{-3}$. However, the mean values reported in Fig. 6.14 differ from 0.5 by much larger amounts, suggesting significant deviations beyond what would be expected from random sampling alone.

The W3 field is just under twice the size of the W1 field. Fig. 6.14 shows large differences in the mean $V/V_{\rm max}$ values returned for these fields. This could reflect differences due to large-scale structures, which can be seen from the spikes in the redshift distribution in Fig. 4.7. Nevertheless, the estimate of the LF from the W1 and W3 fields agree quite well with one another.

Another issue could be errors in the k-correction. These could lead to gradients in the $V/V_{\rm max}$ distribution. The k-correction does not affect the $V_{\rm max}$ value for most galaxies, as the maximum volume is simply the volume of the redshift slice used to measure the luminosity function.

Finally, the $V/V_{\rm max}$ could be affected by evolution in the galaxy population. However, we expect this to be small as the redshift shells used to the measure the LF are small, and the LF does not evolve substantially between shells.

Another consideration is the impact of what are believed to be low quality photometric redshifts on the recovered LFs. We have already argued that outliers

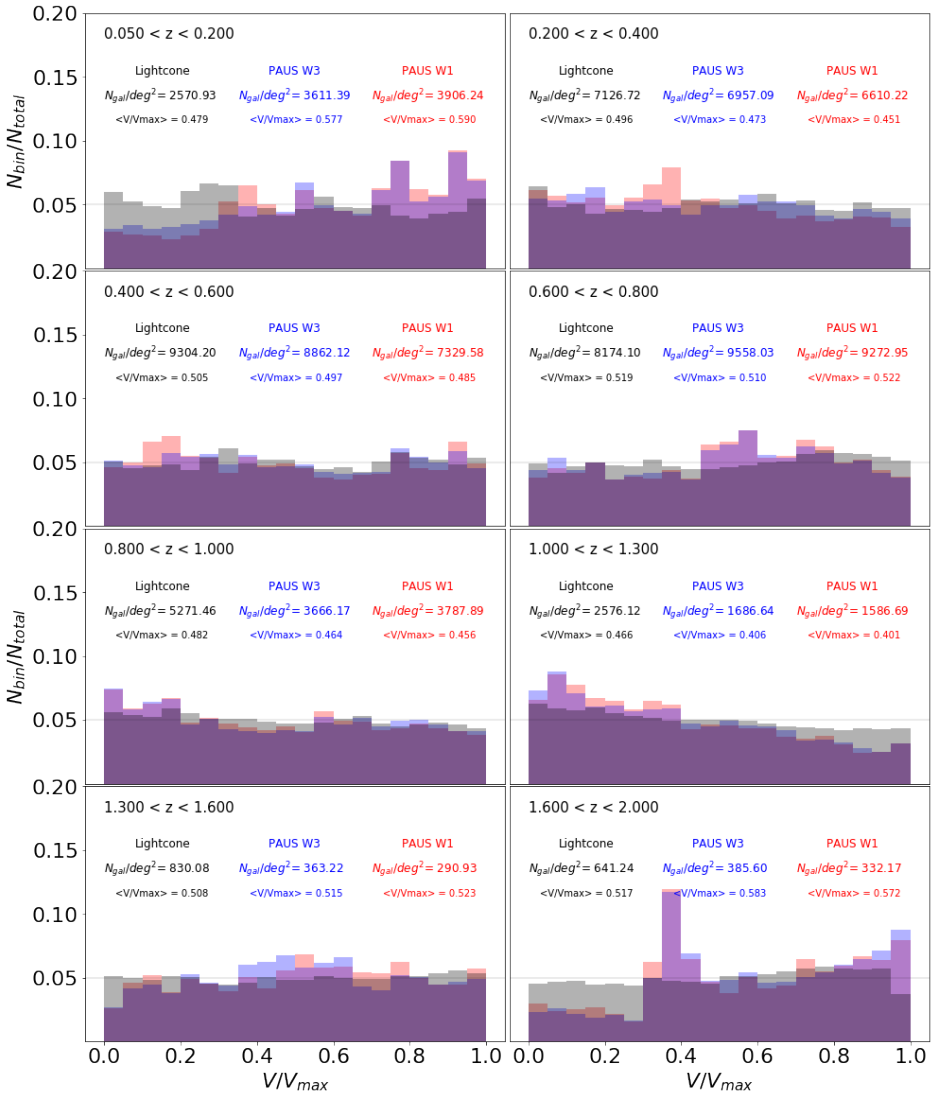


Figure 6.14: The distribution of $V/V_{\rm max}$ for different redshift bins, as labelled in each panel. Distributions are shown for the mock catalogue (grey histogram), without any errors in the photometry or photometric redshifts of the model galaxies, the PAUS W3 (blue histogram) and W1 (red histogram) fields. The labels give the number of galaxies per unit area in each case and the mean value of $\langle V/V_{\rm max}\rangle$.

affect the shape of the bright end of the LF at high redshift, by applying different scenarios for photo-z errors to the mock galaxies. Explicitly including redshift outliers changes the shape of the bright end of the LF recovered from the mocks and makes it look more similar to the LF recovered from the observations.

We can isolate the impact of low-quality photo-z by applying cuts on the quality factor (as defined in Equation 16 of Eriksen et al. 2019) and recomputing

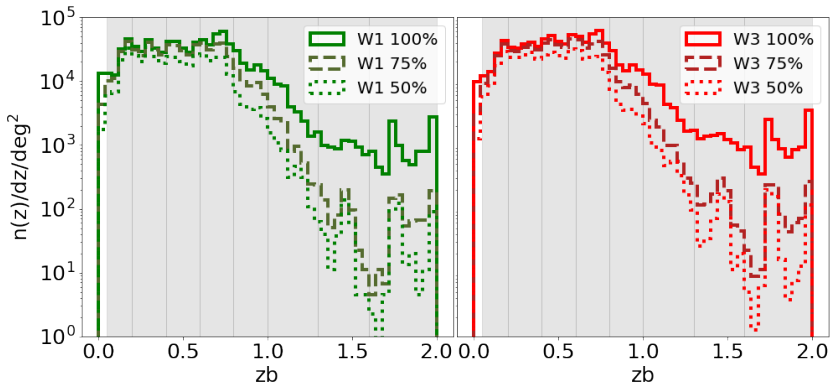


Figure 6.15: The redshift distributions in the PAUS W1 (left) and W3 (right) fields. The solid line histogram shows the distribution for all galaxies brighter than $i_{\rm AB} = 23.0$. The dashed line shows a cut on quality factor which retains the 75 per cent of galaxies with the best photometric redshifts and the dotted line shows the distribution for the best 50 per cent of redshifts.

both the redshift distribution and the luminosity function. Fig. 6.15 shows how these quality cuts affect the redshift distribution of galaxies in both the W1 and W3 fields. Retaining only best 75% or 50% of photo-z estimates results in a sharp drop in galaxy counts at high redshift. This is consistent with the redshift dependence of the outlier rate seen in Fig. 2.5, where the fraction of outlier photo-z increases significantly beyond $z \sim 1.2$.

The form of the redshift distribution changes substantially with quality cuts– not just in amplitude, but in shape. To test whether this effect count be accounted for by a simple rescaling, we apply normalisation factors to the redshift distributions of the 75% and 50% subsamples. Fig. 6.16 shows the result: even after multiplying the counts by constant factors (4/3 for best75, 2 for best50), the redshift distributions remain systematically suppressed at high redshift compared to the full sample.

To further quantify the impact of the quality cut on the redshift distributions, we performed the Kolmogorov-Smirnov (K-S: Massey 1951) test between samples with different cuts. When comparing two large samples, the K-S test often rejects the null hypothesis–that the two samples are drawn from the same distribution– even when the distributions are visually quite similar. As shown in Fig. 6.17, the

cumulative redshift distributions of galaxies with different quality cuts differ only slightly by eye, suggesting that they could possibly be drawn from the same parent distribution. However, the K-S test returns very low p-values (p-value $\ll 0.5$) for every pairwise comparison, as listed in Table 6.1, leading to systematic rejection of the null hypothesis. For samples restricted to higher redshift selections, the p-values even drop to zero in all cases. We reach the same conclusion when applying the Kuiper's test (Kuiper 1960). These results highlight a limitation of applying such statistical tests to very large samples: they are overly sensitive to small deviations and may not be effective as a qualitative diagnostic. In contrast, visual inspection more clearly shows the divergence in shape at the high-redshift end.

Overall, this analysis confirms that the impact of the quality cuts cannot be compensated for by adjusting the overall normalisation. Instead, the redshift-dependent incompleteness indicates that galaxies with low-quality redshifts are preferentially located at high redshifts and are thus disproportionately removed by quality cuts.

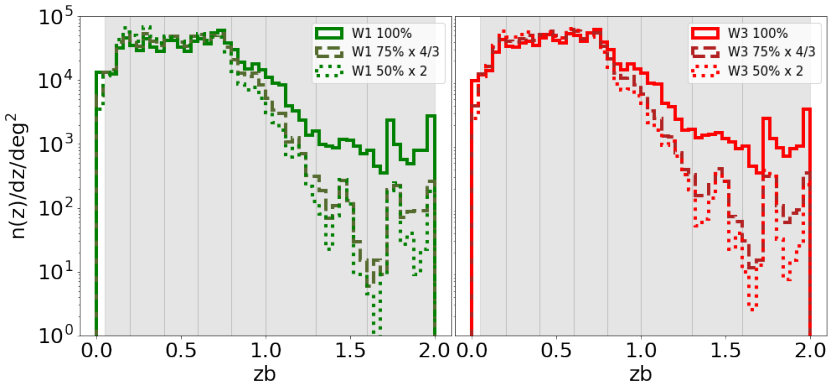


Figure 6.16: The redshift distributions in the PAUS W1 (left) and W3 (right) fields after applying simple normalisation factors. The solid line histogram shows the distribution for all galaxies brighter than $i_{\rm AB} = 23.0$. The dashed line shows the distribution of the best 75 per cent of photometric redshifts (scaled by a factor of 4/3), and the dotted line corresponds to the best 50 per cent (scaled by a factor of 2). Despite this normalisation, a redshift-dependent suppression remains, particularly at high redshift.

The corresponding effect on the luminosity function is shown in Fig. 6.18. Removing the worst 25% or 50% of galaxies based on photo-z quality results in noticeable suppression of the LF, especially at faint magnitudes and higher redshifts.

