



Observatório  
Nacional

DOCTOR THESIS

DATA-DRIVEN COSMOLOGICAL RECONSTRUCTION USING GAUSSIAN  
PROCESSES AND MACHINE LEARNING

AGRIPINO SEGUNDO DE SOUSA NETO

RIO DE JANEIRO  
2026

Ministério da Ciência, Tecnologia, Inovações e Comunicações  
Observatório Nacional  
Programa de Pós-Graduação

Doctor Thesis

DATA-DRIVEN COSMOLOGICAL RECONSTRUCTION USING GAUSSIAN  
PROCESSES AND MACHINE LEARNING

by

Agripino Segundo de Sousa Neto

Thesis submitted to the Faculty of the  
Graduate Program in Astronomy at the  
Observatório Nacional, as part of the necessary  
requirements for the Degree of Doctor in  
Astronomy.

Advisor: Dr. Jailson Souza de Alcaniz

Rio de Janeiro, RJ – Brazil  
April 2026



S725

Segundo de Sousa Neto, Agripino

Data-Driven Cosmological Reconstruction Using Gaussian Processes and Machine Learning [Rio de Janeiro] 2026.

xx, 102 p. 29,7 cm: graf. il. tab.

Tese (doutorado) - Observatório Nacional - Rio de Janeiro, 2026.

1. Observational cosmology. 2. Dark energy. 3. Galaxy formation. 4. Gaussian processes. 5. Symbolic regression. I. Observatório Nacional. II. Título.

CDU 000.000.000

“DATA-DRIVEN COSMOLOGICAL RECONSTRUCTION USING GAUSSIAN  
PROCESSES AND MACHINE LEARNING”

AGRIPINO SEGUNDO DE SOUSA NETO

THESIS SUBMITTED TO THE FACULTY OF THE GRADUATE PROGRAM  
IN ASTRONOMY AT THE OBSERVATÓRIO NACIONAL AS PART OF THE  
NECESSARY REQUIREMENTS FOR A DOCTOR DEGREE IN ASTRONOMY.

Approved by:

---

Dr. Jailson Souza de Alcaniz – ON  
(Advisor)

---

Dr. Cássio Bruno Magalhães Pigozzo – UFBA

---

Dr. Rodrigo de Sousa Gonçalves – UFRRJ

---

Dra. Lilianne Mariko Izuti Nakazono – ON

---

Dr. Jose Carlos Neves de Araujo – INPE

RIO DE JANEIRO, RJ – BRAZIL  
APRIL 29, 2026



---

## Acknowledgments

---

Agradeço ao meu pai, Aldeirton, por sempre priorizar minha educação e nunca deixar faltar nada, colocando muitas vezes meus sonhos à frente dos seus próprios. À minha mãe, Ana, que, mesmo à distância, enviava suas receitas para que eu pudesse reproduzi-las aqui e sentir um pouco do gostinho de casa. Aos meus irmãos, Gabriel e Aldelany, e aos demais familiares, por estarem sempre dispostos a ajudar, mesmo de longe.

Aos meus amigos de infância, por todo o apoio nos momentos em que precisei e pelos inúmeros momentos de descontração. Aos amigos da universidade, com quem, mesmo após a formatura, continuo compartilhando experiências das mais diversas, mostrando que a vida na pós-graduação é árdua, independentemente do lugar. Aos novos amigos do Observatório Nacional, com quem compartilhei não apenas a ciência, mas também os desafios de viver no Rio de Janeiro estando longe de casa. Em especial, aos colegas da Casa Branca, com quem convivi por quatro anos e dividi não apenas o espaço, mas também muitas histórias e aprendizados. À minha “sous-chef” Camila, que esteve ao meu lado muito além dos corredores do Observatório, compartilhando a rotina, os desafios e os momentos mais leves dessa caminhada.

À Ayslana, com quem divido a sala e tantas conversas ao longo dos dias; ao Bruno, meu amigo desde a UERN, com quem morei por dois anos e compartilhei uma fase importante dessa jornada; ao Rayff, que aceitou dividir apartamento com um completo desconhecido; e à Shao, minha amiga chinesa, com quem sempre aprendo sobre a cultura chinesa.

Ao meu orientador, Jailson Alcaniz, por aceitar o desafio de trabalhar comigo, pela confiança depositada e pela orientação ao longo deste percurso.

Aos meus colaboradores, em especial Javier, o “Jedi” do Gaussian Processes, e Carlos Bengaly, sempre disposto a esclarecer minhas dúvidas nos mais variados temas. À Aldinez, minha ex-orientadora e amiga, por ter sido fundamental em minha formação acadêmica, pelos ensinamentos, pela confiança e pelo apoio ao longo da minha trajetória.

Aos pesquisadores e técnicos do Observatório Nacional, que contribuíram para minha formação acadêmica, e aos funcionários que fazem com que o ON funcione diariamente, em especial à Silvinha e ao Chris, que me acolheram quando cheguei ao Rio de Janeiro.

Ao Observatório Nacional, pela oportunidade de cursar uma pós-graduação de excelência, oferecendo sempre a estrutura necessária e mantendo suas portas abertas quando precisei de um espaço para estudo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro fundamental para a realização deste trabalho.

*Há tempo, muito tempo  
Que eu estou longe de casa  
E nessas ilhas cheias de distância  
O meu blusão de couro se estragou  
  
Ouvi dizer num papo da rapaziada  
Que aquele amigo que embarcou comigo  
Cheio de esperança e fé, já se mandou  
  
– Belchior, “Tudo Outra Vez”.*

DATA-DRIVEN COSMOLOGICAL RECONSTRUCTION USING GAUSSIAN  
PROCESSES AND MACHINE LEARNING

**RESUMO**

A cosmologia moderna fornece uma descrição consistente do Universo em larga escala, embora questões fundamentais, como a natureza da energia escura e a consistência entre diferentes estimativas cosmológicas, permaneçam em aberto. Levantamentos de nova geração, como o DESI, impõem novos desafios ao modelo  $\Lambda$ CDM, enquanto observações de idades de objetos cósmicos fornecem uma via complementar para investigar a história de expansão.

Nesse contexto, técnicas de aprendizado de máquina têm ganhado destaque na inferência de observáveis cosmológicos. Assim, esta tese investiga sua aplicação na reconstrução dessas quantidades, com ênfase na Regressão Simbólica (RS). A análise baseia-se na aplicação conjunta dessa abordagem e de Processos Gaussianos (PG), estes últimos empregados como um método estatístico não paramétrico, permitindo avaliar a consistência e a robustez das reconstruções. Utilizamos dados recentes de oscilações acústicas de bárions do DESI, combinados com supernovas do tipo Ia, bem como dados de idades de galáxias passivas.

No contexto da energia escura, realizamos a reconstrução da equação de estado  $w(z)$  diretamente a partir de observáveis cosmológicos, com o objetivo de investigar uma possível evidência de energia escura dinâmica, constituindo uma das primeiras aplicações de regressão simbólica a dados do DESI para esse fim. A reconstrução foi inicialmente realizada com medidas de  $D_H/r_d$ , onde  $D_H$  é a distância de Hubble e  $r_d$  o horizonte acústico na época de dragagem. Em seguida, a análise foi refinada utilizando  $D_M/r_d$ , onde  $D_M$  é a distância comóvel transversal, combinada com dados de supernovas.

Os resultados mostram que, dadas as incertezas atuais, não há evidências conclusivas de desvios em relação ao modelo  $\Lambda$ CDM. A função  $w(z)$  apresenta sensibilidade significativa a parâmetros cosmológicos, particularmente  $\Omega_m$  (densidade de matéria) e  $r_d$ , cujas variações impactam o comportamento inferido. Ainda assim, as reconstruções permanecem compatíveis com o  $\Lambda$ CDM dentro de  $3\sigma$ , com apenas desvios marginais em regiões específicas.

No estudo do tempo de formação de galáxias, reconstruímos  $t_f(z)$ , definido como o tempo cósmico no qual uma determinada fração da massa estelar final de uma galáxia já foi formada, utilizando duas abordagens independentes baseadas em diferentes conjuntos de dados. Ambas produzem resultados consistentes e indicam uma evolução de  $t_f(z)$  com o redshift, especialmente em baixos redshifts. Verificamos ainda que essa estimativa é

pouco sensível ao modelo cosmológico, sendo dominada principalmente pelas incertezas nos parâmetros cosmológicos, em particular  $H_0$ .



# DATA-DRIVEN COSMOLOGICAL RECONSTRUCTION USING GAUSSIAN PROCESSES AND MACHINE LEARNING

## ABSTRACT

Modern cosmology provides a consistent description of the Universe on large scales, although fundamental questions, such as the nature of dark energy and the consistency among different cosmological estimates, remain open. Current surveys, such as DESI, pose new challenges to the  $\Lambda$ CDM model, while observations of the ages of cosmic objects offer a complementary way to probe the expansion history.

In this context, machine learning techniques have gained increasing attention in the inference of cosmological observables. This thesis investigates their application to the reconstruction of such quantities, with emphasis on Symbolic Regression (SR). The analysis is based on the combined use of this approach and Gaussian Processes (GP), the latter employed as a non-parametric statistical method, allowing us to assess the consistency and robustness of the reconstructions. We use recent baryon acoustic oscillation data from DESI, combined with Type Ia supernovae, as well as galaxy age data from passively evolving systems.

In the context of dark energy, we reconstruct the equation-of-state parameter  $w(z)$  directly from cosmological observables, aiming to investigate possible evidence for dynamical dark energy. This constitutes one of the first applications of symbolic regression to DESI data for this purpose. The reconstruction is initially performed using  $D_H/r_d$  measurements, where  $D_H$  is the Hubble distance and  $r_d$  is the sound horizon at the drag epoch, and later refined using  $D_M/r_d$ , where  $D_M$  is the transverse comoving distance, combined with supernova data.

Our results show that, given current uncertainties, there is no conclusive evidence for deviations from the  $\Lambda$ CDM model. The function  $w(z)$  exhibits significant sensitivity to cosmological parameters, particularly  $\Omega_m$  (matter density) and  $r_d$ , whose variations impact the inferred behavior. Nevertheless, the reconstructions remain consistent with  $\Lambda$ CDM at the  $3\sigma$  level, with only marginal deviations in specific regions.

In the study of galaxy formation time, we reconstruct  $t_f(z)$ , defined as the cosmic time at which a given fraction of a galaxy's final stellar mass has formed, using two independent approaches based on different datasets. Both yield consistent results and indicate an evolution of  $t_f(z)$  with redshift, particularly at low redshift. We also find that this estimate is weakly dependent on the assumed cosmological model, being mainly driven by uncertainties in cosmological parameters, especially  $H_0$ .

---

## List of Figures

---

|     |                                                                                                                                             |    |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Parallel timeline of key developments in machine learning and cosmology . . . . .                                                           | 3  |
| 1.2 | Constraints in the $w_0$ – $w_a$ plane from DESI-BAO combined with CMB and<br>different SNe datasets . . . . .                              | 4  |
| 2.1 | Spacetime diagram illustrating the concept of luminosity distance . . . . .                                                                 | 13 |
| 2.2 | Preliminary CMB spectrum measured by the FIRAS instrument . . . . .                                                                         | 14 |
| 2.3 | Cartoon of the recombination . . . . .                                                                                                      | 17 |
| 2.4 | Angular power spectrum of the CMB temperature anisotropies from the<br>Planck . . . . .                                                     | 19 |
| 2.5 | Large-scale two-point correlation function $\xi(s)$ obtained from the SDSS<br>luminous red galaxy sample. . . . .                           | 20 |
| 2.6 | Representative light curves of Type Ia supernovae. . . . .                                                                                  | 21 |
| 3.1 | Examples of strong-lens candidates . . . . .                                                                                                | 26 |
| 3.2 | Top symbolic-regression reconstructions . . . . .                                                                                           | 27 |
| 3.3 | Hierarchical structure of AI . . . . .                                                                                                      | 28 |
| 3.4 | Reconstruction of the Hubble parameter $H(z)$ . . . . .                                                                                     | 30 |
| 3.5 | Schematic illustration of the inner loop of PySR . . . . .                                                                                  | 33 |
| 3.6 | Hubble linear relation . . . . .                                                                                                            | 36 |
| 4.1 | Reconstruction of the Hubble function $H(z)$ from CC data for different<br>values of the correlation-length hyperparameter $\ell$ . . . . . | 42 |
| 4.2 | Reconstruction of $H(z)$ and derived quantities from GP . . . . .                                                                           | 45 |
| 5.1 | Comparison of performance across different operator sets using <b>CustomLoss</b> . . . . .                                                  | 52 |
| 5.2 | Comparison of the Mean Squared Error (MSE) distribution per cross-<br>validation fold for the four tested loss functions. . . . .           | 52 |
| 5.3 | PySR reconstruction of $D_H/r_d$ and the derived $w(z)$ . . . . .                                                                           | 54 |
| 5.4 | PySR reconstruction of $D_M/r_d$ and the derived $w(z)$ using DESI + Pan-<br>theon+ dataset . . . . .                                       | 57 |
| 5.5 | PySR reconstruction of $D_M/r_d$ and the derived $w(z)$ using the DESI + DESY5 . . . . .                                                    | 58 |

|      |                                                                                                                                                                |     |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.6  | Reconstruction ( $3\sigma$ ) of $w(z)$ for DESI + Pantheon+ considering variations in $\Omega_m$ and $r_d$ for different scenarios . . . . .                   | 60  |
| 5.7  | Same as Figure 5.6, but using DESI + DESY5. . . . .                                                                                                            | 61  |
| 5.8  | GP reconstruction of $D_M/r_d$ and the derived $w(z)$ using the DESI + pantheon+ using GP . . . . .                                                            | 62  |
| 5.9  | GP reconstruction of $D_M/r_d$ and the derived $w(z)$ using the DESI + DESY5 . . . . .                                                                         | 63  |
| 5.10 | Reconstruction of $w(z)$ using PySR and GP for two combined data sets . . . . .                                                                                | 63  |
| 6.1  | GP and SR reconstructions of the age-redshift relation . . . . .                                                                                               | 69  |
| 6.2  | Age-redshift curves and $t_f(z)$ reconstruction . . . . .                                                                                                      | 71  |
| 6.3  | Age-redshift curves and $t_f(z)$ reconstruction for PySR . . . . .                                                                                             | 72  |
| 6.4  | Cosmic age as a function of redshift for the CPL model and $\Lambda$ CDM (SH0ES) . . . . .                                                                     | 74  |
| 6.5  | Contributions to the uncertainty in the age of the Universe at redshift $z$ from the cosmological parameters $H_0$ , $w$ , $\Omega_M$ and $\Omega_k$ . . . . . | 75  |
| 6.6  | Distance, lookback-time, and formation-time relations . . . . .                                                                                                | 76  |
| 6.7  | Variance contributions of $t_0$ , $t_L$ , and $t_g$ to $t_f(z)$ . . . . .                                                                                      | 78  |
| 6.8  | Comparison of the reconstructed formation time $t_f$ obtained with different approaches and techniques . . . . .                                               | 79  |
| A.1  | The reconstruction results for $D_M/r_d$ using DESI + Pantheon+ dataset . . . . .                                                                              | 102 |
| A.2  | The same as Figure A.1, but using the DESI + DESY5, and a CPL model with the DESI (FS + BAO) + CMB + DESY5 . . . . .                                           | 102 |

---

## List of Tables

---

|     |                                                                                                          |    |
|-----|----------------------------------------------------------------------------------------------------------|----|
| 3.1 | Final model returned by PySR . . . . .                                                                   | 35 |
| 5.1 | BAO measurements from the DESI (DR2) collaboration . . . . .                                             | 49 |
| 5.2 | List of unary and binary operators used in PySR . . . . .                                                | 51 |
| 5.3 | Top 3 PySR results for different values of $\alpha$ . . . . .                                            | 53 |
| 5.4 | DESI-only . . . . .                                                                                      | 55 |
| 5.5 | DESI + Pantheon+. . . . .                                                                                | 57 |
| 5.6 | DESI + DESY5. . . . .                                                                                    | 61 |
| 6.1 | Log-likelihood values for different Gaussian Process kernels. . . . .                                    | 68 |
| 6.2 | Symbolic regression results for $t_g$ . . . . .                                                          | 73 |
| 6.3 | Values of $t_f$ at three redshifts obtained from the different methods discussed<br>in the text. . . . . | 78 |