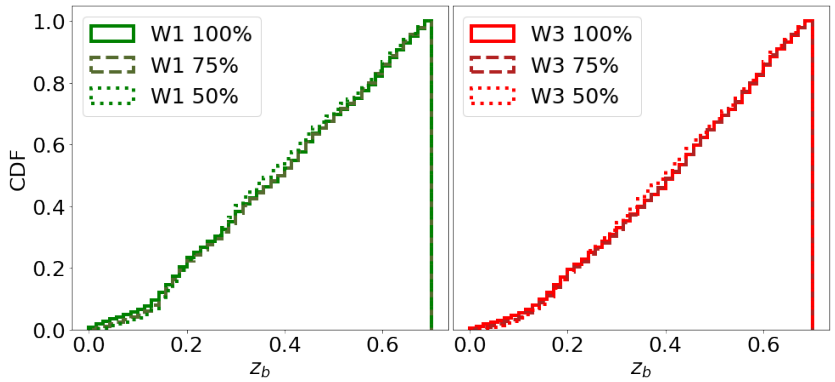


Figure 6.17: The cumulative distribution of the redshifts for PAUS W1 (left panel) and W3 (right panel) fields for galaxies with redshifts $z < 0.7$ with different photometric redshift quality cuts (solid: no quality cut, dashed: best 75%, and dotted: best 50%, corresponding to the redshift distribution in Fig 6.16).

K-S test: Pairs ($z < 0.7$)	p-value	
	W1	W3
Best 100 vs Best 75	2.16×10^{-33}	9.84×10^{-33}
Best 100 vs Best 50	6.45×10^{-99}	3.81×10^{-150}
Best 75 vs Best 50	1.86×10^{-90}	2.02×10^{-132}

Table 6.1: The p-value of the K-S test for each sample comparison for each field. Only galaxies with $z < 0.7$ are selected for the test, see Fig 6.17.

These differences become particularly significant from $z \sim 0.8$ onward. This trend is expected, since the galaxies with the lowest signal-to-noise ratios (and hence lower redshift quality) tend to be the faintest ones in a given redshift bin. While the faint end is most affected, there is also a modest reduction at the bright end. Importantly, these effects are not correctable by a simple amplitude rescaling of the LF–further emphasising the need to understand and model photometric redshift uncertainties and completeness in a redshift-dependent way.

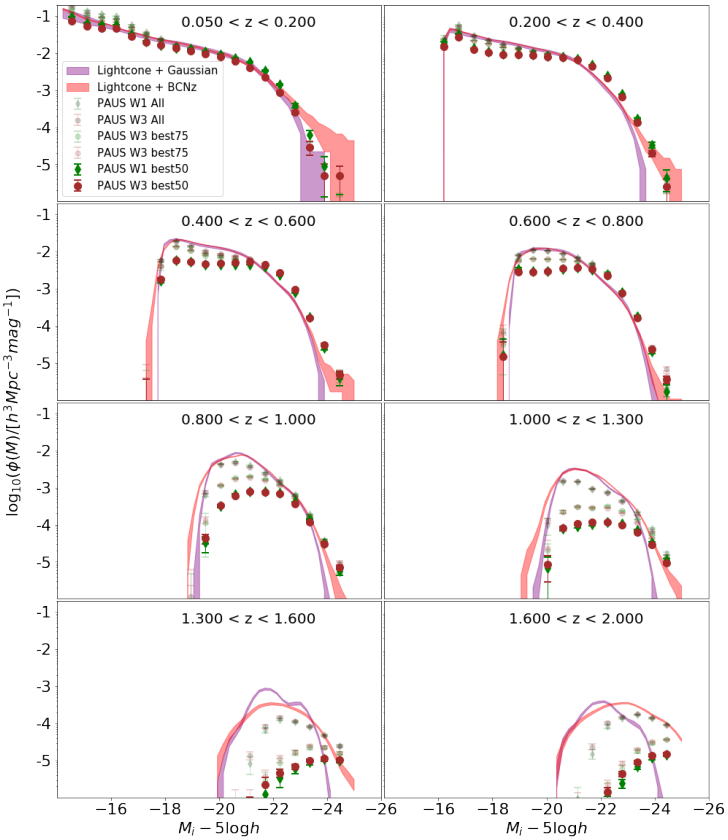


Figure 6.18: The $i-$band galaxy luminosity function for multiple redshift bins ranging from $z = 0.05$ to $z = 1.95$. The LF is estimated for different subsamples of galaxies with $i_{\rm AB} = 23.0$: all galaxies, the best 75 per cent of photometric redshifts as ranked by the quality factor and the best 50 per cent (i.e. discarding half of the galaxies). The LFs estimated from the lightcone are also shown, for all galaxies, but with different models for the errors in the photometric redshifts: Gaussian error (purple), BCNz-like errors, including outliers (red).

6.9 The Luminosity Function in ugrz- Filters

To complement our analysis in the i-band, here we present luminosity functions estimated in the u, g, r, and z bands. This multi-band view allows us to explore how the shape and evolution of the LF vary with wavelength. As in earlier sections, we compare the observational estimates from the PAUS W1 and W3 fields with predictions from the `GALFORM` mock catalogue. Both Gaussian and `BCNz2`-like photo-z error models are shown.

6.9.1 u-band

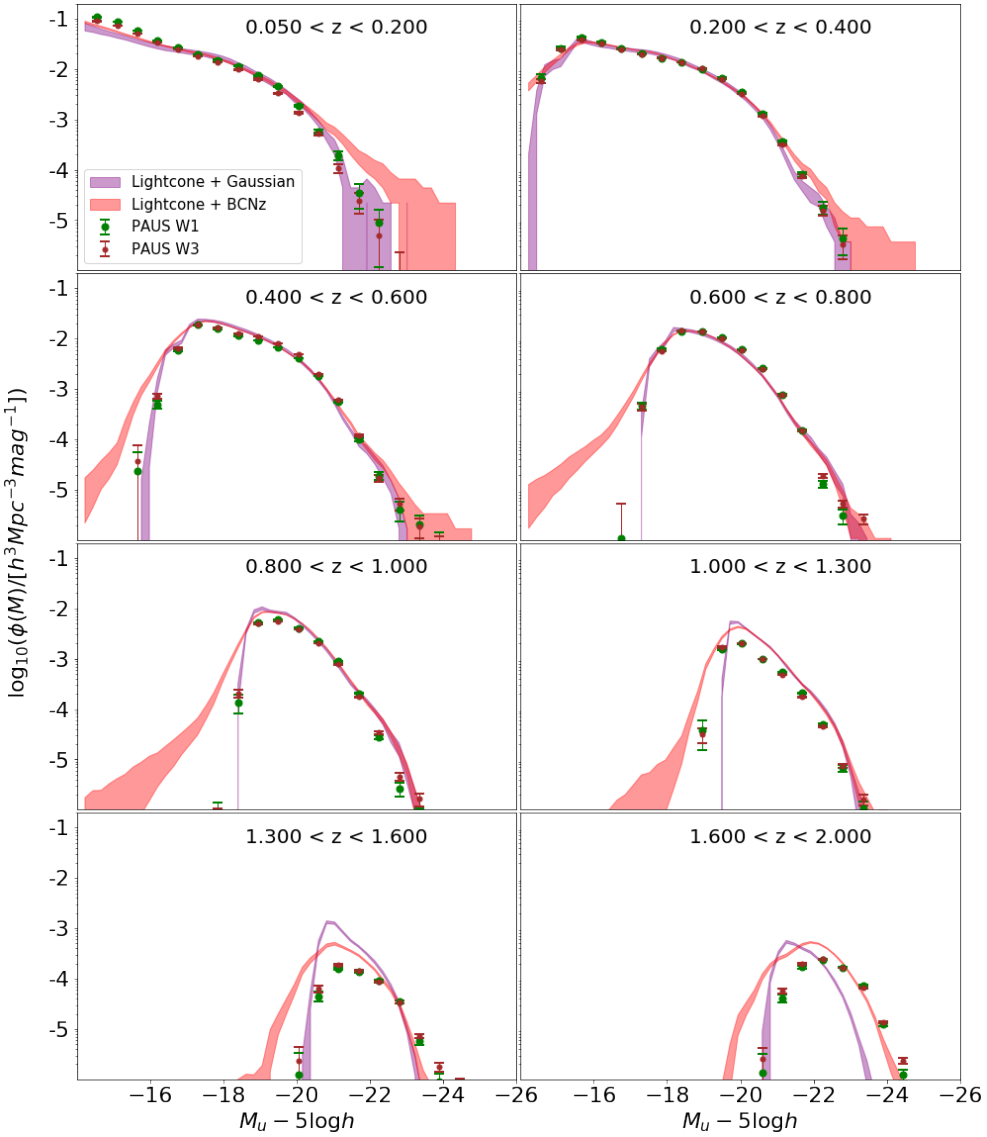


Figure 6.19: The rest-frame u-band luminosity function from PAUS data (green: W1, red: W3) compared to the `GALFORM` lightcone predictions with Gaussian (purple) and `BCNz2`-like (pink) photo-z uncertainties.

The u-band LF (Fig. 6.19) probes the Ultraviolet (UV)-optical transition and is particularly sensitive to recent star formation. As expected, the faint end is noisier and less complete, especially beyond $z \sim 1$ due to the increasing impact of the apparent magnitude limit and the blueward shift of the observed bandpass. Nonetheless, the observed and predicted LFs agree well over the redshift range

$z < 1$. At higher redshift, the bright end appears systematically overestimated in the GALFORM mock when using the BCNz-like photo-z errors.

6.9.2 g-band

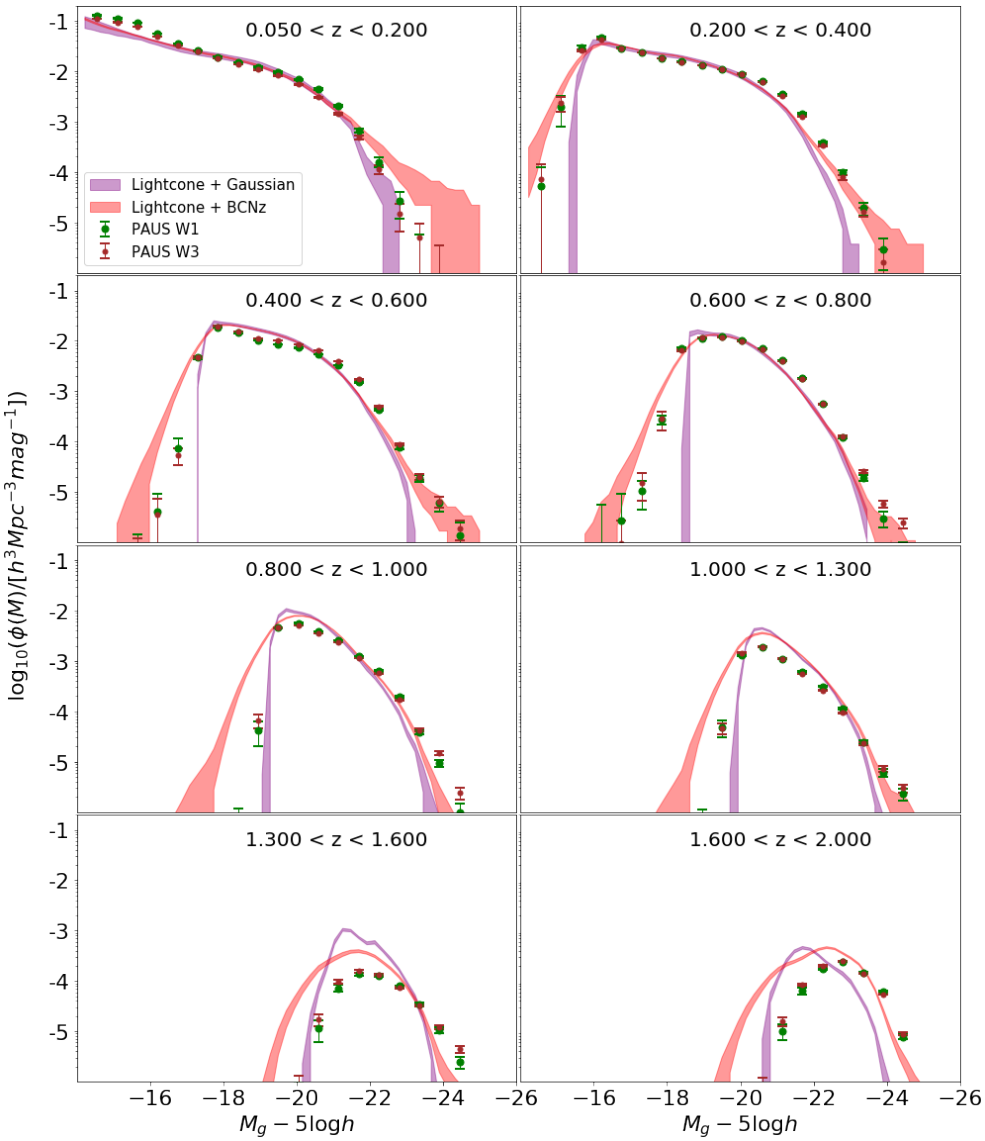


Figure 6.20: Same as Figure 6.19, but for the rest-frame g-band.

The g-band (Fig. 6.20) captures both young stellar populations and evolved stars, offering a cleaner LF than in the u-band. The agreement between the PAUS estimates and the GALFORM mock is excellent up to $z \sim 1$, both in shape and normalisation. Beyond this redshift, small deviations appear at the bright end,

again linked to the effect of redshift outliers. Compared to the i-band, the turnover at the faint end emerges at slightly brighter absolute magnitudes due to increased k-corrections in the blue bands.

6.9.3 r-band

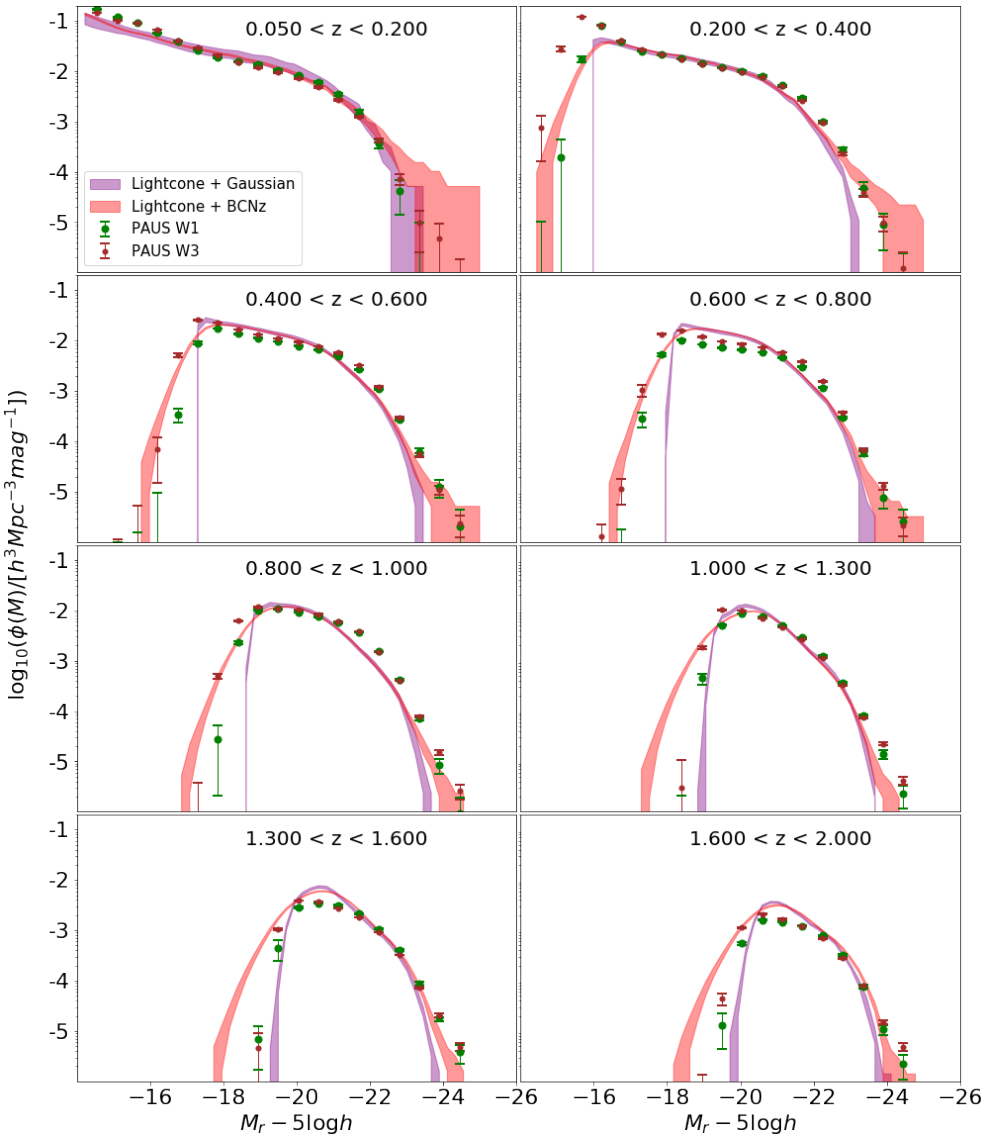


Figure 6.21: Same as Figure 6.19, but for the rest-frame r-band.

The r-band (Fig. 6.21) shows similarly strong agreement between observation and model. The bright end remains well constrained up to $z \sim 1.3$. The smooth-

ness of the LF at intermediate magnitudes reflects the increased completeness and photometric accuracy in this band.

6.9.4 z-band

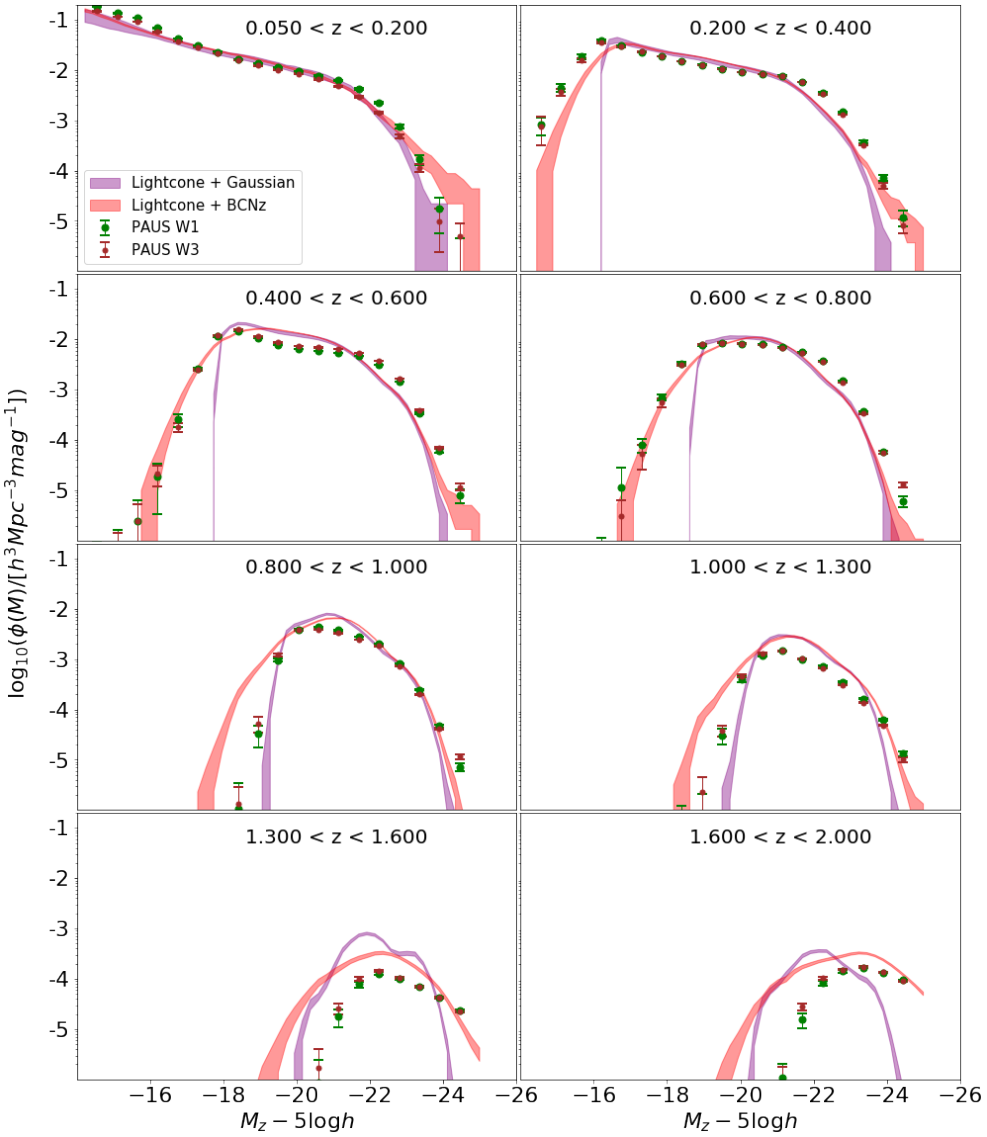


Figure 6.22: Same as Figure 6.19, but for the rest-frame z-band.

The z-band (Fig. 6.22) LF extends the analysis to redder wavelengths, probing older stellar populations and being less sensitive to recent star formation. Up to $z = 1$, the agreement between PAUS and the GALFORM predictions is excellent. At the highest redshifts, the effects of redshift scatter and the increased incompleteness

at faint magnitudes are again evident. The overall LF shape in the z-band evolves more slowly than in the g-bands, consistent with the expectation that older stellar populations dominate this part of the SED.

Together, these $ugrz$-band luminosity functions confirm the robustness of the LF estimates from PAUS and further validate the predictions from the `GALFORM` model across the optical range. Differences at the faint end are driven by varying completeness limits and photometric uncertainties, while bright-end discrepancies remain dominated by photo-z outliers. The consistency between bands suggests that selection effects and systematic uncertainties are well controlled in our analysis.