---

## List of Symbols

---

|                             |                                                                                           |
|-----------------------------|-------------------------------------------------------------------------------------------|
| $C$                         | Observational covariance matrix, p. <a href="#">39</a>                                    |
| $D_H$                       | Hubble distance, p. <a href="#">10</a>                                                    |
| $D_M$                       | Transverse comoving distance, p. <a href="#">12</a>                                       |
| $D_p$                       | Proper distance, p. <a href="#">11</a>                                                    |
| $G$                         | Newton’s gravitational constant, p. <a href="#">8</a>                                     |
| $G_{\mu\nu}$                | Einstein tensor, p. <a href="#">8</a>                                                     |
| $H$                         | Hubble–Lemaître parameter, p. <a href="#">8</a>                                           |
| $H_0$                       | Present-day Hubble constant, p. <a href="#">9</a>                                         |
| $K(\mathbf{X}, \mathbf{X})$ | Covariance matrix evaluated at the observed data points, p. <a href="#">39</a>            |
| $M$                         | Absolute magnitude, p. <a href="#">13</a>                                                 |
| $R$                         | Ricci scalar, p. <a href="#">8</a>                                                        |
| $R_{\mu\nu}$                | Ricci tensor, p. <a href="#">8</a>                                                        |
| $T_{\mu\nu}$                | Energy–momentum tensor, p. <a href="#">8</a>                                              |
| $\Omega_{DE,0}$             | Present-day normalized dark energy density parameter, p. <a href="#">10</a>               |
| $\Omega_{i,0}$              | Present-day normalized density parameter of component $i$ , p. <a href="#">9</a>          |
| $\Omega_{k,0}$              | Present-day normalized curvature density parameter, p. <a href="#">10</a>                 |
| $\Omega_{m,0}$              | Present-day normalized matter density parameter, p. <a href="#">10</a>                    |
| $\chi$                      | Comoving radial distance, p. <a href="#">11</a>                                           |
| $\mathbf{K}^*$              | Cross-covariance matrix between observed and reconstruction points, p. <a href="#">40</a> |
| $\mathbf{K}^{**}$           | Covariance matrix evaluated at reconstruction points, p. <a href="#">40</a>               |

|                                |                                                                                               |
|--------------------------------|-----------------------------------------------------------------------------------------------|
| $\mathbf{X}$                   | Vector of observed input data points, p. <a href="#">39</a>                                   |
| $\mathbf{X}^*$                 | Vector of input points where the function is reconstructed, p. <a href="#">39</a>             |
| $\mathbf{f}^*$                 | Vector of reconstructed function values at test points $\mathbf{X}^*$ , p. <a href="#">40</a> |
| $\mathbf{f}'^*$                | Vector of reconstructed first derivative values at $\mathbf{X}^*$ , p. <a href="#">43</a>     |
| $\mathbf{y}$                   | Vector of observed data values, p. <a href="#">39</a>                                         |
| $\mathcal{GP}$                 | Gaussian Process, p. <a href="#">39</a>                                                       |
| $\mathcal{L}$                  | Likelihood function of the observed data, p. <a href="#">41</a>                               |
| $\mathcal{N}$                  | Multivariate Gaussian (normal) distribution, p. <a href="#">39</a>                            |
| $\text{cov}(\mathbf{f}^*)$     | Covariance matrix of the reconstructed function, p. <a href="#">41</a>                        |
| $\text{cov}(f^{(i)}, f^{(j)})$ | Covariance between the $i$ -th and $j$ -th derivatives of the function, p. <a href="#">44</a> |
| $\mu$                          | Distance modulus, p. <a href="#">13</a>                                                       |
| $\mu(x)$                       | Mean function of the Gaussian Process, p. <a href="#">39</a>                                  |
| $\rho$                         | Mass density, p. <a href="#">8</a>                                                            |
| $\rho_m$                       | Matter density, p. <a href="#">9</a>                                                          |
| $\rho_r$                       | Radiation density, p. <a href="#">9</a>                                                       |
| $\rho_{DE,0}$                  | Present-day dark energy density, p. <a href="#">9</a>                                         |
| $\rho_{DE}$                    | Dark energy density, p. <a href="#">9</a>                                                     |
| $\rho_{c0}$                    | critical density today, p. <a href="#">9</a>                                                  |
| $\rho_{m,0}$                   | Present-day matter density, p. <a href="#">9</a>                                              |
| $\rho_{r,0}$                   | Present-day radiation density, p. <a href="#">9</a>                                           |
| $\sigma_i^2$                   | Variance of the $i$ -th observational data point, p. <a href="#">40</a>                       |
| $\tau$                         | proper time, p. <a href="#">8</a>                                                             |
| $\tilde{\mathbf{K}}$           | Total covariance matrix including observational errors, p. <a href="#">40</a>                 |
| $a(t)$                         | Scale Factor, p. <a href="#">7</a>                                                            |
| $c$                            | Speed of light, p. <a href="#">7</a>                                                          |

|                   |                                                                                                       |
|-------------------|-------------------------------------------------------------------------------------------------------|
| $d$               | Physical distance, p. <a href="#">10</a>                                                              |
| $f(x)$            | Latent function reconstructed with Gaussian Processes, p. <a href="#">38</a>                          |
| $g(f, f', \dots)$ | Derived observable depending on the reconstructed function and its derivatives, p. <a href="#">44</a> |
| $k(x, \tilde{x})$ | Covariance (kernel) function of the Gaussian Process, p. <a href="#">39</a>                           |
| $m$               | Apparent magnitude, p. <a href="#">13</a>                                                             |
| $n$               | Number of observed data points, p. <a href="#">40</a>                                                 |
| $n^*$             | Number of reconstruction points, p. <a href="#">40</a>                                                |
| $p$               | Pressure of the fluid, p. <a href="#">8</a>                                                           |
| $q_0$             | Present-day deceleration parameter, p. <a href="#">10</a>                                             |
| $r_d$             | sound horizon at the drag epoch, p. <a href="#">48</a>                                                |
| $t_0$             | Present cosmic time, p. <a href="#">10</a>                                                            |
| $u_\mu$           | Four-velocity, p. <a href="#">8</a>                                                                   |
| $K$               | Curvature Parameter, p. <a href="#">7</a>                                                             |

---

## List of Abbreviations

---

|               |                                                                |
|---------------|----------------------------------------------------------------|
| $\Lambda$ CDM | $\Lambda$ -Cold Dark Matter, p. <a href="#">3</a>              |
| BAO           | Baryon Acoustic Oscillation, p. <a href="#">3</a>              |
| BA            | Barboza Alcaniz, p. <a href="#">4</a>                          |
| BOSS          | Baryon Oscillation Spectroscopic Survey, p. <a href="#">3</a>  |
| CMB           | Cosmic Microwave Background, p. <a href="#">1</a>              |
| CNN           | Convolutional Neural Networks, p. <a href="#">25</a>           |
| COBE          | Cosmic Background Explorer, p. <a href="#">16</a>              |
| CPL           | Chevallier-Polarski-Linder, p. <a href="#">3</a>               |
| CP            | Cosmological Principle, p. <a href="#">7</a>                   |
| DESI          | Dark Energy Spectroscopic Instrument, p. <a href="#">3</a>     |
| DES5Y         | Dark Energy Survey Supernova 5-Year, p. <a href="#">3</a>      |
| DETF          | Dark Energy Task Force, p. <a href="#">3</a>                   |
| ETG           | Early-Type Galaxies, p. <a href="#">5</a>                      |
| FIRAS         | Far InfraRed Absolute Spectrophotometer, p. <a href="#">16</a> |
| FLRW          | Friedmann–Lemaître–Robertson–Walker, p. <a href="#">7</a>      |
| GA            | Genetic Algorithms, p. <a href="#">26</a>                      |
| GR            | General Relativity, p. <a href="#">7</a>                       |
| MAE           | Mean Absolute Error, p. <a href="#">29</a>                     |
| MDL           | minimum description length, p. <a href="#">26</a>              |
| MLP           | Multi-Layer Perceptrons, p. <a href="#">28</a>                 |



|      |                                                      |
|------|------------------------------------------------------|
| MSE  | Mean Squared Error, p. <a href="#">29</a>            |
| ODE  | Ordinary Differential Equation, p. <a href="#">9</a> |
| PCA  | Principal Component Analysis, p. <a href="#">2</a>   |
| RMSE | Root Mean Squared Error, p. <a href="#">29</a>       |
| SDSS | Sloan Digital Sky Survey, p. <a href="#">3</a>       |
| SR   | Symbolic Regression, p. <a href="#">6</a>            |
| SVM  | Support Vector Machine, p. <a href="#">28</a>        |
| SVR  | Support Vector Regression, p. <a href="#">28</a>     |

---

# Contents

---

|                                                                       |             |
|-----------------------------------------------------------------------|-------------|
| <b>List of Figures</b>                                                | <b>xi</b>   |
| <b>List of Tables</b>                                                 | <b>xiii</b> |
| <b>List of Symbols</b>                                                | <b>xiv</b>  |
| <b>List of Abbreviations</b>                                          | <b>xvii</b> |
| <b>1 Introduction</b>                                                 | <b>1</b>    |
| <b>2 Theoretical Foundations of Modern Cosmology</b>                  | <b>7</b>    |
| 2.1 Dynamics of the Universe . . . . .                                | 7           |
| 2.2 Cosmological Distances . . . . .                                  | 10          |
| 2.2.1 Hubble Distance . . . . .                                       | 10          |
| 2.2.2 Proper Distance . . . . .                                       | 11          |
| 2.2.3 Radial comoving distance . . . . .                              | 11          |
| 2.2.4 Transversal comoving distance . . . . .                         | 12          |
| 2.2.5 Luminosity distance . . . . .                                   | 12          |
| 2.3 Cosmological Observational Probes . . . . .                       | 14          |
| 2.3.1 Cosmic Microwave Background . . . . .                           | 14          |
| 2.3.2 Baryon Acoustic Oscillations . . . . .                          | 18          |
| 2.3.3 Type Ia Supernovae . . . . .                                    | 21          |
| 2.3.4 Age and lookback time . . . . .                                 | 23          |
| <b>3 Supervised learning in Cosmology</b>                             | <b>25</b>   |
| 3.1 Introduction . . . . .                                            | 25          |
| 3.2 Supervised Regression in Cosmology: Conceptual Overview . . . . . | 27          |
| 3.3 Symbolic Regression . . . . .                                     | 31          |
| 3.4 Challenges and Perspectives . . . . .                             | 37          |
| <b>4 Gaussian Processes</b>                                           | <b>38</b>   |
| 4.1 Introduction . . . . .                                            | 38          |
| 4.2 Reconstructing a function . . . . .                               | 38          |

|          |                                                                                        |            |
|----------|----------------------------------------------------------------------------------------|------------|
| 4.3      | Reconstructing the derivative of a function . . . . .                                  | 43         |
| 4.4      | Functions of a GP and Error Propagation . . . . .                                      | 44         |
| 4.5      | A Practical Example: Reconstruction of $H(z)$ via GP . . . . .                         | 45         |
| <b>5</b> | <b>Symbolic regression analysis of dynamical dark energy with DESI-DR2 and SN data</b> | <b>47</b>  |
| 5.1      | Introduction . . . . .                                                                 | 47         |
| 5.2      | Basic Equations . . . . .                                                              | 47         |
| 5.3      | Data . . . . .                                                                         | 48         |
| 5.3.1    | Baryonic acoustic oscillations . . . . .                                               | 49         |
| 5.3.2    | Type Ia Supernova . . . . .                                                            | 50         |
| 5.3.2.1  | Pantheon+ . . . . .                                                                    | 50         |
| 5.3.2.2  | Dark Energy Survey Supernova Year 5 (DESY5) . . . . .                                  | 50         |
| 5.4      | Symbolic regression . . . . .                                                          | 50         |
| 5.4.1    | Operators Used in PySR . . . . .                                                       | 50         |
| 5.4.2    | Loss Function . . . . .                                                                | 51         |
| 5.4.3    | Model Selection . . . . .                                                              | 53         |
| 5.5      | Results . . . . .                                                                      | 54         |
| 5.5.1    | DESI-only . . . . .                                                                    | 54         |
| 5.5.2    | DESI + Pantheon+ . . . . .                                                             | 56         |
| 5.5.3    | DESI + DESY5 . . . . .                                                                 | 58         |
| 5.6      | Gaussian processes . . . . .                                                           | 62         |
| 5.7      | Conclusion . . . . .                                                                   | 64         |
| <b>6</b> | <b>Cosmological constraints on galaxy formation time</b>                               | <b>66</b>  |
| 6.1      | Introduction . . . . .                                                                 | 66         |
| 6.2      | Galaxy formation time . . . . .                                                        | 66         |
| 6.3      | Methodology and Data . . . . .                                                         | 67         |
| 6.3.1    | Gaussian Processes . . . . .                                                           | 68         |
| 6.3.2    | High-performance Symbolic Regression . . . . .                                         | 68         |
| 6.3.3    | Age Data . . . . .                                                                     | 69         |
| 6.3.4    | Type Ia Supernova Data . . . . .                                                       | 70         |
| 6.4      | Results . . . . .                                                                      | 70         |
| 6.5      | Robustness tests of the formation time $t_f$ . . . . .                                 | 78         |
| 6.6      | Conclusion . . . . .                                                                   | 80         |
| <b>7</b> | <b>Conclusions</b>                                                                     | <b>81</b>  |
|          | <b>Bibliography</b>                                                                    | <b>84</b>  |
| <b>A</b> | <b>Analyses with DESI DR1</b>                                                          | <b>101</b> |

# CHAPTER 1

---

## Introduction

---

Cosmology emerged as a result of human curiosity about the universe and the desire to comprehend its beginnings, development, and final destiny. With the development of General Relativity (GR; EINSTEIN, 1915) at the start of the 20th century, contemporary cosmology became a scientific field, transcending purely philosophical conjecture and building a framework based on theoretical and observational principles. This advancement made it possible to apply physical rules to scales that are similar to the size and age of the observable universe.

Currently, the standard cosmological model ( $\Lambda$ CDM), based on General Relativity and the Cosmological Principle, provides an excellent description of the evolution of the Universe. It is strongly supported by several key observational results, including the expansion of the Universe inferred from the recession velocities of distant galaxies (HUBBLE, 1929), the primordial abundances of light elements predicted by Big Bang nucleosynthesis (GAMOW, 1946), and the existence of the Cosmic Microwave Background (CMB) radiation (PENZIAS and WILSON, 1965), as well as the late-time accelerated expansion of the Universe revealed by Type Ia supernova observations (PERLMUTTER *et al.*, 1999; RIESS *et al.*, 1998). The observed cosmic acceleration is commonly interpreted as the manifestation of a dominant cosmic component characterized by a negative equation-of-state (EoS) parameter  $w$ ; within the standard cosmological model, this component is identified with a cosmological constant  $\Lambda$ , for which  $w = -1$ .

Despite its remarkable success, the  $\Lambda$ CDM model faces several open challenges. Among them are the unknown nature of dark energy and the persistent Hubble tension (KAMIONKOWSKI and RIESS, 2022; VERDE *et al.*, 2023). In addition, there are mild discrepancies in the clustering amplitude, commonly referred to as the  $S_8$  tension (HILDEBRANDT *et al.*, 2017; WRIGHT *et al.*, 2025), which may indicate residual inconsistencies between different cosmological probes, although with lower statistical significance. Furthermore, questions related to neutrino physics, such as constraints on their total mass (ELBERS *et al.*, 2025), remain an active area of investigation.

The increasing precision and volume of cosmological data have demanded the devel-

opment of more sophisticated analysis techniques to address these challenges. A range of methodological techniques, from theory-driven models to more adaptable, data-driven frameworks, have been used in this setting. These methods seek to give reliable and objective reconstructions of cosmic observables in addition to improving parameter estimation, allowing for a better comprehension of the underlying physical processes.

Parametric and non-parametric methods represent the two primary methodological approaches used in the analysis of cosmological observations. The parametric approach consists of either introducing a parametrization for the quantity of interest or assuming a specific functional form motivated by theoretical models. This strategy is particularly effective in constraining model parameters, often leading to tighter uncertainties. However, the results may be biased by the choice of functional form and the underlying assumptions.

Non-parametric approaches provide a more flexible alternative to mitigate these limitations. Rather than assuming a predefined functional form, these methods aim to reconstruct smooth functions directly from the data. Their objective is to capture the underlying behavior of the observables while minimizing theoretical bias.

Several non-parametric techniques have been widely employed in cosmology, such as Gaussian Processes (GP; RASMUSSEN and WILLIAMS, 2006; SEIKEL *et al.*, 2012), Principal Component Analysis (PCA; PEARSON, 1901), and smoothing splines (SIMONOFF, 2012). For a more detailed discussion of non-parametric techniques in cosmology, see GONZALEZ (2018). More recently, machine learning techniques have naturally taken on this role, providing flexible and scalable frameworks for reconstructing cosmological quantities and extracting information from increasingly complex datasets.

In this context, machine learning can be viewed as a modern extension of non-parametric approaches, providing enhanced flexibility and the ability to model complex, high-dimensional datasets. However, despite its early theoretical foundations dating back to the mid-20th century, its widespread application in scientific problems is relatively recent, largely driven by advances in computational power and the availability of large datasets.

This aspect becomes particularly evident when compared to the development of modern cosmology. While the theoretical framework of machine learning was already being established, cosmology had already undergone major conceptual and observational breakthroughs, including the formulation of the Big Bang model and, later, the discovery of the accelerated expansion of the Universe. Only in more recent decades has machine learning become a practical and competitive tool in the analysis of cosmological data.

Although cosmology and machine learning originated as distinct fields, their parallel evolution has led to a growing synergy, with artificial intelligence now playing a central role in addressing some of the most challenging open problems in cosmology. This parallel development is illustrated in Fig. 1.1, which highlights key milestones in both fields and their progressive convergence.

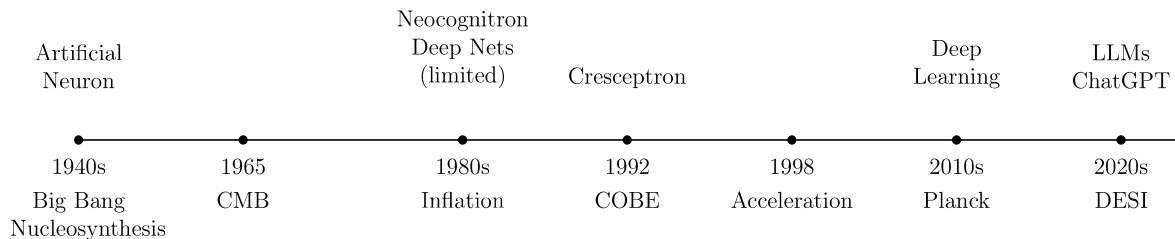


Figure 1.1: Parallel timeline of key developments in machine learning and cosmology, highlighting their independent origins and subsequent convergence in the era of data-driven science.

In light of these considerations, two representative scenarios illustrate the application of these ideas in modern cosmology. The first concerns recent results from the Dark Energy Spectroscopic Instrument (DESI) and their implications for dynamical dark energy, while the second focuses on the problem of galaxy formation time. In both cases, machine learning techniques are employed as complementary tools to aid the analysis and interpretation of the data.

Initiatives such as the Dark Energy Task Force (DETF; ALBRECHT *et al.*, 2006) outlined a roadmap to improve our understanding of dark energy by classifying cosmological surveys into distinct stages according to their precision and constraining power. Early surveys, such as the Sloan Digital Sky Survey (SDSS), are typically associated with Stage II experiments, providing the first robust measurements of large-scale structure and baryon acoustic oscillations. Subsequent Stage III surveys, including Baryon Oscillation Spectroscopic Survey (BOSS) and eBOSS, significantly improved these constraints through larger volumes and higher-quality data. These efforts paved the way for the current generation of Stage IV experiments (see, e.g., BRAUN *et al.* (2015); LSST DARK ENERGY SCIENCE COLLABORATION (2012); SCARAMELLA *et al.* (2022)), which are designed to achieve substantially higher precision. Among these, DESI (AGHAMOUSA *et al.*, 2016) stands out as one of the leading efforts.