6.10 Conclusion

In this chapter, we have presented a comprehensive measurement of the galaxy luminosity function in the PAUS W1 and W3 fields using rest-frame magnitudes derived from the random forest regression technique. Our analysis focuses primarily on the i-band LF, but we extended the results across the $urgz$ bands and explored subsamples defined by colour and redshift quality.

We first described the methodology used to estimate the LF based on the $1/V_{\rm max}$ technique, and validate this approach using the `GALFORM` lightcone mock catalogue. This mock provided a consistent baseline to evaluate selection effects, observational errors, and the impact of photometric redshift uncertainties. By applying different photo-z error models–including Gaussian and more realistic `BCNz`-like errors–we showed how redshift outliers can affect the bright end of the LF, especially at high redshift.

The rest-frame i-band LF from PAUS is in excellent agreement with the `GALFORM` prediction out to $z \sim 1$. At higher redshifts, we identified departures at both the bright and faint ends of the LF, particularly for red galaxies. These discrepancies correlate with increasing photo-z errors and growing incompleteness in faint galaxy populations. We also analyses the LF for red and blue galaxies separately,

and showed that their evolution follows distinct paths: blue galaxies maintain a steep faint-end slope, while red galaxies exhibit a more rapid decline and a stronger dependence on redshift.

We further assessed the completeness of the sample by examining $V/V_{\max}$ distributions and photometric redshift quality. We found that apparent magnitude cuts and selection based on narrow-band coverage are properly accounted for when normalising by the correct effective area. However, quality cuts on photo-z estimates introduce redshift-dependent biases, which cannot be corrected by simple global scaling. This underscores the importance of modelling photometric redshift uncertainty when interpreting galaxy statistic from photometric redshift surveys.

Finally, we extended the LF measurement to the u, g, r, and z bands. These multi-band LFs show excellent internal consistency and further validate the reliability of the PAUS data. Across all bands, the key trends observed in the i-band–particularly the influence of redshift outliers and faint-end incompleteness–are seen to persist, reinforcing the need for careful modelling of selection effects.

This chapter lays the foundation for subsequent analysis of galaxy stellar masses and stellar mass functions in the next chapter.

Chapter 7

The Prediction of Galaxy Stellar Masses using Broad Band Photometry

7.1 Overview

Stellar mass is one of the most fundamental properties of a galaxy, reflecting its star formation history and providing key insights into the physical processes that govern galaxy formation and evolution. Accurate and efficient estimation of stellar masses is therefore essential for constructing stellar mass functions (SMFs), investigating galaxy populations across cosmic time, and testing theoretical models of galaxy formation.

Traditionally, stellar masses are estimated through spectral energy distribution (SED) fitting techniques, which compare observed photometry and sometimes spectra with synthetic models derived from stellar population synthesis codes. While widely used, these methods require assumptions about star formation histories, metallicities, dust attenuation, and stellar initial mass functions, and can be computationally expensive for large surveys. They also offer limited interpretability, with the relationship between input photometric features and the derived stellar

mass often obscured by complex modelling choices.

In this chapter, we explore an alternative approach based on symbolic regression– a machine learning technique that aims to find compact analytical expressions that best describe the relationship between input features and target variables. Specifically, we use the symbolic expressions developed by Kumar et al. (in prep), who trained models on a perturbed version (i.e. including observational and estimation errors) of the `GALFORM` lightcone mock catalogue to estimate stellar masses from a limited set of broad-band photometric features. These expressions provide a transparent and computationally efficient method for stellar mass estimation, directly linking observed magnitudes and colours to stellar mass through closed-form equations.

My role in this collaborative work focused on preparing the input catalogue for model training and validation, based on the mock galaxy sample described in Chapter 4. I applied photometric and photometric redshift uncertainties to simulate PAUS-like observational conditions, and computed rest-frame magnitudes for use in computing volumes for the $1/V_{\rm max}$ method. I then applied Kumar et al.'s symbolic regression models to both the lightcone mock and the PAUS observational dataset to predict stellar masses and compare them against those obtained using the `CIGALE` SED fitting method (Csizi et al., 2024). The resulting stellar mass functions were used to assess the consistency between theoretical predictions and observations.

The structure of this chapter is as follows: § 7.2 describes the preparation of the mock and observational datasets used in this work, along with the symbolic regression methodology. § 7.3 presents a comparison between the predicted stellar masses and those derived from SED fitting for PAUS galaxies. In § 7.4, we compute stellar mass functions using the predicted masses and compare results between the mock and observational data. § 7.5 summarises the main findings and outlines their relevance for future galaxy formation studies.

7.2 Data Preparation and Methodology

This work makes use of the `GALFORM` lightcone mock catalogue developed by Manzoni et al. (2024), which was described in detail in Chapter 4. To facilitate a fair comparison with the Physics of the Accelerating Universe Survey (PAUS) observations, introduced in Chapter 2, we applied the survey-specific photometry and photometric redshift uncertainties to the mock data following the procedures outlined in § 4.6 in Chapter 4. These included Gaussian perturbations to broad-band and narrow-band fluxes, and `BCNz2`-like photometric redshift errors to the redshifts of mock galaxies.

The machine learning framework used to predict stellar masses from galaxy observables was developed by Kumar et al. (in prep), who trained a suite of regression models on the lightcone catalogue to infer stellar masses from a limited set of photometric features. My role in this project was to prepare the input dataset for model training by selecting galaxies from the lightcone with $i_{\rm AB} < 22.5$ and within the redshift range $0.00 < z < 2.00$. Galaxies were further classified into red and blue types based on the observer-frame colours, using the colour-redshift separation defined by Manzoni et al. (2024).

For later $1/V_{\rm max}$ calculations, rest-frame i-band absolute magnitudes were also predicted for each galaxy using the random forest regression method described in Chapter 5, which is used to estimate the k-correction based on the galaxy's photometry and redshift. These rest-frame magnitudes were used to compute magnitude limits in the estimation of the SMF.

The symbolic regression expressions derived by Kumar et al. were then applied to both the mock and PAUS catalogues to infer galaxy stellar masses. The final form of these expressions depends on a small set of input features: redshift $z_{\rm obs}$, observed u- and i-band magnitudes, and the $(g - r)$ colour, making them directly application to both mock and real observational data. The expressions are as

follows:

Red galaxies:

$$\log(M_*/h^{-1}M_\odot) = 2.738 + 0.004 \times u + 0.811 \times (g-r) - 0.334 \times i - 0.498 \times z_{\rm obs}, \quad (7.1)$$

Blue galaxies:

$$\log(M_*/h^{-1}M_\odot) = 3.524 + 0.208 \times u + 0.114 \times (g-r) - 0.493 \times i - 0.195 \times z_{\rm obs}, \quad (7.2)$$

where M_* is the stellar mass.

7.3 Application to PAUS Data

To evaluate the performance of the machine learning-based stellar mass predictions on real observational data, we applied the symbolic regression expressions derived by Kumar et al. (in prep) to the galaxy catalogue from PAUS. The input features required for these predictions–photometric redshift $z_{\rm obs}$, observed u- and i-band magnitudes, and the observer-frame $(g-r)$ colour–were obtained from the PAUS production photometry.

Rest-frame magnitudes for PAUS galaxies were computed using the random forest regression method described in Chapter 5, consistent with the approach used for the lightcone mock. Galaxies were classified into red and blue types using the same colour-redshift relation as described in Chapter 4. The appropriate symbolic regression formula was then applied to estimate the stellar mass for each galaxy.

To assess the validity of these machine learning-based predictions, the resulting stellar masses were compared to those derived from spectral energy distribution (SED) fitting using the `CIGALE` code. These `CIGALE`-based stellar masses were provided as part of the official PAUS production data. The comparison was performed separately for red and blue galaxies in the W1 and W3 fields.

Fig. 7.1 shows the difference between the predicted stellar mass (from Kumar's expressions) and the `CIGALE` SED-fitted stellar mass, as a function of the reference

stellar mass $\log M_{*,\mathrm{ref}}$, for red and blue galaxies in both fields. The reference stellar masses correspond to the CIGALE SED-fitted stellar mass for the PAUS galaxies and the "true" mock stellar masses for the GALFORM lightcone mock galaxies, respectively. The 2D density histograms show the distribution of individual galaxies from the PAUS data, while the black solid lines indicate the running median trend, with the shaded grey region showing the interquartile (25th-75th percentile) range. The red and blue solid lines represent the mock validation for red and blue galaxies, as discussed in Kumar et al (in prep). The general agreement between the two methods is good in the mass range $10^8 < M_*/h^{-1}M_\odot < 10^{11.5}$, with systematic offsets generally below 0.1 dex. Discrepancies become more significant at the low-mass end, where PAUS photometry is less complete, and at the high-mass end, where sample statistics are limited.

These results show that Kumar et al.'s symbolic regression model provides a fast and reasonably accurate approximation to SED-based stellar mass estimates, with significant advantages in computational speed.

How closely should we expect the two estimates of the stellar masses to agree? The CIGALE and GALFORM models both assume solar neighbourhood stellar initial mass functions. Nevertheless, there are sufficient differences between the forms of these IMFs to contemplate applying a correction. For the case of long-lived stars, galaxies in GALFORM, for which a Kennicutt (1998) IMF is assumed, need to have their stellar mass reduced by a factor of 0.81 to compare directly with the masses estimated using CIGALE (see Table B1 from Lacey et al. 2016). This factor alone brings the red galaxies into better agreement for the two approaches to inferring the stellar mass. Applying the same correction to blue galaxies actually makes the correspondence between the two mass estimates slightly worse. The other differences in model assumptions, namely dust attenuation, treatment of stellar metallicity and the assumption of the form of star formation history used in CIGALE are harder to pick apart and contribute to the remaining biases and offset, and some component of the scatter (for a discussion of some of these points see

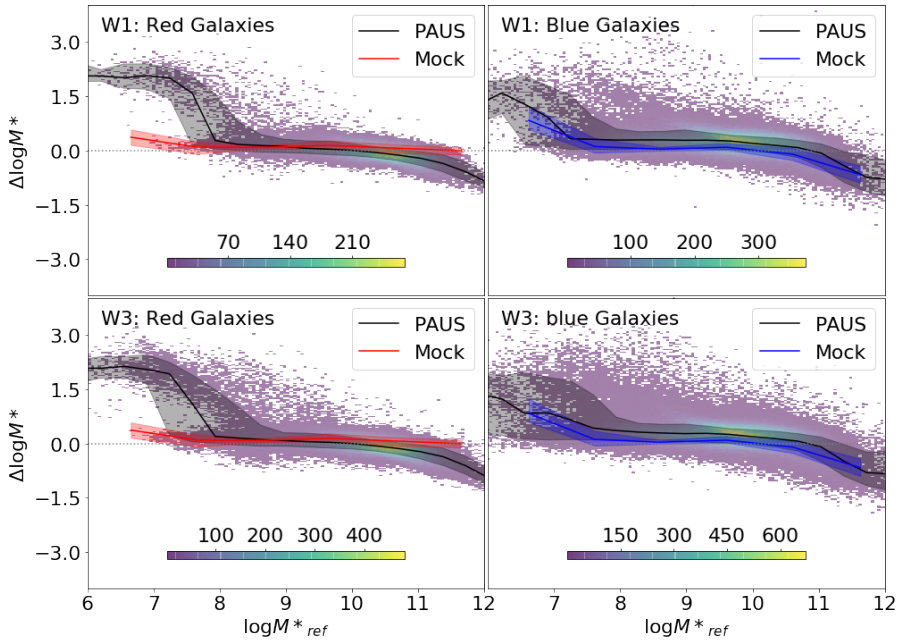


Figure 7.1: Comparison between stellar masses predicted using Kumar et al.'s symbolic regression expressions and those derived from SED fitting using the `CIGALE` code, for PAUS galaxies in the W1 (top row) and W3 (bottom row) fields. Each panel shows the difference in the log stellar mass as a function of the `CIGALE`-derived reference mass (for the PAUS data) or "true" mock stellar masses (for the `GALFORM` lightcone mock galaxies). The left and right columns correspond to red and blue galaxies, respectively. Solid lines show the running median trends for PAUS (black) and mock predictions (red or blue), with shaded regions indicating the 25th-75th percentile range. The background shows the 2D histogram density of the PAUS sample. Overall agreement is generally good, with median residuals below 0.1 dex across most of the stellar mass range

for example Mitchell et al. 2013). It is interesting to note that for low mass red galaxies, there is a step in the accuracy of the predictions from `CIGALE` with some bimodality.

7.4 Stellar Mass Function Estimation

Using the predicted stellar masses derived from Kumar et al.'s symbolic regression expressions, we estimate the galaxy stellar mass function (SMF) for both the PAUS observational sample and the perturbed `GALFORM` lightcone mock catalogue. The SMFs were calculated in multiple redshift bins between $z = 0.05$ and $z = 2.00$ using

the $1/V_{\rm max}$ estimator described in Chapter 6, this time with a magnitude limit of $i_{\rm AB} = 22.5$, to match the limit used when performing the symbolic regression.

The PAUS SMF was computed separately for the W1 and W3 fields, applying the same redshift and magnitude selection criteria as used for the mock sample. The mock SMF was derived from the perturbed lightcone catalogue with survey-like uncertainties applied, ensuring consistency with the PAUS data in both selection and observations.

Fig. 7.2 presents the stellar mass functions for the PAUS W1 and W3 fields and the lightcone mock across eight redshift bins spanning $0.05 < z < 2.00$. In each panel, black, red, and green curves correspond to W1, W3 and lightcone mock, respectively. At low and intermediate redshift ($0.2 < z < 1.3$, the agreement between the mock and PAUS measurements is generally good over the mass range $10^{9.5} < M_* h^{-1}/M_\odot < 11$, where both completeness and number statistics are sufficient. At lower stellar masses, incompleteness in the PAUS data leads to a suppression of the observed SMF, while at higher masses ($M_* > 10^{11} h^{-1} M_\odot$) the fluctuations and deviations become significant.

At higher redshifts ($z > 1.3$), the SMF from PAUS becomes increasingly incomplete at lower masses. Nevertheless, the high-mass end in PAUS and the mock remains broadly consistent.

These results demonstrate that the machine learning-based stellar masses enable a consistent recovery of the underlying galaxy stellar mass distribution, providing further validation of the symbolic regression model for population-level studies.

Moreover, this comparison between observed and mock stellar mass functions provides a valuable test of the underlying galaxy formation model implemented in `GALFORM`. Since the symbolic regression expressions are calibrated directly on the mock and applied to both datasets using consistent input features, the comparison isolates differences arising from the galaxy population itself, rather than from the stellar mass estimation procedure. In future work, this framework can be used to

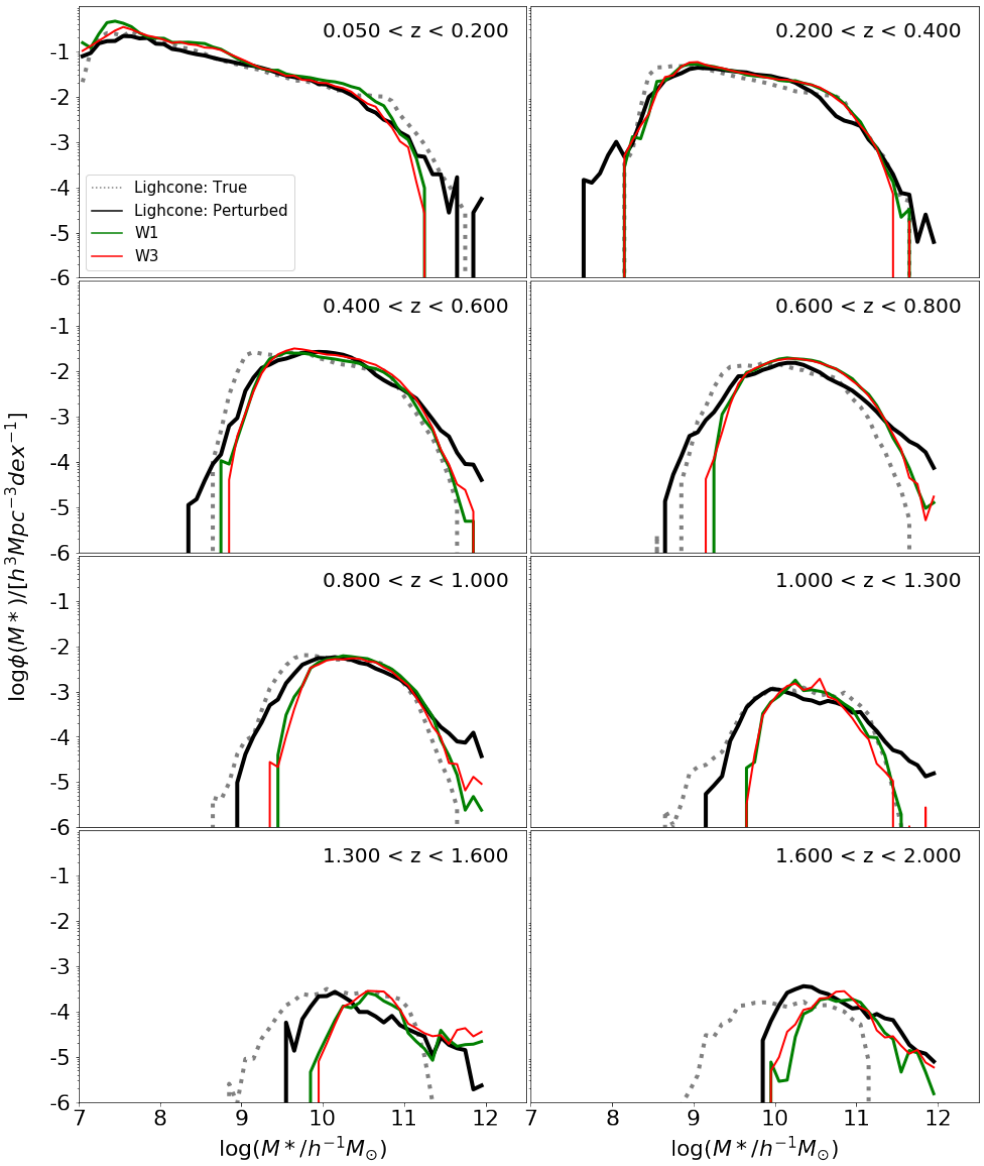


Figure 7.2: Galaxy stellar mass functions derived using Kumar et al.'s symbolic regression–predicted stellar masses for PAUS galaxies in the W1 (black) and W3 (red) fields, and for the perturbed `GALFORM` lightcone mock catalogue (green). The "true" SMF obtained from the `GALFORM` lightcone mock is also plotted as grey dotted lines for reference. Each panel corresponds to a different redshift bin from $0.05 < z < 2.00$, as labelled. The SMFs are computed using the $1/V_{\max}$ estimator with magnitude limits based on predicted rest-frame i-band magnitudes. The agreement between PAUS and the mock is generally good over the mass range where completeness is high.

evaluate the performance of semi-analytic models against observational data across redshift and environment, enabling constraints on the physical process governing stellar mass assembly.

7.5 Conclusion

In this chapter, we demonstrated how the galaxy stellar masses were predicted from broad-band photometry using symbolic regression models. The symbolic expressions were developed by Kumar et al. (in prep), who trained regression models on a perturbed version of the `GALFORM` lightcone mock catalogue to estimate stellar masses from a limited set of observable features: photometric redshift, observed u- and i-band magnitudes, and the observer-frame $(g - r)$ colour.

We validated the predicted stellar masses by comparing them against SED-based estimates from the `CIGALE` code, provided by PAUS production data. The agreement between the two methods was generally good over the stellar mass range $10^8 < M_* h^{-1} M_\odot < 10^{11.5}$, with typical residuals smaller than 0.1 dex. We also constructed stellar mass functions for both PAUS and the mock catalogue using the predicted stellar masses. The comparison shows good agreement in the intermediate mass regime ($10^{9.5} < M_* h^{-1} M_\odot < 10^{11}$) across a wide redshift range, with deviations at the low- and high-mass ends.

The results presented in this chapter demonstrate that symbolic regression provides a fast and effective alternative to traditional SED-fitting methods for estimating stellar masses. The consistent application of the same model to both mock and real data enables a direct comparison between theoretical predictions and observations, offering a tool for testing galaxy formation models in future work.

Chapter 8

Conclusions and Future Work

8.1 Thesis conclusions

The work presented in this thesis has explored the galaxy luminosity function (LF) across cosmic time using high-precision photometric redshift data from the Physics of the Accelerating Universe Survey (PAUS). Leveraging both observational data and theoretical modelling, I have measured the LF in multiple rest-frame bands out to $z \sim 2$, and tested the predictions of a semi-analytic model of galaxy formation, `GALFORM`, under realistic observational conditions.

To estimate rest-frame absolute magnitudes required for LF calculations, I trained a random forest regression model to predict the k-correction directly from observable quantities, including apparent magnitudes and photometric redshift. This non-parametric approach allowed k-correction estimates to be made without relying on galaxy templates, and was validated against true rest-frame magnitudes in the `GALFORM` mock catalogue. The same model was applied to PAUS data, ensuring consistency in the treatment of observed and simulated samples.