The DESI collaboration, in its first data release in 2024 of baryon acoustic oscillation (BAO) measurements, provided important results which challenge the  $\Lambda$ -Cold Dark Matter ( $\Lambda$ CDM) paradigm (ADAME *et al.*, 2024a), as further intensified by subsequent works (BOUSIS and PERIVOLAROPOULOS, 2024; CORTÊS and LIDDLE, 2024; LUONGO and MUCCINO, 2024; PARK *et al.*, 2024; YIN, 2024). Assuming that the dark energy EoS is given by the Chevallier-Polarski-Linder parameterization (CPL; CHEVALLIER and POLARSKI, 2001; LINDER, 2003),  $w(z) = w_0 + w_a z/(1+z)$ , some analyses of the DESI-BAO data, when combined with various Type Ia Supernova (SNe) datasets, such as the Dark Energy Survey Supernova 5-Year compilation (DESY5) ABBOTT *et al.* (2024), Union3 (RUBIN *et al.*, 2024), and Pantheon+ (SCOLNIC *et al.*, 2022), in addition to CMB priors from Planck (AGHANIM *et al.*, 2020), showed a preference for a dynamical dark energy (with  $w_0 > -1$  and  $w_a < 0$ ) compared to  $\Lambda$ CDM (see Fig. 1.2;

see also EFSTATHIOU (2025); ROY CHOUDHURY and OKUMURA (2024) for different perspectives).

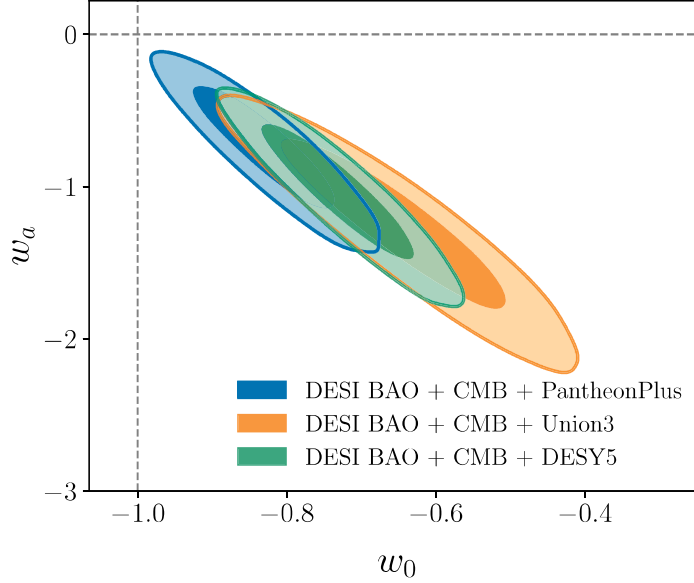


Figure 1.2: Constraints in the  $w_0$ - $w_a$  plane from DESI-BAO combined with CMB and different SNe datasets. The  $\Lambda$ CDM model corresponds to  $(w_0, w_a) = (-1, 0)$ . Reproduced from ADAME *et al.* (2024a).

Additionally, this conclusion is supported by analyses that explore different EoS parameterizations (BHATTACHARYA *et al.*, 2024; DE SOUZA *et al.*, 2025; GHOSH and BENGALY, 2024b; GIARÈ *et al.*, 2024; RAMADAN *et al.*, 2024; REBOUÇAS *et al.*, 2024; RODRIGUES *et al.*, 2025; WOLF *et al.*, 2024), with the best-fit model provided by the Barboza-Alcaniz parameterization (BA, BARBOZA and ALCANIZ, 2008),  $w(z) = w_0 + w_a z(1 + z)/(1 + z^2)$ .

It is worth emphasizing that these indications emerge within specific parametric descriptions of the dark energy equation of state. The CPL and BA parametrizations, for instance, provide a convenient phenomenological expansion of  $w(z)$  around the present epoch, but they impose a predefined functional structure on the redshift evolution. As with any low-dimensional parametrization, this choice restricts the space of possible behaviors and may influence the inferred dynamical features.

This naturally raises the question of whether the apparent preference for an evolving dark energy component reflects genuine physical information encoded in the data or is, at least in part, driven by the assumed functional form. In this sense, approaches that relax parametric assumptions become particularly valuable.

From this perspective, non-parametric analyses (DINDA and MAARTENS, 2024; GHOSH and BENGALY, 2024a; JIANG *et al.*, 2024; MUKHERJEE and SEN, 2024) provide a complementary strategy, as they allow one to investigate a possible redshift evolution of the dark energy equation of state without imposing a specific functional form

for  $w(z)$ . Methods such as Gaussian Processes (GP, RASMUSSEN and WILLIAMS, 2006; SEIKEL *et al.*, 2012) and Genetic Algorithms (GA, ARJONA and NESSERIS, 2020; BOGDANOS and NESSERIS, 2009) have been widely employed for this purpose, offering a flexible framework for reconstructing cosmological functions in a largely model-independent manner.

Interestingly, some recent non-parametric reconstructions of the DESI-BAO measurements, combined with SNe datasets and CMB priors (DINDA and MAARTENS, 2024), have reported a smaller tension with the  $\Lambda$ CDM model than those obtained within the CPL framework in ADAME *et al.* (2024a). This difference further emphasizes how methodological assumptions can influence the inferred dynamical behavior of dark energy.

In a different but complementary direction, the formation time of galaxies constitutes an important probe of cosmic evolution. Defined as the time elapsed between the Big Bang and the epoch at which a galaxy assembles most of its stellar mass, this quantity is intrinsically related to both the age of galaxies and the age of the Universe.

Galaxy formation itself represents a central theme in cosmology and has been extensively investigated through both theoretical modeling and observational data. In this context, the formation time,  $t_f$ , plays a key role, as it provides a direct link between observations of galaxy ages and the underlying cosmological model. This timescale is crucial for constraining scenarios of cosmic structure formation and for testing the consistency of the current cosmological paradigm, thereby offering valuable insights into galactic evolution and its connection to the expansion history of the Universe.

Over the past few decades, our understanding of galaxy formation has advanced significantly, based on both observations and theoretical progress. Seminal works have profoundly shaped our knowledge of galaxy formation and evolution, including those by BLUMENTHAL *et al.* (1984); EGGEN *et al.* (1962); PEEBLES (1965); PEEBLES (1984); PRESS and SCHECHTER (1974); WHITE and REES (1978). These foundational studies elucidated vital mechanisms governing the formation of cosmic structures and introduced key concepts such as hierarchical clustering, the decisive role of dark matter, and the importance of cooling processes in baryonic matter. They continue to be essential references for comprehending the timing and mechanisms of galaxy formation within the broader cosmological context.

In principle, galaxy formation does not necessarily occur on a universal timescale; instead, it may vary with redshift and environment. This study aims to investigate the  $t_f$  redshift dependence and characterize its behavior throughout cosmic evolution. Here, we specifically focus on a particular class of galaxies, namely, passively evolving galaxies, like early-type galaxies (ETG). These objects exhibit little to no ongoing star formation, are dominated by older stellar populations that evolve slowly, and possess well-characterized colors, spectra, and luminosities. Their age can be calculated by modeling the stellar populations using either simple or composite stellar population models. Although our



calculations rely on observational data of ETGs, it is important to acknowledge that the method employed is not entirely model-independent; the formation time can only be inferred from galaxy ages by assuming a cosmological model.

The main objective of this thesis is to investigate the applicability and robustness of supervised learning techniques — particularly Symbolic Regression (SR) implemented through the PySR framework — for the reconstruction of cosmological observables and the derivation of secondary physical quantities. By allowing the data to determine the functional form of the reconstructed relations, this approach seeks to reduce prior assumptions about the underlying cosmological model and to provide flexible yet physically interpretable descriptions of the data.

A central component of this work is the joint application of Symbolic Regression and Gaussian Processes to the same cosmological observables. While Gaussian Processes provide a statistically well-established and fully probabilistic reconstruction framework, Symbolic Regression yields analytic functional forms with explicit interpretability. By analyzing both approaches side by side, we examine the consistency of the inferred cosmological behavior and assess the stability and robustness of the reconstructed results across different methodologies. In this way, Gaussian Processes serve both as a complementary reconstruction technique and as a statistical reference against which the performance and reliability of Symbolic Regression can be critically evaluated.

The remainder of this thesis is organized as follows. Chapter 2 reviews the theoretical foundations of modern cosmology relevant to this work, focusing on the dynamics of the expanding Universe, the cosmological redshift, and the main distance measures used in observational cosmology, as well as the relation between redshift, cosmic age, and lookback time. Chapter 3 provides an overview of machine learning techniques and discusses their growing role in cosmological data analysis. In Chapter 4, we introduce Gaussian Processes as a non-parametric reconstruction method, outlining their formalism and applicability to cosmological inference. Chapter 5 presents the symbolic regression reconstruction of cosmological observables using recent DESI measurements, with particular emphasis on the dark energy EoS parameter  $w(z)$ . In Chapter 6, we investigate the formation time of galaxies through two independent reconstruction approaches. Finally, Chapter 7 summarizes the main results and discusses future perspectives.

# CHAPTER 2

---

## Theoretical Foundations of Modern Cosmology

---

In this chapter, we review the theoretical foundations of modern cosmology, focusing on the dynamics of the expanding Universe, the cosmological redshift, and the main distance measures employed in observational cosmology. We also discuss the relation between redshift, cosmic age, and lookback time. The material presented here follows closely the standard treatments found in the literature (BAUMANN, 2022; HOGG, 1999; WEINBERG, 1972).

### 2.1 Dynamics of the Universe

To understand the dynamics of the Universe, we assume that it is governed by General Relativity (GR) and satisfies the Cosmological Principle (CP). CP means that it is homogeneous, which means that the Universe appears the same in every point, and isotropic, meaning that its properties are independent of direction. A homogeneous Universe does not imply an isotropic Universe. Assuming that the Universe is filled with a perfect fluid on large scales, the Friedmann–Lemaître–Robertson–Walker (FLRW) metric can describe the distances between objects, given by

$$ds^2 = c^2 dt^2 - a^2(t) \left[ \frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2 \theta d\phi^2) \right] \quad (2.1)$$

where  $a(t)$  is the scale factor, which quantifies the change in distances due to the expansion of the Universe, and  $K$  is the curvature parameter, which determines the geometry of spatial sections:  $K = 0$  corresponds to a flat space,  $K = 1$  to a closed space, and  $K = -1$  to an open space.

General Relativity describes how the geometry of spacetime responds to the presence of matter and energy. This relationship between curvature and the energy–momentum content of the Universe is encoded in the field equations formulated by Einstein, expressed as

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{8\pi G T_{\mu\nu}}{c^4}, \quad (2.2)$$

where  $G_{\mu\nu}$  is the Einstein tensor,  $R_{\mu\nu}$  is the Ricci tensor,  $R$  is the Ricci scalar defined by  $R \equiv g^{\mu\nu}R_{\mu\nu}$ ,  $G$  is Newton's gravitational constant, and  $T_{\mu\nu}$  is the energy-momentum tensor. For a perfect fluid, we have

$$T_{\mu\nu} = (\rho c^2 + p) \frac{u_\mu u_\nu}{c^2} - p g_{\mu\nu}, \quad (2.3)$$

in which  $p$  is the pressure of the fluid,  $\rho$  is the mass density of the fluid, and  $u_\mu$  is the four-velocity of the fluid, given by  $u_\mu = dx_\mu/d\tau$ , where  $\tau$  is the proper time (see LANDAU and LIFSHITZ (1987)). For a comoving reference frame, the four-velocity is  $u^\mu = (c, 0, 0, 0)$ . By substituting the FLRW metric (2.1) into the Einstein equations (2.2), one obtains two independent equations,

$$H^2 = \frac{8\pi G}{3} \rho - \frac{c^2 K}{a^2}, \quad (2.4)$$

where the Hubble-Lemaître parameter is defined as

$$H \equiv \frac{\dot{a}}{a}.$$

The second independent equation is given by

$$\frac{\ddot{a}(t)}{a(t)} = -\frac{4\pi G}{3} \left[ \rho(t) + \frac{3p(t)}{c^2} \right]. \quad (2.5)$$

Furthermore, the energy conservation equation can be obtained by taking the time derivative of Equation (2.4) and substituting Equation (2.5) into it,

$$\dot{\rho}(t) + 3H \left( \rho(t) + \frac{p}{c^2} \right) = 0. \quad (2.6)$$

By defining the equation of state parameter  $w \equiv p/\rho c^2$ , which connects pressure  $p$  to energy density  $\rho c^2$ , the conservation equation can be reformulated as

$$\dot{\rho}(t) + 3H (1 + w) \rho(t) = 0. \quad (2.7)$$

The redshift  $z$  is defined as the ratio between the observed and emitted wavelengths of light,  $\lambda_0$  and  $\lambda$ , respectively, and can be written as

$$1 + z \equiv \frac{\lambda_0}{\lambda} = \frac{a(t_0)}{a(t)}, \quad (2.8)$$

where  $a(t)$  is the scale factor. This relation encodes the expansion of the Universe, with increasing (decreasing)  $a(t)$  corresponding to redshift (blueshift). This relation can be

connected to the infinitesimal variation of the redshift of a given object during a time interval  $dt$  as follows:

$$\frac{d}{dt} = \frac{dz}{dt} \frac{d}{dz}, \quad \frac{dz}{dt} = -(1+z) \frac{\dot{a}}{a}, \quad (2.9)$$

thus Eq. (2.7) becomes

$$-(1+z) \frac{\dot{a}}{a} \frac{d\rho}{dz} + 3 \frac{\dot{a}}{a} (1+w) \rho = 0 \quad \Rightarrow \quad \frac{d\rho}{dz} = \frac{3(1+w)}{1+z} \rho. \quad (2.10)$$

This is a separable ordinary differential equation (ODE); integrating from  $z = 0$  (today) to  $z$  yields

$$\rho(z) = \rho_0 \exp \left[ 3 \int_0^z \frac{1+w(z')}{1+z'} dz' \right]. \quad (2.11)$$

The concordance cosmological model,  $\Lambda$ CDM, posits that matter, radiation, and dark energy are the three primary components that constitute the universe's material content. Matter may be either baryonic or non-relativistic (cold dark matter). Using the equation above, different components evolve according to:

- For non-relativistic matter ( $w = 0$ ), the integral yields

$$\rho_m(z) = \rho_{m,0} (1+z)^3.$$

- For radiation ( $w = 1/3$ ), the density evolves as

$$\rho_r(z) = \rho_{r,0} (1+z)^4.$$

- For a cosmological constant ( $w = -1$ ), the exponential factor becomes unity, leading to

$$\rho_{DE}(z) = \rho_{DE,0}.$$

Returning to the Equation (2.4), we can rewrite the Friedmann equation with the sum of the components,

$$H(z)^2 = \frac{8\pi G}{3} \sum_i \rho_i(z) - \frac{c^2 K}{a^2}. \quad (2.12)$$

Another important quantity is the present-day critical density,  $\rho_{c,0}$ , together with the corresponding normalized density parameters evaluated at the present epoch,  $\Omega_{i,0}$ ,

$$\rho_{c,0} \equiv \frac{3H_0^2}{8\pi G}, \quad \Omega_{i,0} \equiv \frac{\rho_{i,0}}{\rho_{c,0}}.$$

Here,  $H_0$  denotes the Hubble constant. With these definitions, the Friedmann equation

can be conveniently written in terms of the normalized Hubble parameter as

$$\frac{H(z)^2}{H_0^2} = \Omega_{m,0}(1+z)^3 + \Omega_{DE,0} \exp \left[ 3 \int_0^z \frac{1+w(z')}{1+z'} dz' \right] + \Omega_{k,0}(1+z)^2, \quad (2.13)$$

where  $\Omega_{k,0} = -c^2 K / H_0^2 a_0^2$  denotes the normalized curvature density parameter evaluated today. We will return to this general expression and discuss its implications in more detail in a later section.

## 2.2 Cosmological Distances

Distance measurements in cosmology constitute a non-trivial task. In an expanding Universe, the distances between comoving objects evolve continuously. For this reason, several distinct definitions of distance are employed in cosmology.

### 2.2.1 Hubble Distance

The first distance we introduce is the *Hubble distance*,  $D_H$ . To this end, we recall that the cosmological redshift, previously defined in Eq. (2.8), is given by

$$1+z \equiv \frac{\lambda_0}{\lambda} = \frac{a(t_0)}{a(t)}, \quad (2.14)$$

where  $t_0$  denotes the present cosmic time. For nearby sources ( $z \ll 1$ ), the scale factor can be expanded around  $t_0$  as

$$a(t) = a(t_0) \left[ 1 - H_0(t_0 - t) - \frac{1}{2} q_0 H_0^2 (t_0 - t)^2 + \mathcal{O}((t_0 - t)^3) \right], \quad (2.15)$$

where  $q_0 \equiv -\ddot{a}(t_0)/[a(t_0)H_0^2]$  is the present-day deceleration parameter. Substituting this expansion into the definition of the redshift yields

$$z \simeq H_0(t_0 - t) + \frac{1}{2}(1+q_0)H_0^2(t_0 - t)^2 + \mathcal{O}((t_0 - t)^3). \quad (2.16)$$

For sufficiently small distances, the spacetime can be approximated as locally Minkowskian. At leading order in  $t_0 - t$ , the light travel time satisfies  $d \simeq c(t_0 - t)$ , which naturally introduces the Hubble distance,

$$D_H \equiv \frac{c}{H_0}. \quad (2.17)$$

The Hubble distance does not represent a directly observable separation, but rather defines a characteristic length scale associated with the present expansion rate.