The PAUS dataset, with its 40 narrow-band filters and high photometric redshift accuracy, offers a new window onto galaxy populations at intermediate redshifts. In Chapter 6, I presented the main results of this thesis: a measurement of the i-band luminosity function in the W1 and W3 fields. These measurements

were performed using the $1/V_{\max}$ method in 8 different redshift bins, with a treatment of photometric errors, photometric redshift uncertainties, and completeness due to photometric redshift quality. In both fields, the luminosity function remains approximately power-law shaped at the faint end out to high redshift and shows relatively little evolution across redshift. The turnover also becomes less sharp at high redshift. There is modest brightening around the characteristic L_*, near $z \sim 0.5$, while at higher redshifts the break at the bright end becomes less distinct. This is likely due to contamination of photo-z outliers. I further examined the LF as a function of colour, revealing that blue galaxies dominate the faint end while red galaxies contribute more at bright magnitudes and low redshift.

To interpret these results, I used the `GALFORM` semi-analytic model applied to a lightcone mock catalogue matched to the PAUS survey geometry and selection. I introduced flux and redshift uncertainties to the mock, allowing a direct comparison with PAUS observation. This approach revealed that the observed turnover at the faint end is largely driven by band-shifting of the observed i-band with redshift, and that redshift outliers systematically smooth the bright-end break, particularly above $z \sim 1$. Despite these observational effects, the shape and redshift evolution of the LF predicted by `GALFORM` agrees well with PAUS data within the limits of observational uncertainty.

In Chapter 7, I extended the analysis to include estimates of stellar mass functions (SMFs) using a symbolic regression model developed by Adarsh Kumar. While I did not train the machine learning model, I prepared the perturbed mock catalogue used in training and applied the trained model to PAUS data to estimate stellar masses. I then measured the SMF from both mock and observational data. These measurements are consistent with earlier SMF estimates and provide an additional constraint on the galaxy formation model.

This thesis demonstrates that high-precision photometric redshift surveys like PAUS, combined with realistic mock catalogues, can yield accurate and meaningful estimates of the LF and SMF. The ability to resolve redshift evolution in the

LF while accounting for photometric systematic highlights the importance of this framework for next generation galaxy evolution studies. Furthermore, the successful application of random forest regression for predicting the rest-frame magnitudes and symbolic regression for estimating stellar mass illustrates the potential of machine learning methods to augment traditional approaches in observational cosmology.

8.2 Future Work

As mentioned in the previous section that the galaxy luminosity function is a powerful observable for testing and refining galaxy formation models. Several promising directions for future work arise from the results presented here.

First, the measured luminosity function can be more directly incorporated into the calibration of galaxy formation models such as `GALFORM` using statistical emulation techniques. Recent work by Madar et al. (2024) has shown that deep learning emulators can effectively mimic `GALFORM` predictions and enable efficient exploration of its parameter space. Applying similar techniques to constrain the model using the LF measured in this thesis–particularly across multiple photometric bands, for distinct galaxy populations (e.g. red and blue), and across redshift range–offers a route to systematically test the physical prescriptions implemented in `GALFORM` and potentially uncover tensions between different observables. A future study could construct a purpose-built emulator that incorporates both the i-band luminosity function evolution and its colour-separated components as calibration targets.

Second, another promising direction would be to revisit the photometric redshift measurements for galaxies at $z > 1$, where a significant fraction of outliers was observed in the sample. Improving the photo-z performance in this regime would directly increase the quality of the high-redshift LF. At $z > 1.3$, key spectral features such as the [OII] emission line and the 4000Åbreak are redshifted beyond

the upper limit of the PAUS NB filter range (8500Å), leaving only weaker features visible in the observed-frame optical. One avenue for improvement would be to incorporate longer-wavelength data–such as near-infrared (NIR) photometry from surveys like Euclid–which could help recover those features and potentially improve the photo-z measurements in this redshift range.

Third, although this thesis accounted for photometric and redshift errors in estimating the LF, the impact of these observational uncertainties on estimated SMF remains an open question. A future work could implement the same methodology used to study LF uncertainty (e.g. Monte Carlo and Jackknife methods) to assess how these uncertainties propagate into SMF estimates.

Finally, improvements to the construction of mock catalogues offer another fruitful avenue for progress. One limitation of the current mocks is that they do not include emission lines in the galaxy spectra (hence, the photometry). This is particularly relevant for narrow-band surveys like PAUS, where emission lines can have a substantial impact on the observational fluxes and, therefore, on photometric redshift estimation. Incorporating physically motivated emission line models into the mock photometry–e.g. by convolving the SFH and gas-phase metallicities predicted by `GALFORM` with an HII region model–would make the simulated catalogue more realistic and allow for more accurate comparison with observed photo-z distributions. This, in turn, could help improve the fidelity of k-correction predictions and redshift estimation pipelines.

Together, these future direction will help deepen our understanding of the connection between observable galaxy populations and the physical processes that shape them.

Bibliography

T. M. C. Abbott et al. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *Phys. Rev. D*, 98(4):043526, Aug. 2018. doi: 10.1103/PhysRevD.98.043526.

V. Acquaviva. How to measure metallicity from five-band photometry with supervised machine learning algorithms. *Monthly Notices of the Royal Astronomical Society*, 456(2):1618–1626, Feb. 2016. doi: 10.1093/mnras/stv2703.

R. Ahumada et al. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *The Astrophysical Journal, Supplement*, 249(1):3, July 2020. doi: 10.3847/1538-4365/ab929e.

A. Alarcon et al. The PAU Survey: an improved photo-z sample in the COSMOS field. *Monthly Notices of the Royal Astronomical Society*, 501(4):6103–6122, Mar. 2021. doi: 10.1093/mnras/staa3659.

M. Ayromlou et al. Comparing galaxy formation in the L-GALAXIES semi-analytical model and the IllustrisTNG simulations. *Monthly Notices of the Royal Astronomical Society*, 502(1):1051–1069, Mar. 2021. doi: 10.1093/mnras/staa4011.

I. K. Baldry et al. Galaxy and mass assembly: the g02 field, herschel–atlas target selection and data release 3. *Monthly Notices of the Royal Astronomical Society*, 474(3):3875–3888, 11 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx3042. URL `https://doi.org/10.1093/mnras/stx3042`.

D. Baron. Machine Learning in Astronomy: a practical overview. *arXiv e-prints*, art. arXiv:1904.07248, Apr. 2019. doi: 10.48550/arXiv.1904.07248.

C. M. Baugh. A primer on hierarchical galaxy formation: the semi-analytical approach. *Reports on Progress in Physics*, 69(12):3101–3156, Dec. 2006. doi: 10.1088/0034-4885/69/12/R02.

C. M. Baugh et al. Can the faint submillimetre galaxies be explained in the Λ cold dark matter model? *Monthly Notices of the Royal Astronomical Society*, 356(3): 1191–1200, Jan. 2005. doi: 10.1111/j.1365-2966.2004.08553.x.

C. M. Baugh et al. Galaxy formation in the Planck Millennium: the atomic hydrogen content of dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 483(4):4922–4937, Mar. 2019. doi: 10.1093/mnras/sty3427.

C. M. Baugh, C. G. Lacey, V. Gonzalez-Perez, and G. Manzoni. Modelling emission lines in star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 510(2):1880–1893, Feb. 2022. doi: 10.1093/mnras/stab3506.

P. S. Behroozi, R. H. Wechsler, and H.-Y. Wu. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. *Astrophysical Journal*, 762(2):109, Jan. 2013a. doi: 10.1088/0004-637X/762/2/109.

P. S. Behroozi et al. Gravitationally Consistent Halo Catalogs and Merger Trees for Precision Cosmology. *Astrophysical Journal*, 763(1):18, Jan. 2013b. doi: 10.1088/0004-637X/763/1/18.

A. J. Benson et al. What Shapes the Luminosity Function of Galaxies? *Astrophysical Journal*, 599:38–49, Dec. 2003. doi: 10.1086/379160.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(null):281–305, Feb. 2012. ISSN 1532-4435.

F. Bigiel et al. A Constant Molecular Gas Depletion Time in Nearby Disk Galaxies. *Astrophysical Journal*, 730(2):L13, Apr. 2011. doi: 10.1088/2041-8205/730/2/L13.

J. Blaizot et al. MoMaF: the Mock Map Facility. *Monthly Notices of the Royal Astronomical Society*, 360(1):159–175, June 2005. doi: 10.1111/j.1365-2966.2005.09019.x.

M. R. Blanton et al. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. *Astrophysical Journal*, 592(2):819–838, Aug. 2003. doi: 10.1086/375776.

L. Blitz and E. Rosolowsky. The Role of Pressure in GMC Formation II: The H_2-Pressure Relation. *Astrophysical Journal*, 650(2):933–944, Oct. 2006. doi: 10.1086/505417.

D. G. Bonfield et al. Photometric redshift estimation using Gaussian processes. *Monthly Notices of the Royal Astronomical Society*, 405(2):987–994, June 2010. doi: 10.1111/j.1365-2966.2010.16544.x.

V. Bonjean et al. Star formation rates and stellar masses from machine learning. *Astronomy & Astrophysics*, 622:A137, Feb. 2019. doi: 10.1051/0004-6361/201833972.

R. G. Bower et al. Breaking the hierarchy of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 370(2):645–655, Aug. 2006. doi: 10.1111/j.1365-2966.2006.10519.x.

G. B. Brammer et al. 3D-HST: A Wide-field Grism Spectroscopic Survey with the Hubble Space Telescope. *The Astrophysical Journal, Supplement*, 200(2):13, June 2012. doi: 10.1088/0067-0049/200/2/13.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, Jan. 2001. doi: 10.1023/A:1010933404324.

S. Carliles et al. Random forests for photometric redshifts. *The Astrophysical Journal*, 712(1):511, mar 2010. doi: 10.1088/0004-637X/712/1/511. URL `https://dx.doi.org/10.1088/0004-637X/712/1/511`.

F. J. Castander et al. The PAU camera and the PAU survey at the William Herschel Telescope. *Ground-based and Airborne Instrumentation for Astronomy IV*, 8446: 84466D, Sept. 2012. doi: 10.1117/12.926234.

S. Chandrasekhar. Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction. *Astrophysical Journal*, 97:255, Mar. 1943. doi: 10.1086/144517.

J. Chu et al. Galaxy stellar and total mass estimation using machine learning. *Monthly Notices of the Royal Astronomical Society*, 528(4):6354–6369, Mar. 2024. doi: 10.1093/mnras/stae406.

S. Cole et al. A recipe for galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 271:781–806, Dec. 1994. doi: 10.1093/mnras/271.4.781.

S. Cole, C. G. Lacey, C. M. Baugh, and C. S. Frenk. Hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 319(1):168–204, Nov. 2000. doi: 10.1046/j.1365-8711.2000.03879.x.

M. Colless et al. The 2dF Galaxy Redshift Survey: spectra and redshifts. *Monthly Notices of the Royal Astronomical Society*, 328(4):1039–1063, Dec. 2001. doi: 10.1046/j.1365-8711.2001.04902.x.