### 2.2.2 Proper Distance

In the Hubble–Lemaître law, the distance  $d$  corresponds to the *proper distance* between comoving objects evaluated at the present cosmic time. This notion of distance follows directly from the FLRW metric given in Eq. (2.1). For radial separations at fixed cosmic time ( $dt = d\theta = d\phi = 0$ ), the line element reduces to

$$ds = a(t) \frac{dr}{\sqrt{1 - Kr^2}}. \quad (2.18)$$

The proper distance between two comoving points is then obtained by integrating the spatial line element,

$$D_p(t) = a(t) \int_0^r \frac{dr'}{\sqrt{1 - Kr'^2}} \equiv a(t) \chi(r), \quad (2.19)$$

where  $\chi(r)$  denotes the comoving radial distance. By construction,  $\chi(r)$  is time-independent for comoving observers. The time dependence of  $D_p(t)$  reflects the expansion of the Universe through the scale factor  $a(t)$ .

### 2.2.3 Radial comoving distance

The comoving distance provides a natural measure of spatial separations in an expanding Universe. For two nearby objects moving with the Hubble flow, the comoving separation is defined as the distance that remains constant with cosmic time. Equivalently, it corresponds to the proper distance measured at a given epoch divided by the scale factor at that time or, using the definition of redshift given in Eq. (2.8), to the proper distance multiplied by  $(1 + z)$  (HOGG, 1999).

This notion can be made precise by considering the propagation of light in an expanding spacetime. The *radial comoving distance* is determined by the propagation of photons along radial null geodesics. Setting  $ds^2 = 0$  and  $d\theta = d\phi = 0$  in the FLRW metric, Eq. (2.1), one finds

$$c dt = a(t) \frac{dr}{\sqrt{1 - Kr^2}}. \quad (2.20)$$

Introducing the comoving radial coordinate,

$$d\chi \equiv \frac{dr}{\sqrt{1 - Kr^2}}, \quad (2.21)$$

The radial comoving distance to a source emitting light at cosmic time  $t$  is given by

$$\chi = \int_t^{t_0} \frac{c dt'}{a(t')}. \quad (2.22)$$

Using Eq. (2.8) to perform a change of integration variable from cosmic time to redshift,

together with the definition of the Hubble parameter, this expression can be rewritten as

$$\chi(z) = c \int_0^z \frac{dz'}{H(z')}. \quad (2.23)$$

Writing the Hubble parameter as  $H(z) = H_0 E(z)$ , where  $E(z)$  is a dimensionless function encoding the expansion history of the Universe, one can rewrite Eq. (2.13) as

$$E(z)^2 = \Omega_{m0}(1+z)^3 + \Omega_{DE0} \exp \left[ 3 \int_0^z \frac{1+w(z')}{1+z'} dz' \right] + \Omega_{k0}(1+z)^2. \quad (2.24)$$

In this form, the radial comoving distance can be expressed as

$$\chi(z) = D_H \int_0^z \frac{dz'}{E(z')}, \quad (2.25)$$

where the Hubble distance  $D_H = c/H_0$  arises naturally as the characteristic length scale set by the present expansion rate of the Universe. The radial comoving distance  $\chi(z)$  plays a fundamental role in cosmography, as all other distance measures can be expressed in terms of it. Physically, it represents the distance that would be measured today between two nearby events if they were comoving with the Hubble flow, making it the appropriate scale for describing large-scale structures in the Universe (HOGG, 1999).

### 2.2.4 Transversal comoving distance

The *transverse comoving distance*,  $D_M$ , also known as the proper motion distance in the literature (WEINBERG, 1972), relates transverse physical separations to observed angular separations.

$$D_M(z) = \begin{cases} \frac{D_H}{\sqrt{\Omega_k}} \sinh \left( \sqrt{\Omega_k} \frac{\chi(z)}{D_H} \right), & \Omega_k > 0, \\ \chi(z), & \Omega_k = 0, \\ \frac{D_H}{\sqrt{|\Omega_k|}} \sin \left( \sqrt{|\Omega_k|} \frac{\chi(z)}{D_H} \right), & \Omega_k < 0. \end{cases} \quad (2.26)$$

In a spatially flat Universe ( $\Omega_k = 0$ ), the transverse and radial comoving distances coincide,  $D_M(z) = \chi(z)$ , while spatial curvature introduces a non-trivial mapping between them for  $\Omega_k \neq 0$ .

### 2.2.5 Luminosity distance

In a Euclidean space, the apparent luminosity of a source at rest at a distance  $d$  is given by  $L/(4\pi d^2)$ . By analogy, following WEINBERG (1972), one may define the *luminosity*

distance  $D_L$  of a light source in a general spacetime through the relation

$$D_L = \sqrt{\frac{L}{4\pi F}} \quad (2.27)$$

In a FLRW spacetime, the luminosity distance is related to the transverse comoving distance  $D_M$  by

$$D_L = (1 + z) D_M, \quad (2.28)$$

where the factor  $(1 + z)$  accounts for the redshift of photon energies and the time dilation of the emission rate, as schematically illustrated in Fig. 2.1. It is often convenient to express the luminosity distance in terms of the *distance modulus*. When  $D_L$  is given in megaparsecs, the distance modulus is defined as

$$\mu \equiv m - M = 5 \log_{10} \left( \frac{D_L}{\text{Mpc}} \right) + 25, \quad (2.29)$$

where  $m$  and  $M$  denote the apparent and absolute magnitudes of the source, respectively. Thus, the luminosity distance can be written as

$$D_L = 10^{\frac{m-M-25}{5}} \text{ Mpc}. \quad (2.30)$$

Combining Eq. (2.28) with the definition of the distance modulus, the transverse comoving distance can be written as

$$D_M = \frac{10^{\frac{m-M-25}{5}}}{(1 + z)} \text{ Mpc}. \quad (2.31)$$

This expression will play an important role in the subsequent analysis.

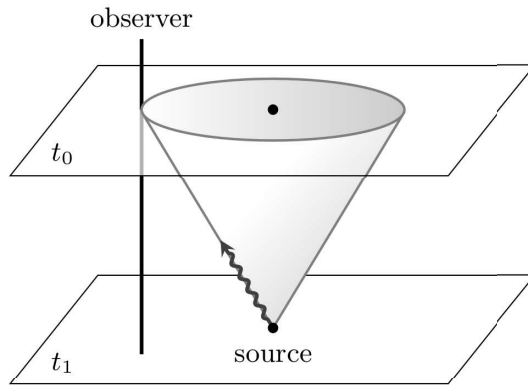


Figure 2.1: Spacetime diagram illustrating the concept of luminosity distance. Light emitted by a source spreads over a spherical surface, while cosmic expansion reduces both the photon energy and arrival rate (BAUMANN, 2022).

Recalling the relation  $D_L = (1 + z) D_M$  given in Eq. (2.28), and assuming a spatially flat FLRW Universe ( $\Omega_k = 0$ ), the transverse comoving distance reduces to the radial



comoving distance,  $D_M(z) = \chi(z)$ . In this case, the luminosity distance can be written explicitly in integral form as

$$D_L(z) = (1+z) \int_0^z \frac{c \, dz'}{H(z')}. \quad (2.32)$$

This integral expression for the luminosity distance will be used throughout the subsequent chapters of this thesis.

## 2.3 Cosmological Observational Probes

In this section, we introduce the primary cosmological observables considered in this thesis. These probes access distinct physical aspects of the Universe across a wide range of redshifts, providing complementary insights into its properties and evolution. We focus on the theoretical description of the Cosmic Microwave Background, baryon acoustic oscillations, Type Ia supernovae, and galaxy age and lookback time.

### 2.3.1 Cosmic Microwave Background

The Cosmic Microwave Background is a nearly perfect blackbody radiation (see Fig. 2.2) field that permeates the Universe, originating from the epoch when photons decoupled from matter in the early Universe. This decoupling occurred approximately  $3.8 \times 10^5$  years after the Big Bang, when the expansion and cooling of the Universe reduced the interaction rate between photons and free electrons to negligible levels.

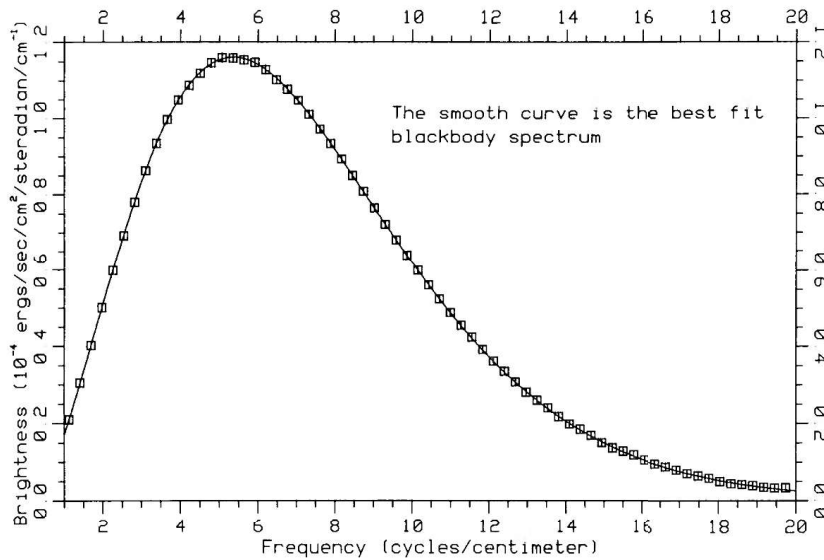


Figure 2.2: Preliminary CMB spectrum measured by the FIRAS instrument, compared with a blackbody spectrum. The boxes represent the observed data points and indicate the size of the 1% error bars (SMOOT *et al.*, 1992).

Within the standard cosmological framework, at temperatures  $T \gtrsim 1 \text{ eV}$ , photons were tightly coupled to the baryonic plasma through Compton scattering with free electrons. These electrons, in turn, were strongly coupled to protons via Coulomb interactions, forming a single photon–baryon fluid. As the temperature decreased and neutral hydrogen formed during recombination (see Fig. 2.3), the mean free path of photons increased dramatically, allowing radiation to propagate freely. The photons observed today as the CMB therefore provide a snapshot of the Universe at the time of last scattering.

Before photon decoupling, the tightly coupled photon–baryon fluid supported pressure-driven acoustic oscillations generated by primordial density perturbations. These oscillations propagated through the plasma until the epoch at which baryons were released from the Compton drag exerted by photons, commonly referred to as the baryon drag epoch. The maximum distance that these sound waves could travel before this epoch defines the comoving sound horizon, which sets a characteristic physical scale in the early Universe.

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)} dz, \quad (2.33)$$

where  $c_s(z)$  is the sound speed in the photon–baryon fluid, and  $z_d$  denotes the redshift of the baryon drag epoch, which occurs at  $z_d \approx 1059$  in the standard cosmological model. The sound speed of the photon–baryon fluid is given by

$$c_s(z) = \frac{c}{\sqrt{3(1 + R(z))}}, \quad (2.34)$$

where

$$R(z) = \frac{3\rho_b(z)}{4\rho_\gamma(z)}, \quad (2.35)$$

with  $\rho_b$  and  $\rho_\gamma$  representing the baryon and photon energy densities, respectively.

The sound horizon constitutes a fundamental standard ruler in cosmology. It is imprinted in the temperature and polarization anisotropies of the CMB and also leaves a characteristic feature in the large-scale distribution of matter at late times, known as baryon acoustic oscillations. Measurements of this scale, therefore, provide a powerful probe of the expansion history of the Universe and play a central role in modern observational cosmology.

In the early Universe, repeated scattering between photons and free electrons kept the radiation tightly coupled to the hot, dense plasma, ensuring thermal equilibrium. Under these conditions, CMB photons with frequencies in the interval  $\nu$  to  $\nu + d\nu$ , at temperature  $T$ , follow a blackbody distribution whose number density  $\mathcal{N}_T$  is given by

$$\mathcal{N}_T(\nu)d\nu = \frac{8\pi\nu^2/c^3}{(\exp[2\pi\hbar\nu/k_B T] - 1)} d\nu \quad (2.36)$$

In this expression,  $\hbar$  denotes the reduced Planck constant and  $k_B$  is the Boltzmann con-

stant. The frequency distribution of the radiation follows a blackbody spectrum, which is the characteristic signature of a system in thermal equilibrium. After photon decoupling, the radiation propagates freely through the expanding Universe. As the expansion proceeds, photon frequencies redshift according to  $\nu(t) \propto a(t)^{-1}$ , while the number density decreases as  $a(t)^{-3}$ . Nevertheless, the spectral shape remains that of a blackbody provided that the temperature scales as  $T \propto a(t)^{-1}$ . In this sense, the relic radiation preserves the imprint of the early thermal equilibrium phase of the hot Big Bang.

The Far InfraRed Absolute Spectrophotometer (FIRAS) experiment, launched in 1989 aboard the Cosmic Background Explorer (COBE) satellite (BOGGESS *et al.*, 1992), provided the first high-precision measurement of the CMB spectrum. The results showed that the present-day temperature of the radiation is  $T_0 = 2.725 \pm 0.001$  K (SMOOT, 1999). After establishing the monopole temperature of the CMB, it is important to note that the observed temperature depends on the relative motion of the observer due to the Doppler effect. For an isotropic blackbody radiation field, the temperature measured by an observer moving with velocity  $v$  relative to the CMB rest frame is given by

$$T_{\text{obs}} = T \frac{\sqrt{1 - \beta^2}}{1 - \beta \cos \theta}, \quad (2.37)$$

where  $\beta \equiv v/c$ , and  $\theta$  denotes the angle between the direction of observation and the observer's velocity vector. Expanding this expression for  $\beta \ll 1$  and using the identity  $2 \cos^2 \theta = 1 + \cos 2\theta$ , one obtains

$$T_{\text{obs}} = T_0 \left( 1 + \beta \cos \theta + \frac{\beta^2 \cos 2\theta}{2} + \dots \right). \quad (2.38)$$

The leading term in this expansion corresponds to the dipole anisotropy, which arises from the motion of the Earth relative to the CMB rest frame. This dipole was first detected by SMOOT *et al.* (1977). The corresponding temperature variation is  $\Delta T/T \sim 10^{-3}$ , implying a relative velocity of approximately  $390 \pm 60$  km/s.

After subtracting the kinematic dipole contribution, the CMB still exhibits small temperature fluctuations across the sky, known as anisotropies. These anisotropies were first detected by the COBE satellite. The measured quadrupole amplitude is  $13 \pm 4$   $\mu$ K (SMOOT ET AL., 1992), corresponding to  $\Delta T/T \approx 5 \times 10^{-6}$ , more than two orders of magnitude smaller than the dipole component.

To characterize the statistical properties of the temperature anisotropies, it is useful to describe the CMB temperature field across the sky. Let the temperature measured in a given direction  $\hat{\mathbf{n}}$  be written as

$$T(\hat{\mathbf{n}}) = \bar{T}_0 [1 + \Theta(\hat{\mathbf{n}})], \quad (2.39)$$

where  $\bar{T}_0$  denotes the mean CMB temperature today and  $\Theta(\hat{\mathbf{n}}) \equiv \delta T(\hat{\mathbf{n}})/\bar{T}_0$  represents the

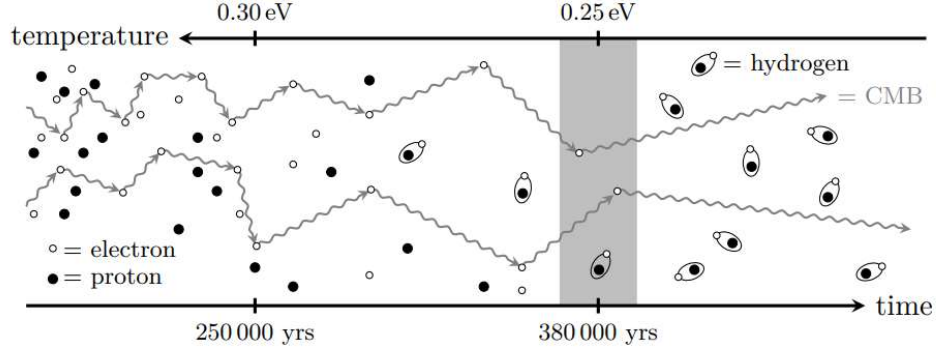


Figure 2.3: Cartoon of the recombination of protons and electrons into neutral hydrogen atoms and the corresponding decoupling of photons. Reproduced from BAUMANN (2022).

fractional temperature fluctuation in that direction.

A convenient way to quantify these fluctuations is through the two-point correlation function. By comparing the temperature fluctuations at two different directions  $\hat{\mathbf{n}}$  and  $\hat{\mathbf{n}}'$  on the sky, one defines

$$C(\theta) \equiv \langle \Theta(\hat{\mathbf{n}}) \Theta(\hat{\mathbf{n}}') \rangle, \quad (2.40)$$

where  $\cos \theta \equiv \hat{\mathbf{n}} \cdot \hat{\mathbf{n}}'$ , so that  $\theta$  represents the angular separation between the two directions on the sky. The angle brackets denote an average over an ensemble of possible universes. Although we observe only a single realization of the Universe, this ensemble average can be estimated by averaging correlations over different regions of the observed sky. However, on large angular scales, the number of statistically independent regions is small, which introduces an intrinsic statistical uncertainty known as cosmic variance. Expanding the temperature field in spherical harmonics is convenient since we see variations on the spherical surface of last-scattering

$$\Theta(\hat{\mathbf{n}}) = \sum_{l=-2}^{\infty} \sum_{m=-l}^l a_{lm} Y_{lm}(\hat{\mathbf{n}}) \quad (2.41)$$

where the expansion coefficients  $a_{lm}$  are called multipole moments. For a statistically isotropic temperature fluctuation field, the ensemble average of the two-point correlation function is fully determined by the angular power spectrum  $C_l$  through the relation

$$\langle a_{lm} a_{l'm'}^* \rangle = C_l \delta_{ll'} \delta_{mm'}. \quad (2.42)$$

Kronecker deltas are a consequence of statistical isotropy. The two-point correlation function in real space is equivalent to the angular power spectrum in harmonic space. In

fact, when we replace Eq. (2.40) with (2.41), we obtain

$$C(\theta) = \sum_l \frac{2l+1}{4\pi} C_l P_l(\cos \theta), \quad (2.43)$$

where  $P_l(\cos \theta)$  are the Legendre polynomials. The orthogonality of  $P_l(\cos \theta)$  allows us to write

$$C_l = 2\pi \int_{-1}^1 d\cos \theta C(\theta) P_l(\cos \theta) \quad (2.44)$$

The information contained in the angular power spectrum  $C_\ell$  is therefore fully equivalent to that described by the correlation function  $C(\theta)$ . In particular, the variance of the temperature anisotropy field can be written in terms of the multipole coefficients as

$$C(0) = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell \approx \int d\ln \ell \frac{\ell(\ell+1)C_\ell}{2\pi}. \quad (2.45)$$

This expression shows that the quantity  $\ell(\ell+1)C_\ell/2\pi$  represents the contribution to the variance per logarithmic interval in multipole  $\ell$ . It is therefore convenient to define the dimensionless quantity

$$\Delta_T^2 \equiv \frac{\ell(\ell+1)}{2\pi} C_\ell \bar{T}_0^2. \quad (2.46)$$

The CMB power spectrum is often plotted in terms of  $\Delta_T^2$ . For scale-invariant primordial fluctuations, this quantity is approximately independent of  $\ell$ . Note that  $\Delta_T^2$  corresponds to the quantity  $\mathcal{D}_\ell$  commonly used in the Planck collaboration papers. An example of the measured CMB temperature power spectrum is shown in Fig. 2.4. The series of acoustic peaks visible in the spectrum arises from oscillations in the photon–baryon fluid before recombination. The position and amplitude of these peaks encode important information about the physical conditions of the early Universe, including the matter content, baryon density, and spatial geometry.

### 2.3.2 Baryon Acoustic Oscillations

Baryon acoustic oscillations arise from sound waves propagating in the photon–baryon plasma in the early Universe, as discussed in the previous section (see, also, BASSETT and HLOZEK (2009); WEINBERG *et al.* (2013) for reviews). Although their physical origin lies in pre-recombination physics, their imprints remain observable at late times in the large-scale distribution of matter.

The clustering of matter acquires a characteristic scale in the presence of baryon acoustic oscillations. This feature can be identified either as a series of oscillatory patterns in the matter power spectrum,  $P(k)$ , which describes clustering in Fourier space, or as a distinct peak in the two-point correlation function,  $\xi(r)$ , defined in configuration space, as

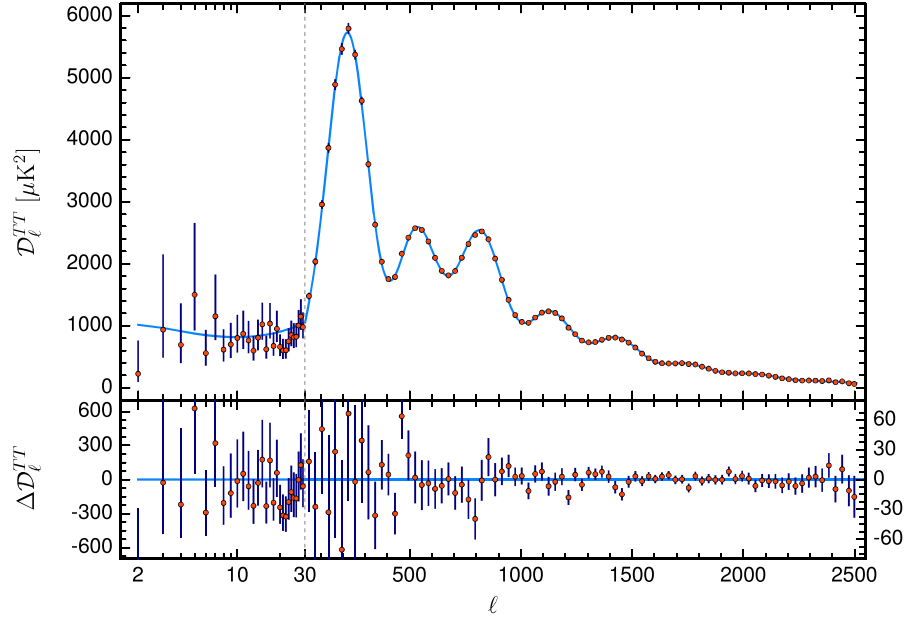


Figure 2.4: Angular power spectrum of the CMB temperature anisotropies from the PLANCK COLLABORATION (2020). The red points denote the measured temperature fluctuations, while the solid blue line corresponds to the best-fitting  $\Lambda$ CDM model. The lower panel illustrates the residuals between the observational data and the theoretical prediction.

first detected in galaxy surveys by EISENSTEIN *et al.* (2005) using luminous red galaxies (LRGs) from the SDSS, as illustrated in Fig. 2.5.

The acoustic scale is relatively large, of order  $\sim 150$  Mpc (or  $\sim 105 h^{-1}$  Mpc), reflecting the fact that primordial sound waves propagate at relativistic speeds. In the early Universe, when the radiation density dominates over baryons, the sound speed approaches  $c/\sqrt{3}$ , allowing these waves to travel substantial distances before recombination.

This large characteristic scale makes the BAO feature particularly robust against the effects of non-linear structure formation at low redshift, preserving its usefulness as a cosmological probe.

The main challenge of the BAO technique is statistical in nature. Since the signal is relatively weak and manifests at large scales, it requires mapping vast volumes of the Universe to be robustly detected and to yield precise distance measurements. Galaxy redshift surveys enable the construction of such three-dimensional maps of the large-scale structure.

Galaxy positions can be reconstructed in three dimensions from their angular coordinates on the sky and their redshifts. This enables independent measurements of the BAO scale in directions perpendicular and parallel to the line of sight, which probe distinct combinations of cosmological distances.

In practice, these measurements are expressed relative to the sound horizon at the

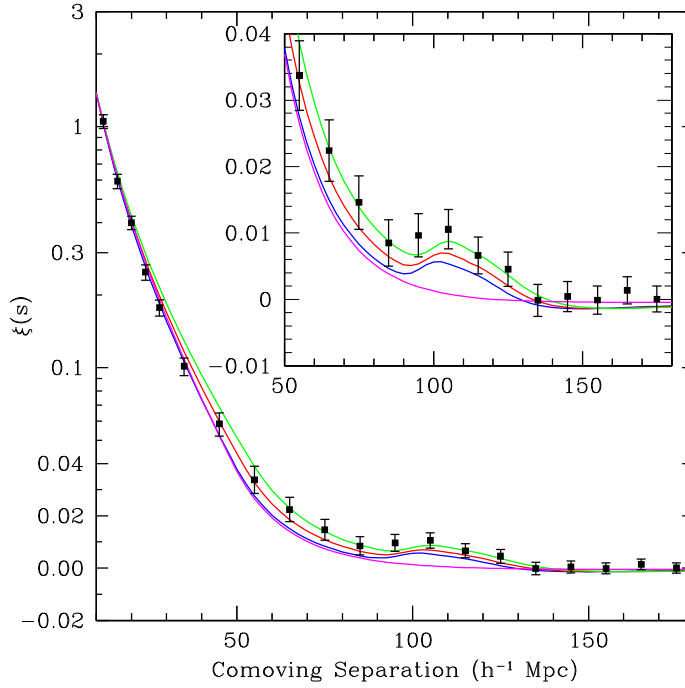


Figure 2.5: Large-scale two-point correlation function  $\xi(s)$  obtained from the SDSS luminous red galaxy sample. A clear peak is observed at  $\sim 100 h^{-1} \text{ Mpc}$ , providing one of the first detections of the baryon acoustic oscillation signal in galaxy surveys. Figure reproduced from EISENSTEIN *et al.* (2005).

drag epoch,  $r_d$ , defined in Eq. (2.33). The transverse component constrains the ratio  $D_M(z)/r_d$ , where  $D_M(z)$  is given in Eq. (2.26), through the observed angular separation  $\Delta\theta$ . The radial component is sensitive to the expansion rate, constraining the combination  $D_H(z)/r_d$ , where  $D_H(z)$  is defined in Eq. (2.17), as inferred from the redshift separation  $\Delta z$ .

Rather than separating the transverse and radial components, BAO measurements are often reported in terms of a single, spherically averaged distance scale that combines both contributions. This effective distance is defined as

$$D_V(z) = [z D_M^2(z) D_H(z)]^{1/3}, \quad (2.47)$$

which encapsulates the information from both the angular diameter distance and the Hubble expansion rate. In practice, observational constraints are typically expressed in terms of the ratio  $D_V(z)/r_d$ .

BAO thus provides a standard ruler to probe the expansion history of the Universe across different redshifts. These measurements can be used to infer cosmological distances and related quantities by thereby placing indirect constraints on the underlying cosmological model.

Because they offer a robust and largely model-independent probe of the background



expansion, BAO measurements play a central role in modern cosmology. We revisit this observable in the context of recent DESI survey data in Chapter 5, where these measurements are employed in the reconstruction of cosmological quantities.

### 2.3.3 Type Ia Supernovae

Supernovae (SNe) are events that occur during the last stages of star evolution. They are distinguished by an abrupt increase in luminosity, usually reaching peak luminosities of order  $10^9$ – $10^{10} L_{\odot}$ , which represent a significant portion of a galaxy’s overall luminosity. Different progenitor systems and explosion mechanisms are reflected in the various observational classes of supernovae. However, we are limited to Type Ia supernovae in this thesis.

A widely accepted interpretation of Type Ia supernovae (SNe Ia) is that they arise from thermonuclear explosions of white dwarfs in binary systems<sup>1</sup>. In the canonical progenitor scenario, the white dwarf accretes matter from a companion star until its mass approaches the Chandrasekhar limit,  $M_{\text{Ch}} \simeq 1.4 M_{\odot}$ , corresponding to the maximum mass supported by electron degeneracy pressure. Once this limit is exceeded, the star becomes gravitationally unstable, leading to a rapid increase in temperature and the ignition of thermonuclear reactions that disrupt the white dwarf.

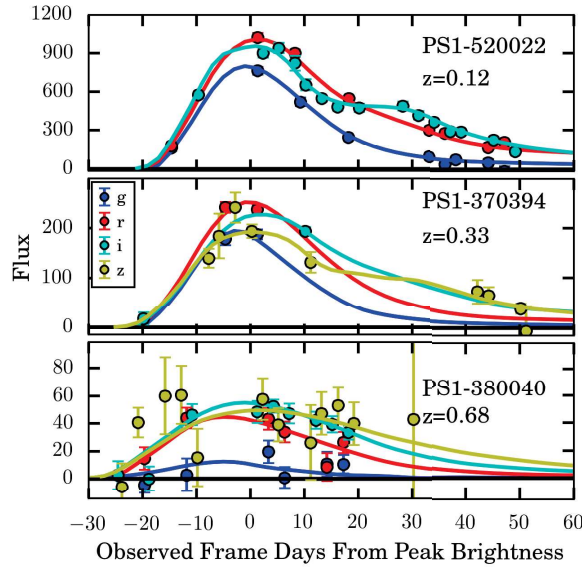


Figure 2.6: Representative multi-band light curves of Type Ia supernovae at different redshifts from the PS1 survey. The data points correspond to observed flux measurements, while the solid curves show best-fit models obtained using the SALT2 light-curve fitter. These fits allow for the standardization of SNe Ia luminosities, making them reliable distance indicators. Figure reproduced from SCOLNIC *et al.* (2018).

<sup>1</sup>An alternative progenitor channel is the double-degenerate scenario, in which the binary system consists of two white dwarfs whose merger may lead to a thermonuclear explosion.



The peak luminosities of SNe Ia show empirical relationships with the form and color of the light-curve, even though they are not inherently perfect standard candles. Specifically, the Phillips relation enables standardization of these occurrences by establishing a correlation between peak luminance and the rate of decrease in the light curve. SNe Ia behave as standardizable candles once these empirical corrections are applied, allowing cosmological distance measurements. Fig. 2.6 displays an example of observed multi-band light curves and their accompanying fits.

As an illustrative example, the Pantheon+ compilation (SCOLNIC *et al.*, 2022) standardizes Type Ia supernova light curves using the SALT2 model (GUY, J. *et al.*, 2007). In this approach, the observed light curves are fitted to determine the color parameter  $c$ , the stretch parameter  $x_1$ , and the overall amplitude  $x_0$ . The rest-frame  $B$ -band peak magnitude is then defined as

$$m_B \equiv -2.5 \log_{10}(x_0). \quad (2.48)$$

These parameters are used to correct the observed magnitude for empirical correlations between luminosity, light-curve width, and color. The distance modulus is then obtained through a modified TRIPP (1998) relation

$$\mu = m_B + \alpha x_1 - \beta c - M - \delta_{\mu-\text{bias}}, \quad (2.49)$$

where  $\alpha$  and  $\beta$  are correlation coefficients,  $M$  is the fiducial absolute magnitude of a standardized SN Ia, and  $\delta_{\mu-\text{bias}}$  represents bias corrections derived from simulations to account for selection effects and other observational systematics.

Modern Type Ia supernova compilations differ not only in sample size but also in the methodology adopted to process the data. For instance, the *Union3* compilation employs the SALT3 light-curve fitter, whereas *Pantheon+* is primarily based on SALT2. In addition, the DESY5 sample incorporates its own photometric calibration and survey-specific modeling of selection effects and intrinsic scatter. Consequently, even when the same supernova appears in more than one catalog, the derived parameters  $(m_B, x_1, c)$ , the global nuisance coefficients  $(\alpha, \beta, M)$ , and the associated covariance matrices are not identical.

Formally, each compilation defines a distinct statistical mapping between the raw photometric observations and the distance moduli used in cosmological inference. Therefore, different catalogs should not be regarded merely as quantitative extensions of one another, but rather as independent statistical reconstructions of the same underlying observational data.

### 2.3.4 Age and lookback time

In the early 1990s, cosmology faced a significant “age problem”: estimates of the Hubble constant around  $H_0 \sim 80 \text{ km s}^{-1} \text{ Mpc}^{-1}$  implied a Universe younger than 10 Gyr, in conflict with the ages of the oldest stellar populations, which exceeded 12 Gyr (CHABOYER, 1998; JIMENEZ, 1997; KRAUSS and CHABOYER, 2003; KRAUSS and TURNER, 1995; OSTRICKER and STEINHARDT, 1995). This discrepancy challenged the standard cosmological model and was later alleviated by the introduction of a cosmological constant,  $\Lambda$ , leading to the success of the  $\Lambda$ CDM model (ALCANIZ and LIMA, 1999; ALCANIZ *et al.*, 2003; KRAUSS and TURNER, 1995; LIMA and ALCANIZ, 2000).

This tension highlighted the importance of accurately determining the ages of cosmic objects for constraining cosmology. In this context, galaxies can be used as cosmic chronometers (JIMENEZ and LOEB, 2002), providing an alternative probe of the expansion history. By comparing galaxy ages at different redshifts (NOLAN *et al.*, 2001; SIMON *et al.*, 2005), one can place constraints on key cosmological parameters, such as the Hubble constant (VAGNOZZI *et al.*, 2022; WEI and MELIA, 2022) and the dark energy equation of state (DANTAS *et al.*, 2007b, 2009, 2011).

To make this connection explicit, we relate the redshift to the cosmic age starting from

$$\int_0^{t_z} dt = \int_0^{a'} \frac{da}{a}. \quad (2.50)$$

Using the relation between the redshift and the scale factor given in Eq. (2.8) to perform a change of variables, we obtain

$$t(z) = H_0^{-1} \int_z^\infty \frac{dz'}{(1+z')E(z')}, \quad (2.51)$$

Complementarily, we define the relationship between the lookback time and the redshift as

$$t_L(z) = H_0^{-1} \int_0^z \frac{dz'}{(1+z')E(z')}, \quad (2.52)$$

Following CAPOZZIELLO *et al.* (2004), we consider an astrophysical object (e.g., a galaxy, a quasar) observed at redshift  $z_i$ . The age of the object, denoted by  $t(z_i)$ , is defined as the difference between the age of the Universe

$$t(z_i) = H_0^{-1} \left[ \int_{z_i}^\infty \frac{dz'}{(1+z')E(z')} - \int_{z_f}^\infty \frac{dz'}{(1+z')E(z')} \right] \quad (2.53)$$

or, equivalently,

$$t(z_i) = t_L(z_f) - t_L(z_i). \quad (2.54)$$

The observed lookback time to an object at  $z_i$  may now be easily defined using the

formulas above as

$$\begin{aligned}
 t_L^{obs}(z_i) &= t_L(z_f) - t(z_i) \\
 &= [t_0 - t(z_i)] - [t_0 - t_L(z_f)] \\
 &= t_0 - t(z_i) - t_f,
 \end{aligned} \tag{2.55}$$

where  $t_0$  is the age of the universe at  $z = 0$ . In the literature, related terms such as formation epoch, delay factor, or incubation time are often used to refer to the formation time  $t_f$ , depending on the specific modeling assumptions or observational context. These relations and definitions will play a central role in [Chapter 6](#).

## CHAPTER 3

---

### Supervised learning in Cosmology

---

#### 3.1 Introduction

Over the past decade, the number of studies in cosmology involving ML has increased significantly, driving the development of new techniques and applications for analyzing astronomical data. These advances have been crucial for handling the large volume and complexity of data from modern observational surveys, such as DES, DESI, KiDS, and LSST, which produce extensive and multidimensional catalogs.

Among the most common applications of ML in astronomy are fast finding of strongly lensed systems (LANUSSE *et al.*, 2018), measuring parameters of early galaxies using the 21-cm signal (GILLET *et al.*, 2019), the search for transient events (ZHANG *et al.*, 2020), and the reconstruction of CMB lensing maps (CALDEIRA *et al.*, 2019), Primordial Black Holes (YUAN *et al.*, 2025), hydrodynamical simulations for cosmology and galaxy cluster physics (SCHAYE *et al.*, 2023), emulate  $N$ -body simulations or accelerate the generation of mock catalogs (HE *et al.*, 2019; VILLAESCUSA-NAVARRO *et al.*, 2022).

Several of the studies described here involve convolutional neural networks (CNN), which are part of the broader class of Deep Learning techniques, multi-layer neural architectures meant to learn hierarchical feature representations directly from input. As an illustrative case, JACOBS *et al.* (2019) deployed a deep residual CNN on DES imaging data to discover high-redshift strong-lens candidates, achieving high purity and completeness. Fig. 3.1 provides exemplary examples of the systems detected by their CNN-based pipeline.