C. Conroy. Modeling the Panchromatic Spectral Energy Distributions of Galaxies. *Annual Review of Astronomy & Astrophysics*, 51(1):393–455, Aug. 2013. doi: 10.1146/annurev-astro-082812-141017.

S. Contreras, C. M. Baugh, P. Norberg, and N. Padilla. How robust are predictions of galaxy clustering? *Monthly Notices of the Royal Astronomical Society*, 432 (4):2717–2730, July 2013. doi: 10.1093/mnras/stt629.

S. Contreras, C. M. Baugh, P. Norberg, and N. Padilla. The galaxy-dark matter halo connection: which galaxy properties are correlated with the host halo mass? *Monthly Notices of the Royal Astronomical Society*, 452(2):1861–1876, Sept. 2015. doi: 10.1093/mnras/stv1438.

B. Csizi et al. The PAU Survey: Galaxy stellar population properties estimates with narrowband data. *Astronomy & Astrophysics*, 689:A37, Sept. 2024. doi: 10.1051/0004-6361/202449838.

I. Davidzon et al. The VIMOS Public Extragalactic Redshift Survey (VIPERS). A precise measurement of the galaxy stellar mass function and the abundance of massive galaxies at redshifts $0.5 < z < 1.3$. *Astronomy & Astrophysics*, 558: A23, Oct. 2013. doi: 10.1051/0004-6361/201321511.

L. J. M. Davies et al. Galaxy and mass assembly (gama): curation and reanalysis of 16.6k redshifts in the g10/cosmos region. *Monthly Notices of the Royal Astronomical Society*, 447(1):1014–1027, 12 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2515. URL `https://doi.org/10.1093/mnras/stu2515`.

M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophysical Journal*, 292:371–394, May 1985. doi: 10.1086/163168.

M. Davis et al. Science Objectives and Early Results of the DEEP2 Redshift Survey. In P. Guhathakurta, editor, *Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II*, volume 4834 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 161–172, Feb. 2003. doi: 10.1117/12.457897.

I. V. Daza-Perilla et al. The PAU survey: Enhancing photometric redshift estimation using DEEPz. *Astronomy & Astrophysics*, 693:A102, Jan. 2025. doi: 10.1051/0004-6361/202452053.

J. T. A. de Jong, G. A. Verdoes Kleijn, K. H. Kuijken, and E. A. Valentijn. The Kilo-Degree Survey. *Experimental Astronomy*, 35(1-2):25–44, Jan. 2013. doi: 10.1007/s10686-012-9306-1.

DESI Collaboration et al. Validation of the Scientific Program for the Dark Energy Spectroscopic Instrument. *The Astronomical Journal*, 167(2):62, Feb. 2024. doi: 10.3847/1538-3881/ad0b08.

H. Domínguez Sánchez et al. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, Feb. 2018. doi: 10.1093/mnras/sty338.

S. P. Driver and S. Phillipps. Is the Luminosity Distribution of Field Galaxies Really Flat? *Astrophysical Journal*, 469:529, Oct. 1996. doi: 10.1086/177801.

S. P. Driver et al. Galaxy And Mass Assembly (GAMA): the $0.013 < z < 0.1$ cosmic spectral energy distribution from 0.1 μm to 1 mm. *Monthly Notices of the Royal Astronomical Society*, 427(4):3244–3264, Dec. 2012. doi: 10.1111/j.1365-2966.2012.22036.x.

G. Efstathiou. A model of supernova feedback in galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 317(3):697–719, Sept. 2000. doi: 10.1046/j.1365-8711.2000.03665.x.

G. Efstathiou, R. S. Ellis, and B. A. Peterson. Analysis of a complete galaxy redshift survey. II. The field-galaxy luminosity function. *Monthly Notices of the Royal Astronomical Society*, 232:431–461, May 1988. doi: 10.1093/mnras/232.2.431.

P. J. Elahi, R. J. Thacker, and L. M. Widrow. Peaks above the Maxwellian Sea: a new approach to finding substructures in N-body haloes. *Monthly Notices*

of the Royal Astronomical Society, 418(1):320–335, Nov. 2011. doi: 10.1111/j.1365-2966.2011.19485.x.

T. Erben et al. CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey - imaging data and catalogue products. *Monthly Notices of the Royal Astronomical Society*, 433(3):2545–2563, Aug. 2013. doi: 10.1093/mnras/stt928.

M. Eriksen and E. Gaztañaga. Combining spectroscopic and photometric surveys: Same or different sky? *Monthly Notices of the Royal Astronomical Society*, 451 (2):1553–1560, 06 2015a. ISSN 0035-8711. doi: 10.1093/mnras/stv1093. URL `https://doi.org/10.1093/mnras/stv1093`.

M. Eriksen and E. Gaztañaga. Combining spectroscopic and photometric surveys using angular cross-correlations – i. algorithm and modelling. *Monthly Notices of the Royal Astronomical Society*, 452(2):2149–2167, 07 2015b. ISSN 0035-8711. doi: 10.1093/mnras/stv1288. URL `https://doi.org/10.1093/mnras/stv1288`.

M. Eriksen and E. Gaztañaga. Combining spectroscopic and photometric surveys using angular cross-correlations - II. Parameter constraints from different physical effects. *Monthly Notices of the Royal Astronomical Society*, 452(2):2168–2184, Sept. 2015c. doi: 10.1093/mnras/stv1075.

M. Eriksen et al. The PAU Survey: early demonstration of photometric redshift performance in the COSMOS field. *Monthly Notices of the Royal Astronomical Society*, 484(3):4200–4215, Apr. 2019. doi: 10.1093/mnras/stz204.

N. Fanidakis et al. Grand unification of AGN activity in the ΛCDM cosmology. *Monthly Notices of the Royal Astronomical Society*, 410(1):53–74, Jan. 2011. doi: 10.1111/j.1365-2966.2010.17427.x.

O. L. Fèvre et al. The Galaxy Merger Rate History (GMRH) since z = 3. In W. H. Sun, C. K. Xu, N. Z. Scoville, and D. B. Sanders, editors, *Galaxy Mergers in an Evolving Universe*, volume 477 of *Astronomical Society of the Pacific Conference Series*, page 133, Oct. 2013.

S. Folkes et al. The 2dF Galaxy Redshift Survey: spectral types and luminosity functions. *Monthly Notices of the Royal Astronomical Society*, 308(2):459–472, Sept. 1999. doi: 10.1046/j.1365-8711.1999.02721.x.

V. Fontirroig et al. A Statistical Study of Lopsided Galaxies using Random Forest. *arXiv e-prints*, art. arXiv:2411.19723, Nov. 2024. doi: 10.48550/arXiv.2411.19723.

J. S. Gómez et al. Halo merger tree comparison: impact on galaxy formation models. *Monthly Notices of the Royal Astronomical Society*, 510(4):5500–5519, Mar. 2022. doi: 10.1093/mnras/stab3661.

V. Gonzalez-Perez et al. How sensitive are predicted galaxy luminosities to the choice of stellar population synthesis model? *Monthly Notices of the Royal Astronomical Society*, 439(1):264–283, Mar. 2014. doi: 10.1093/mnras/stt2410.

A. J. Griffin et al. The evolution of SMBH spin and AGN luminosities for $z < 6$ within a semi-analytic model of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 487(1):198–227, July 2019. doi: 10.1093/mnras/stz1216.

Q. Guo et al. Galaxies in the EAGLE hydrodynamical simulation and in the Durham and Munich semi-analytical models. *Monthly Notices of the Royal Astronomical Society*, 461(4):3457–3482, Oct. 2016. doi: 10.1093/mnras/stw1525.

A. H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D*, 23(2):347–356, Jan. 1981. doi: 10.1103/PhysRevD.23.347.

J. Han, Y. P. Jing, H. Wang, and W. Wang. Resolving subhaloes' lives with the Hierarchical Bound-Tracing algorithm. *Monthly Notices of the Royal Astronomical Society*, 427(3):2437–2449, Dec. 2012. doi: 10.1111/j.1365-2966.2012.22111.x.

F. Hernandez Vivanco, R. Smith, E. Thrane, and P. D. Lasky. A scalable random forest regressor for combining neutron-star equation of state measurements: a

case study with GW170817 and GW190425. *Monthly Notices of the Royal Astronomical Society*, 499(4):5972–5977, Dec. 2020. doi: 10.1093/mnras/staa3243.

C. Heymans et al. CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey. *Monthly Notices of the Royal Astronomical Society*, 427(1):146–166, Nov. 2012. doi: 10.1111/j.1365-2966.2012.21952.x.

H. Hildebrandt et al. KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 465(2):1454–1498, Feb. 2017. doi: 10.1093/mnras/stw2805.

H. Hoekstra, M. Viola, and R. Herbonnet. A study of the sensitivity of shape measurements to the input parameters of weak-lensing image simulations. *Monthly Notices of the Royal Astronomical Society*, 468(3):3295–3311, July 2017. doi: 10.1093/mnras/stx724.

K. Hoffman, J. Y. Sung, and A. Zazzera. Multi-output random forest regression to emulate the earliest stages of planet formation. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6, 2021. doi: 10.1109/SIEDS52267.2021.9483749.

D. W. Hogg, I. K. Baldry, M. R. Blanton, and D. J. Eisenstein. The K correction. *arXiv e-prints*, art. astro-ph/0210394, Oct. 2002. doi: 10.48550/arXiv.astro-ph/0210394.

O. Ilbert et al. The VIMOS-VLT deep survey. Evolution of the galaxy luminosity function up to z = 2 in first epoch data. *Astronomy & Astrophysics*, 439(3):863–876, Sept. 2005. doi: 10.1051/0004-6361:20041961.

O. Ilbert et al. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *Astronomy & Astrophysics*, 457(3):841–856, Oct. 2006. doi: 10.1051/0004-6361:20065138.

C. Y. Jiang et al. A fitting formula for the merger timescale of galaxies in hierarchical clustering. *The Astrophysical Journal*, 675(2):1095, mar 2008. doi: 10.1086/526412. URL `https://dx.doi.org/10.1086/526412`.

L. Jiang, J. C. Helly, S. Cole, and C. S. Frenk. N-body dark matter haloes with simple hierarchical histories. *Monthly Notices of the Royal Astronomical Society*, 440(3):2115–2135, May 2014. doi: 10.1093/mnras/stu390.

R. C. Kennicutt, Jr. The rate of star formation in normal disk galaxies. *Astrophysical Journal*, 272:54–67, Sept. 1983. doi: 10.1086/161261.

R. C. Kennicutt, Jr. The Global Schmidt Law in Star-forming Galaxies. *Astrophysical Journal*, 498(2):541–552, May 1998. doi: 10.1086/305588.

D. Kereš, N. Katz, D. H. Weinberg, and R. Davé. How do galaxies get their gas? *Monthly Notices of the Royal Astronomical Society*, 363(1):2–28, Oct. 2005. doi: 10.1111/j.1365-2966.2005.09451.x.