Machine learning has also been applied more broadly in cosmology, including investigating observational degeneracies between nine models of modified gravity and massive neutrinos (MERTEN *et al.*, 2019), using deep learning methods for measuring cosmological parameters from density fields (SCHMELZLE *et al.*, 2017), inferring cosmological parameters and reconstructing the cosmic distance ladder through deep-learning approaches (SHAH *et al.*, 2024), as well as estimating cosmological parameters from observational data using supervised regression techniques (BENGALY *et al.*, 2023; PAL and SAHA,

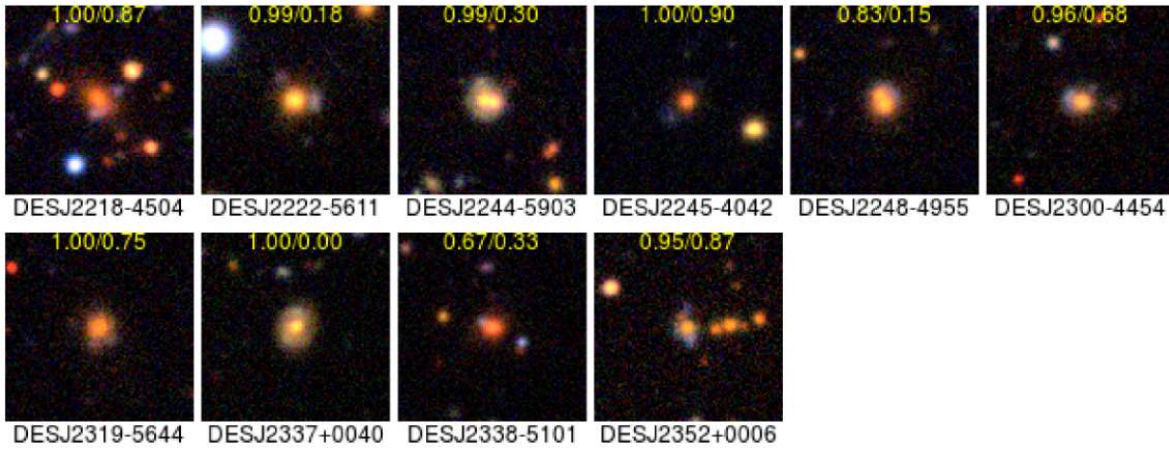


Figure 3.1: Examples of strong-lens candidates identified in DES imaging data using CNNs. The CNN scores are shown in yellow text. Image reproduced from Figure 12 of JACOBS *et al.* (2019).

2024; SOUSA-NETO and DANTAS, 2025).

There is also a class of machine learning methods known as Genetic Algorithms (GA), which have been applied in cosmology, for example, in Type Ia supernova analyses (BOGDANOS and NESSERIS, 2009), dark-energy modeling (NESSERIS and GARCÍA-BELLIDO, 2012), null tests of the cosmological constant (NESSERIS and SHAFIELOO, 2010), and constraints on the distance–duality relation using standard sirens (HOGG *et al.*, 2020). For a comprehensive overview of ML applications in cosmology, see the reviews by CARLEO *et al.* (2019); DVORKIN *et al.* (2022); HLOŽEK (2023); MORIWAKI *et al.* (2023).

Of particular relevance here are the studies that employ SR techniques, which have been used to reconstruct cosmological data (DESMOND, 2025) and to explore inflationary potentials (SOUSA *et al.*, 2024), to investigate bounded dark energy models (BORGHETTO *et al.*, 2025), and to constrain dark matter halo profiles (MARTÍN *et al.*, 2025). We will return to SR in more detail later.

As an illustrative example, BARTLETT *et al.* (2024b) used symbolic regression to recreate the Hubble expansion rate  $H(z)$  from Cosmic Chronometer data. The top SR-derived functions are color-coded according to their minimum description length<sup>1</sup> (MDL) value in Figure 3.2. The ensemble of solutions demonstrates the range of functional forms compatible with the data.

In this chapter, we will revisit supervised learning to see what it is and how it works, focusing on its applications in cosmology. In particular, we will explore supervised regression methods, with a special emphasis on SR, and how they can infer cosmological parameters from data, illustrating the methodology with recent examples from the literature.

<sup>1</sup>In ESR, model selection is performed using the MDL principle, which is based on information theory.

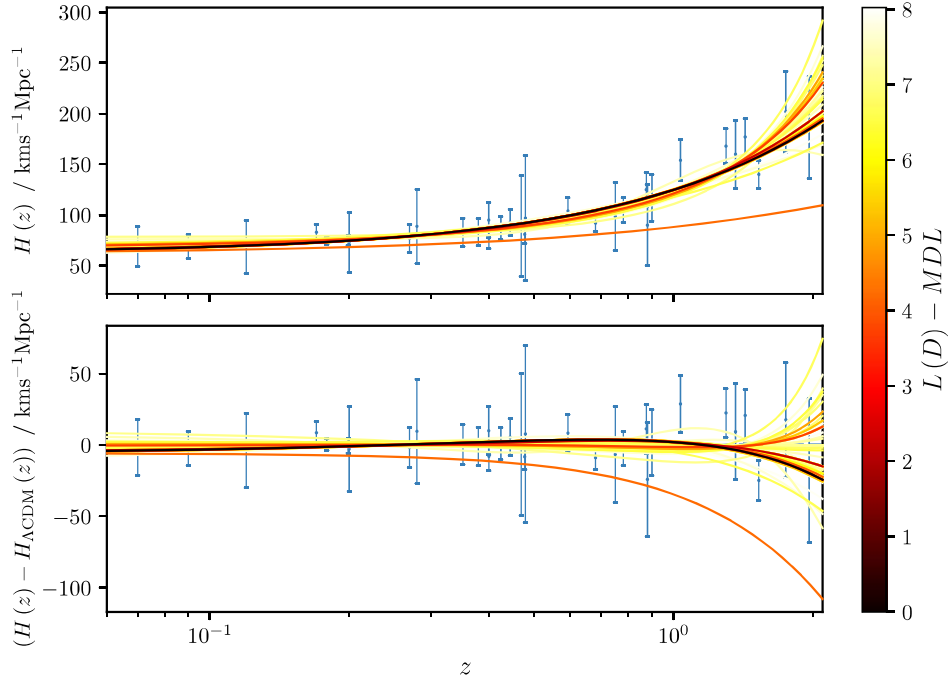


Figure 3.2: Top symbolic-regression reconstructions of  $H(z)$  from Cosmic Chronometer data, colored by their MDL score. Figure extracted from BARTLETT *et al.* (2024b).

## 3.2 Supervised Regression in Cosmology: Conceptual Overview

Machine learning is a subfield of artificial intelligence concerned with systems that learn from data without being explicitly programmed (BISHOP, 2006). It is traditionally divided into three main categories: supervised, unsupervised, and reinforcement learning (see Figure 3.3). We focus on supervised learning.

Given a training set  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ , where  $N$  is the number of training samples, its objective is to learn a mapping between inputs  $x_i$  and outputs  $y_i$ . The inputs  $x_i$  may be simple  $D$ -dimensional vectors—representing features such as color or size—or more complex structures such as images or time series. The outputs  $y_i$  can be categorical variables,  $y_i \in \{1, 2, \dots, \mathcal{C}\}$ , where  $\mathcal{C}$  is the number of classes (e.g., a stellar spectral type), or real-valued scalars (e.g., the age of a galaxy). The former defines a *classification* task, while the latter corresponds to *regression* (MURPHY, 2012).

In both cases, we assume an underlying relationship  $y = f(x)$  for some unknown function  $f$ . Given the training data, we seek to estimate this function and use it to make predictions  $\hat{y} = \hat{f}(x)$ . The main objective is to generalize to unseen data (the test set), since predicting the training labels is trivial. Properly splitting the dataset into training and test subsets is therefore essential; a common choice in the literature is a 70/30 train–test split.



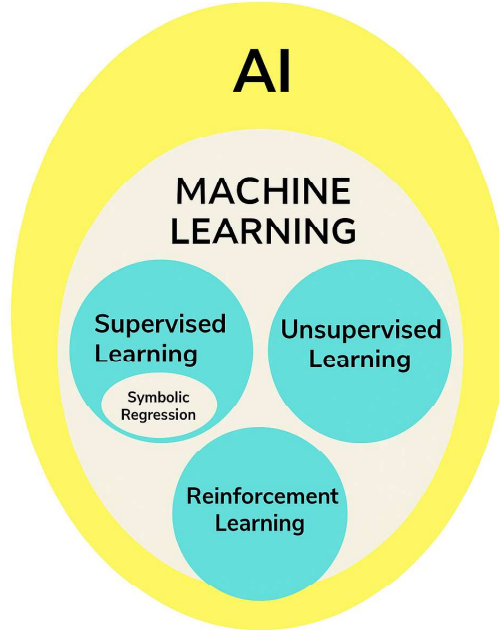


Figure 3.3: Hierarchical structure of AI showing ML as a subset, which includes supervised, unsupervised, and reinforcement Learning. Symbolic regression is highlighted as a specific technique within Supervised Learning.

Classic supervised learning tools, such as Support Vector Machines (SVM; CORTES and VAPNIK, 1995), Multi-Layer Perceptrons (MLP), and Decision Trees, have been widely applied in cosmology (BENGALY *et al.*, 2023; CAGLIARI *et al.*, 2024; CONCEIÇÃO *et al.*, 2024; LOCHNER *et al.*, 2016; SOUSA-NETO and DANTAS, 2025). These studies have addressed both regression and classification tasks; here, the focus is primarily on regression, aiming at predicting continuous cosmological parameters such as the Hubble constant or matter density.

To demonstrate the general procedure of a supervised learning regression task, we employ Support Vector Regression (SVR<sup>2</sup>), the regression counterpart of SVM, as a practical example. In this approach, the reconstruction is determined by a subset of the training data known as support vectors, which correspond to the data points that most strongly constrain the regression function. The algorithm seeks to reconstruct the underlying relation by maximizing the margin around the regression curve while allowing deviations smaller than a specified tolerance parameter. As our dataset, we use the  $H(z)$  measurements obtained from Cosmic Chronometers.

This type of reconstruction using machine learning techniques is already well estab-

---

<sup>2</sup>For further details, see <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>.

lished in the literature (BARTLETT *et al.*, 2024a; BENGALY *et al.*, 2023; DIALEK-TOPOULOS *et al.*, 2022; WANG *et al.*, 2020). The inputs for the model consist of the redshift  $z$ , the corresponding  $H(z)$  values, and a weight for each data point, defined here as the inverse of the squared measurement error.

First, we divide the dataset into training and testing subsets, using 70% of the data for training and 30% for testing. Since the  $H(z)$  dataset from Cosmic Chronometers contains 34 data points, this corresponds to approximately 24 points for training and 10 for testing. This division allows the model to be trained on one subset while being independently evaluated on unseen data, ensuring an unbiased assessment of its predictive performance. Before training, the data were standardized to ensure consistent scaling, improving the performance of the SVR model. Another important step is the choice of hyperparameters, which are the intrinsic parameters that define the behavior of our model—in this case, the SVR.

To ensure the proper selection of these values, one can explore the entire hyperparameter space by training the model with all possible combinations within a predefined range. This process, commonly known as a **grid search**, allows us to identify the combination of hyperparameters that yields the best performance for our dataset.

Another important aspect is to train the model comprehensively, that is, to split the data into training and testing sets in multiple ways (**folds**), thereby reducing the risk of bias.

In order to guarantee that the results not only replicate the average behavior of the data but also represent the statistical confidence associated with the predictions, it is crucial to talk about how uncertainties can be measured and how the reconstruction process can account for the errors in the original data.

Using weights in the model training process is a good way to take data mistakes into account. This method allows the model to learn more robustly since data points with large related uncertainties contribute less than those with smaller uncertainty.

In our example, the weights were defined as  $w = 1/\sigma^2$ , where  $\sigma$  represents the uncertainty associated with each data point. Consequently, observations with larger uncertainties have less influence during model training, whereas more reliable points contribute more to the final reconstruction. Furthermore, setting a random seed is a crucial step to ensure the reproducibility of the results.

Since machine learning methods do not automatically produce error estimates, this is, in fact, one of the main critiques leveled at them.

Several metrics can be employed to assess model performance, such as the Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination ( $R^2$ ), Root Mean Squared Error (RMSE), and the ratio between RMSE and the mean of the observed values ( $\text{RMSE}/\bar{y}$ ).



Continuing with our example, the following results were obtained:

$$\text{MSE} = 252.054, \quad \text{RMSE} = 15.876, \quad \text{MAE} = 10.921, \quad R^2 = 0.828, \quad \frac{\text{RMSE}}{\bar{y}} = 0.147.$$

These results indicate that the model exhibits good predictive performance, with relatively low average error and a high coefficient of determination, suggesting that it is able to explain about 83% of the variance in the observed data. However, it is important to maintain a balance between model accuracy and generalization. Extremely high accuracy values may indicate that the model is overfitting the training data, capturing noise rather than the underlying physical behavior of  $H(z)$ .

In this context, the relatively high value of the MSE should not be interpreted as poor performance. Since the reconstruction of the Hubble parameter involves values on the order of hundreds of km/s/Mpc, a large MSE is expected. Therefore, it is more meaningful to evaluate the model using relative quantities, such as the ratio  $\text{RMSE}/\bar{y}$ , which in this case is approximately 0.15, in good agreement with previous reconstructions of  $H(z)$  using similar techniques (BENGALY *et al.*, 2023; DIALEKTOPOULOS *et al.*, 2022).

Figure 3.4 shows the reconstruction of  $H(z)$  obtained with the SVR. Blue points correspond to the observational data from CC, displayed with their respective uncertainties. The black solid line represents the SVR prediction for  $H(z)$ , while the shaded regions correspond to the  $1\sigma$  (gray) and  $2\sigma$  (orange) confidence intervals, estimated using the RMSE.

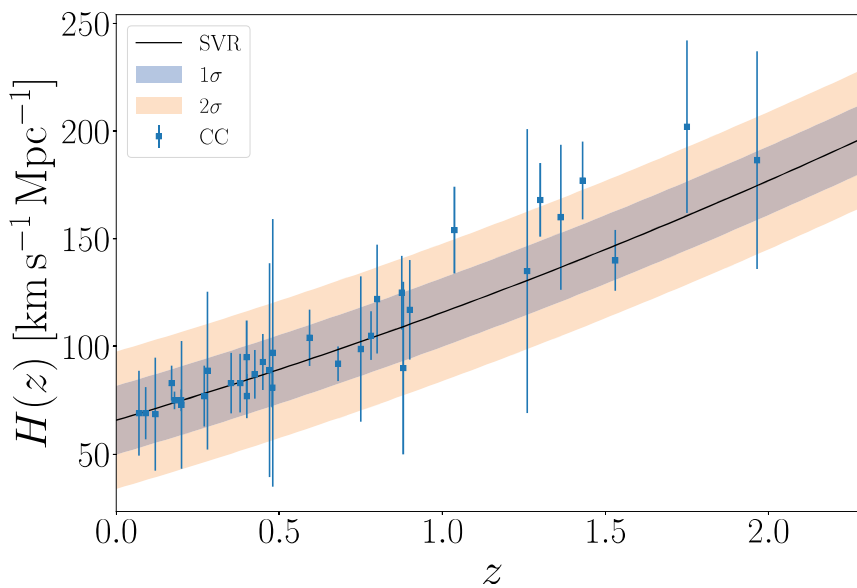


Figure 3.4: Reconstruction of the Hubble parameter  $H(z)$ .

By using the example of reconstructing  $H(z)$ , it was possible to illustrate the general

procedure for data reconstruction using supervised learning techniques. It is essential to note that, although the SVR approach was employed in this example, the same methodology applies to other supervised learning algorithms. The process involves fundamental steps such as dividing the dataset into training and testing samples, normalizing the variables, searching for the optimal hyperparameters, performing cross-validation, selecting and computing performance metrics (such as MSE, RMSE, MAE,  $R^2$ , and  $\text{RMSE}/\bar{y}$ ), and using weights to incorporate observational uncertainties. The main difference among the techniques lies in the specific hyperparameters of each model and in how they influence the final reconstruction.

### 3.3 Symbolic Regression

Symbolic Regression (SR) is a supervised learning task in which the model space consists of analytic expressions (KOZA, 1994; SCHMIDT and LIPSON, 2009). This idea is actually intuitive from the perspective of the history of science: long before modern machine learning, scientists effectively performed a form of “manual” symbolic regression when inferring functional laws directly from empirical data. A classic example is the discovery of the linear relation between recessional velocity and distance — today known as the Hubble-Leaître law — which emerged from searching for the analytic expression that best described the observational trend.

This search for analytic expressions is commonly approached through a multi-objective optimization framework, where the goal is to minimize prediction error and model complexity simultaneously. Here, complexity plays a central role in SR, and we adopt the definition provided by CRANMER (2023b). By default, complexity is equal to the number of nodes in an expression tree, regardless of each node’s content. This means that two expression trees containing, for example, the operators  $[+, -, *]$  and  $[\exp, \cos, \log]$  would both have complexity equal to 3, since only the number of nodes matters, not their specific content. Instead of adjusting specific parameters within an overparameterized general model, SR explores the space of simple analytic expressions to identify accurate and interpretable models.

Among the various SR libraries available in the literature, such as Eureqa (SCHMIDT and LIPSON, 2009), GPLEarn<sup>3</sup>, Exhaustive Symbolic Regression (ESR; BARTLETT *et al.*, 2022), OPERON, and AI Feynman (UDRESCU and TEGMARK, 2020), we will utilize PySR (CRANMER, 2023b). PySR is a powerful open-source library for symbolic regression, available in both Julia and Python. The Python interface is maintained at the official GitHub repository<sup>4</sup>, and extensive documentation with examples is provided

<sup>3</sup><https://github.com/trevorstephens/gplearn>

<sup>4</sup><https://github.com/MilesCranmer/PySR>

online<sup>5</sup>.

The algorithmic core of PySR combines evolutionary search to explore the space of candidate symbolic expressions with gradient-based optimization to refine their parameters, enabling the efficient discovery of interpretable models. A detailed description of the methodology and its theoretical foundations is presented in CRANMER (2023b).

In this framework, the evolutionary component can be summarized through the following steps (see Fig. 3.5):

1. Assume a population of individuals, a fitness function, and a set of mutation operators.
2. Randomly select a subset of size  $n_s$  from the population (e.g., classical tournament selection uses  $n_s = 2$ , although larger values are also possible).
3. Evaluate the fitness of each individual in the selected subset.
4. Select the fittest individual with probability  $p$ . Otherwise, remove it from the subset and repeat the selection procedure. If only one individual remains, select it. This leads to probabilities approximately given by  $p$ ,  $p(1 - p)$ ,  $p(1 - p)^2$ , and so on for increasingly less fit individuals.
5. Create a copy of the selected individual and apply a randomly chosen mutation from the available set of mutation operators.
6. Replace an individual in the population with the mutated one, typically the least fit individual.

This procedure is computationally efficient and naturally suited for large-scale parallelization. In particular, the population can be partitioned into independent subgroups that evolve separately, with occasional asynchronous migration of individuals between them.

As a modification to the standard scheme, instead of replacing the least fit individual after mutation, one may choose to replace the oldest member of the population. In what follows, we focus on the key practical aspects of PySR that are most relevant to its application in cosmological data analysis.

With PySR, various operators construct and evolve symbolic expressions. The operators are divided into unary and binary to explore possible solutions and identify the best-fitting models. These operators are used to construct the symbolic expressions that are evaluated and optimized to find the most accurate and interpretable models for the given data. The number of operators chosen has a direct impact on the computational time. Moreover, the chosen operator set constitutes a bias, as it shapes the functional

---

<sup>5</sup><https://astroautomata.com/PySR/v1.5.9/api.html>

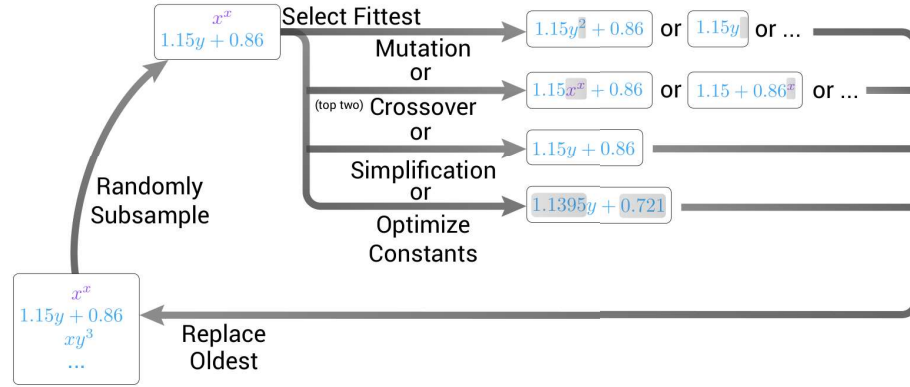


Figure 3.5: Schematic illustration of the inner loop of PySR. A subset of expressions is randomly sampled from the population, within which a tournament selection is performed. The selected individual is then used to generate new candidates through mutation, crossover, simplification, or parameter optimization. Image reproduced from CRANMER (2023a).

space that SR is allowed to explore. For this reason, it is often useful to train the model with different operator sets to assess how robustly each choice fits the data.

In this sense, the operator set constitutes a key hyperparameter<sup>6</sup>, as it shapes the hypothesis space of symbolic regression. An additional hyperparameter of central importance is the loss function, which quantifies how well a model reproduces the target behavior. During the training process, the objective is to minimize this loss. In PySR, this minimization is performed using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (BROYDEN, 1970), which determines the descent direction by iteratively updating an approximation of the inverse Hessian and refining the search toward a local minimum. Consequently, an important requirement is that the chosen loss function be differentiable.

For the selection of the final equation, PySR gives three distinct criteria—**accuracy**, **best**, and **score**—each presenting a different technique for balancing predictive performance and model complexity. The equation with the lowest loss out of all the contenders is chosen based on the accuracy requirement. The best criterion identifies models whose loss is below a preset threshold (usually 1.5 times the minimum loss) and then selects the one that achieves the most advantageous trade-off between loss and complexity. The scoring criterion is defined as the negated derivative of the log-loss with respect to complexity; an expression is picked if its loss is significantly better at a slightly greater complexity. According to the stated selection rule, the chosen expression consequently provides the most suitable empirical description of the data<sup>7</sup>.

<sup>6</sup>A hyperparameter is any parameter that governs the training process but is not directly learned from the data.

<sup>7</sup>Further details on these criteria can be found in the official PySR documentation: <https://ai.damtp.cam.ac.uk/pysr/api/>.

Although the resulting expressions generally exhibit strong statistical performance, they do not inherently guarantee physical interpretability, since the selection process is guided solely by data fitting and structural simplicity.

As discussed earlier, Hubble's law can be interpreted as a manual form of SR. Therefore, an interesting exercise is to test whether SR is capable of recovering Hubble's law itself — a test previously performed in CRANMER (2023a). For this purpose, we use the data presented in Table 1 of HUBBLE (1929) (see the data points shown in Fig. 3.6). Although these data do not correspond to the final set adopted in his analysis, they provide a suitable approximation for didactic purposes.

Our input variables for the model are the recession velocities of the galaxies and their corresponding distances. The next step is to choose the hyperparameters of the SR algorithm. In particular, we select the following operator set:

- **Binary operators:**  $\{+, -, \times, /\}$
- **Unary operators:**  $\{\text{square}, \text{cube}\}$

For the model selection procedure, we adopt the "best" criterion. To simplify our analysis, we fix the number of iterations to `niterations = 100` and the population size to `populations = 50`. Regarding the loss function, it is worth noting that PySR provides several built-in options for model fitting, including robust, smooth, scale-invariant, and personalized losses.<sup>8</sup>

In what follows, we briefly present three examples of loss functions available in PySR. Among these options, the **HuberLoss**, a widely used robust loss designed to mitigate the influence of outliers, will be adopted in our analysis.

The **Huber loss** is defined as

$$L(r) = \begin{cases} \frac{r^2}{2}, & \text{if } |r| \leq \alpha, \\ \alpha|r| - \frac{\alpha^2}{2}, & \text{otherwise,} \end{cases} \quad (3.1)$$

where  $r = y - \hat{y}$  is the residual and  $\alpha$  is a transition parameter controlling the boundary between the quadratic and linear regimes.

**L1 Epsilon-Insensitive Loss** ignores residuals smaller than  $\epsilon$ , focusing the fit on larger deviations.

$$L(r) = \max\{0, |r| - \epsilon\}; \quad (3.2)$$

**LP Distance Loss** generalizes L1 loss; the exponent  $P$  tunes the sensitivity to residuals and outliers.

$$L(r) = |r|^P \quad (3.3)$$

---

<sup>8</sup>A complete list of the available loss functions is provided in the official PySR documentation: <https://ai.damtp.cam.ac.uk/symbolicregression/v1.12.0/losses>.

This loss behaves as an  $L_2$  loss for small residuals ( $|r| \leq \alpha$ ), ensuring smooth optimization, while transitioning to an  $L_1$ -like penalty for large residuals, thus reducing sensitivity to outliers.

| Equation                                                              | Complexity | Loss        |
|-----------------------------------------------------------------------|------------|-------------|
| $v = 292.$                                                            | 1          | 295.        |
| <b><math>v = 425 \times D</math></b>                                  | <b>3</b>   | <b>178.</b> |
| $v = 443 \times D - 36.6$                                             | 5          | 177.        |
| $v = 420 \times D + \frac{5.02}{D}$                                   | 7          | 167.        |
| $v = 425 \times D + \frac{0.163}{D^2}$                                | 8          | 166.        |
| $v = 432 \times D - \frac{24.9}{D - 0.126}$                           | 9          | 142.        |
| $v = 432 \times D + \frac{-24.9 - 0.126}{D - 0.126}$                  | 11         | 142.        |
| $v = 424 \times D - \frac{17.2}{-D^6 + D - 0.133}$                    | 13         | 138.        |
| $v = \frac{D^3}{(D - 0.315)^3} + 422 \times D - \frac{5.01}{D}$       | 14         | 135.        |
| $v = 423 \times D + \frac{0.0295}{(D - 0.319)^3} + \frac{0.172}{D^2}$ | 15         | 132.        |
| $v = 424 \times D + \frac{0.0295}{(D - 0.319)^3} + \frac{0.161}{D^2}$ | 16         | 132.        |
| $v = 424 \times D - \frac{17.2}{D - (-D + 0.0167D)^6 - 0.133}$        | 17         | 124.        |

Table 3.1: Final model returned by PySR, where the **bold** expression corresponds to the best equation.

With the dataset specified and the hyperparameters properly defined, we can now perform the symbolic regression using PySR. The resulting expressions from this illustrative application to the HUBBLE (1929) data are summarized in Table 3.1. It is important to emphasize that the complexity reported in the table corresponds to the internal expression-tree complexity used by PySR during the search process, and not necessarily to the simplified equation shown in the table. This distinction is relevant because different tree representations may lead to expressions that look simpler (or more compact) once written explicitly, potentially confusing the actual model complexity evaluated by the algorithm. The equation highlighted in bold,

$$v = 425 \times D \quad (3.4)$$

is identified by PySR as the best equation under the `model_selection = "best"` criterion, with a complexity equal to 3. Moreover, the balance between loss and model complexity does not change sufficiently to favor any of the higher-complexity expressions returned

by the algorithm. This expression reproduces the functional form of the Hubble law,

$$v = H_0 \times D, \quad (3.5)$$

yielding an estimated value of  $H_0 = 425 \text{ km s}^{-1} \text{ Mpc}^{-1}$ . Furthermore, the blue curve shown in Fig. 3.6 corresponds precisely to this symbolic regression model. This result illustrates that, even when applied to the highly noisy historical measurements of 1929, the symbolic regression framework is capable of recovering the fundamental linear relation originally identified by Hubble.

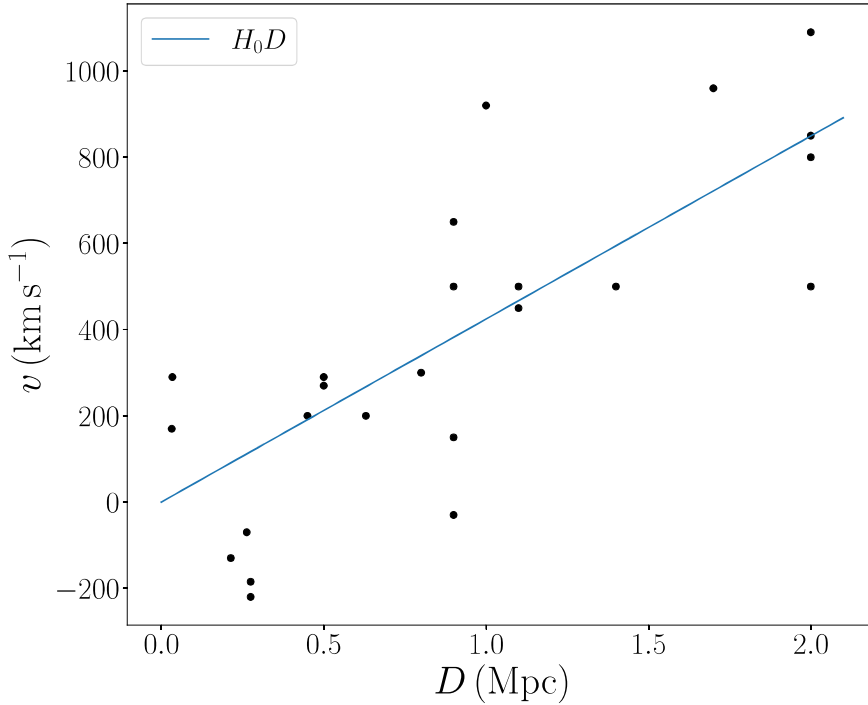


Figure 3.6: Scatter plot of the recession velocities of galaxies as a function of their distances. The data points were extracted from Table 1 of HUBBLE (1929), and the solid line represents the linear relation  $v = H_0 D$ .

Once again, this example makes it easy to understand the ML process step-by-step and pinpoint the hyperparameters that require special care during training. It is crucial to note that, in contrast to the previous SVR example, PySR manages these processes internally, so we don't divide the data into training and test sets or do a grid search for hyperparameter tuning. However, the same procedure as in the preceding example can still be employed if desired; in fact, we do so in our analysis, as covered in chapters 6 and 5. This additional step provides an extra degree of assurance that the chosen hyperparameter set is suitable for the data.



### 3.4 Challenges and Perspectives

In the previous sections, we reviewed a variety of machine-learning applications, particularly supervised learning and SR. A fundamental question that arises concerns how uncertainties are treated in such analyses. Traditionally, a wide range of metrics is available to assess the quality of a regression model and ensure that it accurately represents the data. However, quantifying the uncertainty associated with the predicted quantities is more subtle, especially when we aim to propagate observational errors into the model predictions.

A straightforward way to incorporate data uncertainties is through weighted loss functions, in which points with smaller errors contribute more heavily to the training process than those with larger uncertainties. Additionally, Monte Carlo sampling can be employed to propagate observational errors and estimate uncertainties in the reconstructed quantities.

There are also Bayesian statistical approaches (e.g., Gaussian Processes RASMUSSEN and WILLIAMS (2006); SEIKEL *et al.* (2012), Bayesian Neural Networks (ESCAMILLA-RIVERA *et al.*, 2020; NEAL, 1996; THUMMEL *et al.*, 2024)) designed to provide explicit uncertainty estimates. Nonetheless, these techniques are not always easily integrated with other supervised-learning methods, which may limit their applicability in certain contexts.

Another important aspect concerns the nature of cosmological datasets. While many machine-learning applications in other fields rely on large datasets, key cosmological probes—such as SNIa, CC, BAO, gravitational-wave standard sirens, and galaxy-age measurements—often consist of relatively small samples that nevertheless encode significant physical information. This limited data volume tends to increase reconstruction uncertainties. Even so, as demonstrated throughout this chapter and supported by numerous works in the literature, machine learning remains a promising and powerful tool for cosmological analysis and inference.

In the next chapter, we will introduce Gaussian Processes, a widely used non-parametric statistical method for function reconstruction and uncertainty estimation, which provides an important point of comparison with the supervised learning techniques discussed in this chapter.



# CHAPTER 4

---

## Gaussian Processes

---

### 4.1 Introduction

In the previous chapter, we presented data reconstruction using machine learning techniques such as SVR and PySR. In this chapter, we turn to another methodology widely used in cosmology for non-parametric reconstruction: Gaussian Processes (GP).

Gaussian Processes provide a unifying probabilistic framework that connects several well-established modeling approaches. Under suitable conditions, they can be shown to be mathematically related to Bayesian linear models, and even certain classes of neural networks. They also share conceptual similarities with support vector machines. From the GP perspective, these models can often be interpreted more transparently and analytically tractable.

The presentation will primarily follow the approach described in RASMUSSEN and WILLIAMS (2006). Although GP are frequently employed within the broader machine learning literature — for instance, through implementations available in libraries such as `scikit-learn` — they are more fundamentally understood as Bayesian non-parametric models.

In this thesis, we use the implementation provided by `GaPP` (Gaussian Processes in Python) (SEIKEL *et al.*, 2012), a tool specifically developed for cosmological applications, to directly reconstruct a function and its derivatives from observational data.

### 4.2 Reconstructing a function

A Gaussian Process is defined as a collection of random variables, any finite subset of which follows a joint Gaussian distribution. In the context of reconstruction, we treat the observed data  $y$  and the latent function  $f(\mathbf{x})$  as realizations of the same Gaussian Process.

Accordingly, the function is modeled as

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, \tilde{x})), \quad (4.1)$$

where the process is fully characterized by its mean and covariance functions. The mean function is defined as

$$\mu(x) = \mathbb{E}[f(x)], \quad (4.2)$$

where  $\mathbb{E}[\cdot]$  denotes the expectation value taken over the Gaussian Process distribution.<sup>1</sup> The covariance function is given by

$$k(x, \tilde{x}) = \mathbb{E}[(f(x) - \mu(x))(f(\tilde{x}) - \mu(\tilde{x}))]. \quad (4.3)$$

By definition, the covariance between function values at two input points satisfies

$$\text{cov}(f(x), f(\tilde{x})) = k(x, \tilde{x}). \quad (4.4)$$

Throughout this chapter, for simplicity, we adopt the squared exponential covariance function,

$$k(x, \tilde{x}) = \sigma_f^2 \exp\left(-\frac{(x - \tilde{x})^2}{2\ell^2}\right). \quad (4.5)$$

This kernel is particularly convenient because it is smooth to all orders, meaning that it can be differentiated infinitely many times. This property makes it especially suitable when the goal is to reconstruct not only a function but also its derivatives. The squared exponential covariance function is specified by two hyperparameters<sup>2</sup>: the amplitude  $\sigma_f$  and the characteristic length scale  $\ell$ .