R. Kugel et al. FLAMINGO: calibrating large cosmological hydrodynamical simulations with machine learning. *Monthly Notices of the Royal Astronomical Society*, 526(4):6103–6127, Dec. 2023. doi: 10.1093/mnras/stad2540.

N. H. Kuiper. Tests concerning random points on a circle. In *Nederl. Akad. Wetensch. Proc. Ser. A*, volume 63, pages 38–47, 1960.

C. G. Lacey et al. A unified multiwavelength model of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 462(4):3854–3911, Nov. 2016. doi: 10.1093/mnras/stw1888.

C. d. P. Lagos et al. The origin of the atomic and molecular gas contents of early-type galaxies - II. Misaligned gas accretion. *Monthly Notices of the Royal Astronomical Society*, 448(2):1271–1287, Apr. 2015. doi: 10.1093/mnras/stu2763.

A. K. Leroy et al. The Star Formation Efficiency in Nearby Galaxies: Measuring Where Gas Forms Stars Effectively. *The Astronomical Journal*, 136(6):2782–2845, Dec. 2008. doi: 10.1088/0004-6256/136/6/2782.

S. J. Lilly et al. The Canada-France Redshift Survey. VI. Evolution of the Galaxy Luminosity Function to Z approximately 1. *Astrophysical Journal*, 455:108, Dec. 1995. doi: 10.1086/176560.

J. Loveday, B. A. Peterson, G. Efstathiou, and S. J. Maddox. The Stromlo–APM Redshift Survey. I. The Luminosity Function and Space Density of Galaxies. *Astrophysical Journal*, 390:338, May 1992. doi: 10.1086/171284.

J. Loveday et al. Galaxy and Mass Assembly (GAMA): ugriz galaxy luminosity functions. *Monthly Notices of the Royal Astronomical Society*, 420(2):1239–1262, Feb. 2012. doi: 10.1111/j.1365-2966.2011.20111.x.

M. S. Madar, C. M. Baugh, and D. Shi. Predictions for the abundance and clustering of H α emitting galaxies. *Monthly Notices of the Royal Astronomical Society*, 535(4):3324–3341, Dec. 2024. doi: 10.1093/mnras/stae2560.

R. K. Malbon, C. M. Baugh, C. S. Frenk, and C. G. Lacey. Black hole growth in hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 382(4):1394–1414, Dec. 2007. doi: 10.1111/j.1365-2966.2007.12317.x.

G. Manzoni et al. The PAU Survey: a new constraint on galaxy formation models using the observed colour redshift relation. *Monthly Notices of the Royal Astronomical Society*, 530(2):1394–1413, May 2024. doi: 10.1093/mnras/stae659.

R. O. Marzke, M. J. Geller, J. P. Huchra, and H. G. Corwin, Jr. The Luminosity Function for Different Morphological Types in the CFA Redshift Survey. *The Astronomical Journal*, 108:437, Aug. 1994. doi: 10.1086/117081.

F. J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2280095`.

T. McNaught-Roberts et al. Galaxy And Mass Assembly (GAMA): the dependence of the galaxy luminosity function on environment, redshift and colour. *Monthly Notices of the Royal Astronomical Society*, 445(2):2125–2145, Dec. 2014. doi: 10.1093/mnras/stu1886.

A. I. Merson et al. Lightcone mock catalogues from semi-analytic models of galaxy formation - I. Construction and application to the BzK colour selection. *Monthly Notices of the Royal Astronomical Society*, 429(1):556–578, Feb. 2013. doi: 10.1093/mnras/sts355.

P. D. Mitchell, C. G. Lacey, C. M. Baugh, and S. Cole. How well can we really estimate the stellar masses of galaxies from broad-band photometry? *Monthly Notices of the Royal Astronomical Society*, 435(1):87–114, Oct. 2013. doi: 10.1093/mnras/stt1280.

M. Moles et al. The Alhambra Survey: a Large Area Multimedium-Band Optical and Near-Infrared Photometric Survey. *The Astronomical Journal*, 136(3):1325–1339, Sept. 2008. doi: 10.1088/0004-6256/136/3/1325.

S. Mucesh et al. A machine learning approach to galaxy properties: joint redshift-stellar mass probability distributions with Random Forest. *Monthly Notices of the Royal Astronomical Society*, 502(2):2770–2786, Apr. 2021. doi: 10.1093/mnras/stab164.

V. F. Mukhanov and G. V. Chibisov. Quantum fluctuations and a nonsingular universe. *Soviet Journal of Experimental and Theoretical Physics Letters*, 33:532, May 1981.

J. F. Navarro and S. D. M. White. Simulations of Dissipative Galaxy Formation in Hierarchically Clustering Universes - Part One - Tests of the Code. *Monthly Notices of the Royal Astronomical Society*, 265:271, Nov. 1993. doi: 10.1093/mnras/265.2.271.

J. F. Navarro, C. S. Frenk, and S. D. M. White. A Universal Density Profile from Hierarchical Clustering. *Astrophysical Journal*, 490(2):493–508, Dec. 1997. doi: 10.1086/304888.

D. Navarro-Gironés et al. The PAU survey: photometric redshift estimation in deep wide fields. *Monthly Notices of the Royal Astronomical Society*, 534(2): 1504–1527, Oct. 2024. doi: 10.1093/mnras/stae1686.

G. L. Nedjati-Gilani et al. Machine learning based compartment models with permeability for white matter microstructure imaging. *NeuroImage*, 150: 119–135, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.02.013. URL `https://www.sciencedirect.com/science/article/pii/S1053811917301179`.

E. Neistein, S. Khochfar, C. Dalla Vecchia, and J. Schaye. Hydrodynamical simulations and semi-analytic models of galaxy formation: two sides of the same coin. *Monthly Notices of the Royal Astronomical Society*, 421(4):3579–3593, Apr. 2012. doi: 10.1111/j.1365-2966.2012.20584.x.

J. A. Newman et al. The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts. *The Astrophysical Journal, Supplement*, 208(1): 5, Sept. 2013. doi: 10.1088/0067-0049/208/1/5.

P. Norberg et al. The 2dF Galaxy Redshift Survey: the b_J-band galaxy luminosity function and survey selection function. *Monthly Notices of the Royal Astronomical Society*, 336(3):907–931, Nov. 2002. doi: 10.1046/j.1365-8711.2002.05831.x.

P. Norberg, C. M. Baugh, E. Gaztañaga, and D. J. Croton. Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics. *Monthly Notices of the Royal Astronomical Society*, 396(1):19–38, June 2009. doi: 10.1111/j.1365-2966.2009.14389.x.

J. B. Oke and J. E. Gunn. Secondary standard stars for absolute spectrophotometry. *Astrophysical Journal*, 266:713–717, Mar. 1983. doi: 10.1086/160817.

C. Padilla et al. The Physics of the Accelerating Universe Camera. *The Astronomical Journal*, 157(6):246, June 2019. doi: 10.3847/1538-3881/ab0412.

J. Pasquet et al. Photometric redshifts from SDSS images using a convolutional neural network. *Astronomy & Astrophysics*, 621:A26, Jan. 2019. doi: 10.1051/0004-6361/201833617.

J. A. Peacock. *Cosmological Physics*. Cambridge, 1999.

F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Oct. 2011. doi: 10.48550/arXiv.1201.0490.

Planck Collaboration et al. Planck 2018 results. X. Constraints on inflation. *Astronomy & Astrophysics*, 641:A10, Sept. 2020. doi: 10.1051/0004-6361/201833887.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

B. H. F. Ramos et al. Evolution of Galaxy Luminosity Function Using Photometric Redshifts. *The Astronomical Journal*, 142(2):41, Aug. 2011. doi: 10.1088/0004-6256/142/2/41.

M. Schaller et al. Baryon effects on the internal structure of ΛCDM haloes in the EAGLE simulations. *Monthly Notices of the Royal Astronomical Society*, 451(2):1247–1267, Aug. 2015. doi: 10.1093/mnras/stv1067.

P. Schechter. An analytic expression for the luminosity function for galaxies. *Astrophysical Journal*, 203:297–306, Jan. 1976. doi: 10.1086/154079.

M. Schmidt. Space Distribution and Luminosity Functions of Quasi-Stellar Radio Sources. *Astrophysical Journal*, 151:393, Feb. 1968. doi: 10.1086/149446.

M. Scodeggio et al. The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2). *Astronomy & Astrophysics*, 609:A84, Jan. 2018. doi: 10.1051/0004-6361/201630114.

N. Scoville et al. COSMOS: Hubble Space Telescope Observations. *The Astrophysical Journal, Supplement*, 172(1):38–45, Sept. 2007. doi: 10.1086/516580.

A. Smith et al. A light-cone catalogue from the Millennium-XXL simulation: improved spatial interpolation and colour distributions for the DESI BGS. *Monthly Notices of the Royal Astronomical Society*, 516(3):4529–4542, Nov. 2022. doi: 10.1093/mnras/stac2519.

V. Springel, S. D. M. White, G. Tormen, and G. Kauffmann. Populating a cluster of galaxies - I. Results at z=0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750, Dec. 2001. doi: 10.1046/j.1365-8711.2001.04912.x.

V. Springel, C. S. Frenk, and S. D. M. White. The large-scale structure of the Universe. *Nature*, 440(7088):1137–1144, Apr. 2006. doi: 10.1038/nature04805.

L. Stothert et al. The PAU Survey: spectral features and galaxy clustering using simulated narrow-band photometry. *Monthly Notices of the Royal Astronomical Society*, 481(3):4221–4235, Dec. 2018. doi: 10.1093/mnras/sty2491.

L. Wenzl et al. Cosmology with the Roman Space Telescope - Synergies with CMB lensing. *Monthly Notices of the Royal Astronomical Society*, 512(4):5311–5328, June 2022. doi: 10.1093/mnras/stac790.

S. D. M. White and C. S. Frenk. Galaxy Formation through Hierarchical Clustering. *Astrophysical Journal*, 379:52, Sept. 1991. doi: 10.1086/170483.

S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183:341–358, May 1978. doi: 10.1093/mnras/183.3.341.

C. Wolf et al. The evolution of faint AGN between z $=_\sim$1 and z $=_\sim$5 from the COMBO-17 survey. *Astronomy & Astrophysics*, 408:499–514, Sept. 2003. doi: 10.1051/0004-6361:20030990.

E. Zucca et al. The ESO Slice Project (ESP) galaxy redshift survey. II. The luminosity function and mean galaxy density. *Astronomy & Astrophysics*, 326: 477–488, Oct. 1997. doi: 10.48550/arXiv.astro-ph/9705096.

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2_ε. It was created using the *memoir* package, maintained by Lars Madsen, with the *madsen* chapter style. The font used is Latin Modern, derived from fonts designed by Donald E. Knuth.