Given the dataset  $(\mathbf{X}, \mathbf{Y}) = \{x_i, y_i\}$ , the covariance matrix associated with the input values  $\mathbf{X}$  is defined as  $[K(\mathbf{X}, \mathbf{X})]_{ij} = k(x_i, x_j)$ . For a given input value  $\mathbf{X}$ , the reconstructed function corresponds to a Gaussian distribution. In this case, the covariance matrix of the data is obtained by adding the observational covariance matrix  $C$  to the chosen kernel function, such that

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, K(\mathbf{X}, \mathbf{X}) + C) \quad (4.6)$$

Here,  $\mathcal{N}$  denotes a Gaussian distribution for a random variable, and  $\boldsymbol{\mu}$  represents the

---

<sup>1</sup>For a finite-dimensional Gaussian distribution with probability density  $p(\mathbf{f})$ , the expectation value is defined as

$$\mathbb{E}[f_i] = \int f_i p(\mathbf{f}) d\mathbf{f}.$$

<sup>2</sup>In analogy with the hyperparameters introduced in the previous chapter for machine learning models (e.g., SVR), these quantities control the flexibility and smoothness of the reconstruction. While in ML algorithms hyperparameters typically regulate model complexity and generalization, in the GP framework they determine the amplitude of correlations and the characteristic correlation length of the function.

prior mean of the function evaluated at the input  $\mathbf{X}$ . When the observational data are assumed to be uncorrelated, the covariance matrix  $C$  becomes diagonal, with elements given by the individual variances  $\sigma_i^2$ . Let  $\mathbf{f}_*$  denote the (unknown) function values we aim to reconstruct, evaluated at the input locations  $\mathbf{X}_*$ . The joint GP describing both the observed data and the function values at the reconstruction points can then be written as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^* \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + C & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix} \right) \quad (4.7)$$

At this stage, the reconstruction problem amounts to determining the conditional distribution of  $\mathbf{f}^*$  given the observed data  $\mathbf{y}$ , namely  $\mathbf{f}^* | \mathbf{y}$ . For convenience, we introduce the following notation,  $\tilde{\mathbf{K}} \equiv \mathbf{K}(\mathbf{X}, \mathbf{X}) + C$ ,  $\mathbf{K}^* \equiv \mathbf{K}(\mathbf{X}, \mathbf{X}^*)$ ,  $\mathbf{K}^{**} \equiv \mathbf{K}(\mathbf{X}^*, \mathbf{X}^*)$  and

$$\mathbf{K} = \begin{bmatrix} \tilde{\mathbf{K}} & \mathbf{K}^* \\ \mathbf{K}^{*T} & \mathbf{K}^{**} \end{bmatrix} \quad (4.8)$$

This corresponds to a square matrix of dimension  $(n + n^*) \times (n + n^*)$ , where  $n$  denotes the number of observed data points and  $n^*$  represents the number of input locations at which the function is to be reconstructed. The associated joint probability distribution is then given by

$$p(\mathbf{y}, \mathbf{f}^*) = \frac{1}{(2\pi)^{N/2} \sqrt{\det(\mathbf{K})}} \exp \left\{ -\frac{1}{2} \begin{bmatrix} (\mathbf{y} - \boldsymbol{\mu})^T, & (\mathbf{f}^* - \boldsymbol{\mu}^*)^T \end{bmatrix} \mathbf{K}^{-1} \begin{bmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \mathbf{f}^* - \boldsymbol{\mu}^* \end{bmatrix} \right\}. \quad (4.9)$$

Based on this expression, and employing the properties of multivariate Gaussian distributions along with basic matrix identities, the marginalized distribution of  $\mathbf{y}$  can be obtained as

$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{f}^*) d\mathbf{f}^* \quad (4.10)$$

and finally the conditional probability distribution

$$\begin{aligned} p(\mathbf{f}^* | \mathbf{y}) &= \frac{p(\mathbf{y}, \mathbf{f}^*)}{p(\mathbf{y})} \\ &= \frac{1}{(2\pi)^{n^*/2} \sqrt{\det(\mathbf{K})}} \exp \left[ -\frac{1}{2} (\mathbf{f}^* - \bar{\mathbf{f}}^*)^T \text{cov}(\mathbf{f}^*)^{-1} (\mathbf{f}^* - \bar{\mathbf{f}}^*) \right] \\ &= \mathcal{N}(\bar{\mathbf{f}}^*, \text{cov}(\mathbf{f}^*)). \end{aligned} \quad (4.11)$$

where

$$\bar{\mathbf{f}}^* = \boldsymbol{\mu}^* + \mathbf{K}^{*T} \tilde{\mathbf{K}}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \quad (4.12)$$

and

$$\text{cov}(\mathbf{f}^*) = \mathbf{K}^{**} - \mathbf{K}^{*T} \mathbf{K}^{-1} \mathbf{K}^*. \quad (4.13)$$

The reconstruction uncertainty, or variance, is given by the diagonal elements of this matrix. An important feature that follows from the previous expressions is the dependence of the expected reconstructed function  $\mathbf{f}^*$  on the choice of the prior mean function  $\boldsymbol{\mu}$ . A conservative approach is to adopt a constant mean function, typically chosen as  $\mu(x) = 0$ .

Examining the reconstruction variance (i.e., the diagonal elements of Eq. (4.13)), and recalling that the covariance function is positive definite Eq. (4.5), such that  $K^{**} \geq 0$  and  $K^* \geq 0$ , one finds that the uncertainty reaches its maximum when  $K^* \rightarrow 0$ . This situation occurs when the points  $x^*$  at which the function is reconstructed lie far from all available observations (WAN and SAPSIS, 2017).

Finally, the values of the hyperparameters must be determined. This can be achieved by maximizing the likelihood function of the observed data, conditioned on the chosen covariance matrix and prior model. The corresponding probability distribution is given by

$$\begin{aligned} \mathcal{L} &= p(\mathbf{y} \mid \mathbf{X}, \sigma_f, \ell) \\ &= \frac{1}{(2\pi)^{n/2} \sqrt{\det(\tilde{\mathbf{K}})}} \exp \left[ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \tilde{\mathbf{K}}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right]. \end{aligned} \quad (4.14)$$

For computational efficiency, it is more convenient to maximize the logarithm of the previous equation rather than the expression itself.

Once the optimal hyperparameters are determined, the mean of the reconstructed function and its associated covariance can be computed using Eqs. (4.12) and (4.13), respectively.

To illustrate the role of hyperparameters in Gaussian Process reconstruction, we revisit the example from the previous chapter, where the Hubble function  $H(z)$  was reconstructed using SVR and ESR techniques.

Here, we vary only the correlation-length hyperparameter  $\ell$ , keeping the others fixed to isolate its impact. As shown in Fig. 4.1, different values of  $\ell$  lead to distinct reconstructed behaviors.

Small  $\ell$  values reduce long-range correlations, allowing rapid local variations that closely follow the data. Larger  $\ell$  values impose stronger correlations over wider intervals, producing smoother curves.

This comparison highlights the sensitivity of the reconstruction to hyperparameter choices and emphasizes the need for marginal likelihood optimization.

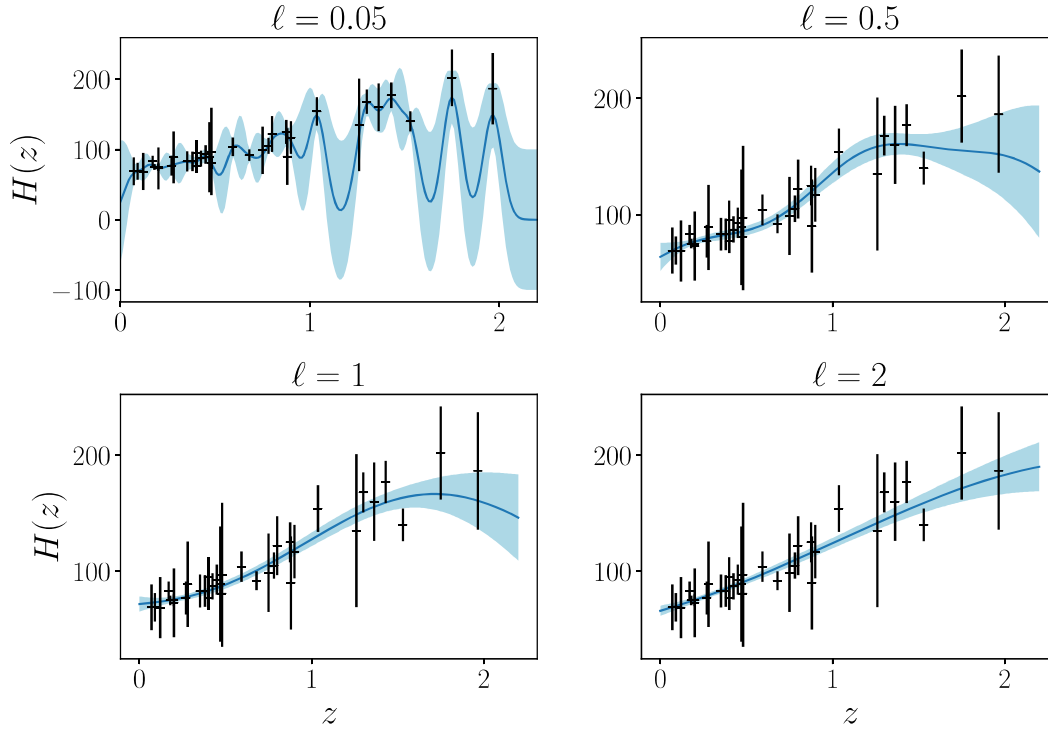


Figure 4.1: Reconstruction of the Hubble function  $H(z)$  from CC data for different values of the correlation-length hyperparameter  $\ell$ . The solid line represents the mean reconstructed function, while the shaded region indicates the  $1\sigma$  confidence interval. The data points with error bars correspond to the observational measurements. The upper-left, upper-right, lower-left, and lower-right panels correspond to  $\ell = 0.05, 0.5, 1$ , and  $2$ , respectively.

According to SEIKEL *et al.* (2012), the key steps involved in constructing a GP are:

- Choose the data set.
- Choose a covariance function.
- Select the points  $x_i^*$  at which the function  $f_i^*$  is to be estimated.
- Define a prior mean function  $\mu(x)$ ; assuming  $\mu = \text{const.}$  is a safe choice.
- Train the hyperparameters:
  - Carefully choose the initial values of the hyperparameters. It is recommended to try different initial values, since the optimization procedure may get stuck in local maxima.
  - Maximize the likelihood function, Eq. (4.14), with respect to the hyperparameters.
- Compute  $\mathbf{f}^*$  from Eq. (4.12), which provides the expected value of the function.

- Compute the diagonal elements of  $\text{cov}(\mathbf{f}^*)$ , Eq. (4.13), which give the variance of  $\mathbf{f}^*$ .

### 4.3 Reconstructing the derivative of a function

The derivative of a function can be inferred by noting that it also follows a Gaussian Process. While preserving the covariance structure of the original data, one must additionally compute the cross-covariance between the function values and their derivatives, as well as the covariance between the derivatives themselves. These quantities are obtained by directly differentiating the original covariance matrix, namely,

$$\text{cov}\left(f_i, \frac{\partial f_j}{\partial x_j}\right) = \frac{\partial k(x_i, x_j)}{\partial x_j}, \quad (4.15)$$

$$\text{cov}\left(\frac{\partial f_i}{\partial x_i}, \frac{\partial f_j}{\partial x_j}\right) = \frac{\partial^2 k(x_i, x_j)}{\partial x_i \partial x_j}. \quad (4.16)$$

The first derivative can be obtained from the joint distribution of the observed data  $\mathbf{y}$  and  $\mathbf{f}^{*'}$ , given by

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}^{*'} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^{*'} \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}'^* \\ \mathbf{K}'^T & \mathbf{K}''^* \end{pmatrix}\right), \quad (4.17)$$

where

$$[\mathbf{K}'^*]_{ij} = \frac{\partial k(x_i, x_j^*)}{\partial x_j^*}, \quad (4.18)$$

and

$$[\mathbf{K}''^*]_{ij} = \frac{\partial^2 k(x_i, x_j^*)}{\partial x_i \partial x_j^*}. \quad (4.19)$$

From these relations, the reconstructed mean of the derivative is

$$\overline{\mathbf{f}'^*} = \boldsymbol{\mu}^{*'} + \mathbf{K}'^T \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \quad (4.20)$$

and its covariance is

$$\text{cov}(\mathbf{f}'^*) = \mathbf{K}''^* - \mathbf{K}'^T \mathbf{K}^{-1} \mathbf{K}'^*. \quad (4.21)$$

The same hyperparameters obtained in the reconstruction of  $\mathbf{f}^*$  can be used to reconstruct its derivative. This follows from the fact that the likelihood depends only on the observed data and on the prior mean function  $\mu$ , rather than on the specific quantity being predicted (see Eq. (4.14)). Higher-order derivatives can be computed analogously.

## 4.4 Functions of a GP and Error Propagation

In many cosmological applications, the quantity of interest is not the reconstructed function itself, but a derived observable depending on it and possibly on its derivatives, e.g.,  $g(f(x), f'(x), \dots)$ . A natural first approach would be to evaluate  $g$  at the mean reconstruction and propagate the associated uncertainties using standard error propagation formulas.

Such a procedure corresponds to expanding  $g$  around the mean values and combining the individual variances in quadrature. This approximation is typically adequate when the uncertainties are small and the function  $g$  varies smoothly over the relevant parameter range.

However, within the Gaussian Process framework, the reconstructed function and its derivatives are jointly Gaussian distributed. Therefore, a more consistent treatment consists of exploiting their full covariance structure when computing derived quantities. In this way, correlations between  $f(x)$  and its derivatives are properly accounted for, leading to a more accurate estimate of the uncertainty in  $g$ .

In a more precise formulation, computing a reconstructed function via GPs and its derivatives requires knowledge of the covariance between the involved functions. In other words, one must know the correlation between  $f^* = f(x^*)$ ,  $f^{*'} = f'(x^*)$ , and higher-order derivatives at each reconstruction point  $x^*$ .

The covariance matrices between  $f$  and each of its derivatives  $f^{(i)}$  (the  $i$ -th derivative of  $f$ ) evaluated at the point  $x^*$  are given by

$$\text{cov}(f^{(i)}, f^{(j)}) = k^{(i,j)}(x^*, x^*) - K^{(i)}(x^*, X) [K(X, X) + C]^{-1} K^{(j)}(X, x^*), \quad (4.22)$$

where  $k^{(i,j)}(x^*, x^*)$  denotes the  $i$ -th derivative of the covariance function (Eq. 4.4) with respect to the first argument and the  $j$ -th derivative with respect to the second argument, both evaluated at  $(x^*, x^*)$ .

With this information, one can obtain  $g(f(x^*), f'(x^*), \dots)$  through a Monte Carlo simulation, in which the values of  $f^{(i)}$  are sampled from a multivariate Gaussian distribution whose correlation matrix has a block structure composed of the covariance matrices defined in Eq. (4.23). Thus, the Monte Carlo simulation used to determine  $g$  at each value of  $x^*$  is drawn from the distribution

$$\begin{pmatrix} f^* \\ f^{*'} \\ \vdots \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \bar{f}^* \\ \bar{f}^{*'} \\ \vdots \end{pmatrix}, \begin{pmatrix} \text{var}(f^*) & \text{cov}(f^*, f^{*'}) & \cdots \\ \text{cov}(f^{*'}, f^*) & \text{var}(f^{*'}) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \right). \quad (4.23)$$

This Monte Carlo procedure may yield asymmetric uncertainties, in contrast to the standard error-propagation approach.

## 4.5 A Practical Example: Reconstruction of $H(z)$ via GP

We now return to the example of reconstructing  $H(z)$ , optimizing the hyperparameters, and subsequently computing the derivatives of the reconstructed function. In this application, we use the CC measurements as our data set. For these data, the observational uncertainties are uncorrelated, so the covariance matrix is purely diagonal.

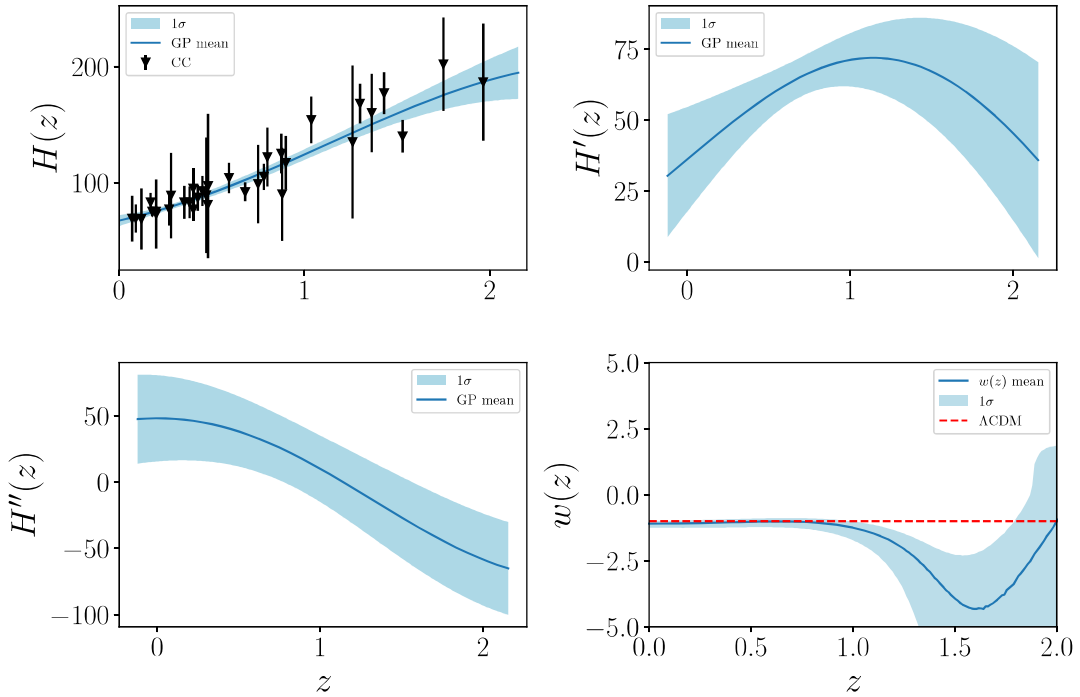


Figure 4.2: Reconstruction of  $H(z)$ ,  $H'(z)$ ,  $H''(z)$ , and the dark energy equation-of-state parameter  $w(z)$  using Gaussian Processes (GP). The shaded regions represent the  $1\sigma$  confidence intervals around the GP mean. The upper-left panel shows  $H(z)$  together with the cosmic chronometer (CC) data. The remaining panels display the reconstructions of  $H'(z)$  and  $H''(z)$ . The lower-right panel presents the reconstructed  $w(z)$  with its  $1\sigma$  uncertainty, compared to the  $\Lambda$ CDM prediction.

We adopt the Squared Exponential (SE) kernel for the Gaussian Process reconstruction and assume a vanishing mean function,  $\mu(z) = 0$ , which corresponds to the standard default choice in GP implementations. The kernel hyperparameters are optimized by maximizing the marginal likelihood, as described in the previous section.



Figure 4.2 presents the GP reconstruction of  $H(z)$ . The upper-right panel displays the first derivative,  $H'(z)$ , while the lower-right panel shows the second derivative,  $H''(z)$ . The derivatives are obtained analytically from the GP, ensuring a fully consistent propagation of uncertainties from the original data to the reconstructed quantities.

Using Eq. (5.3), and expressing the result in terms of the reconstructed function and its derivative, we can obtain  $w(z)$ , as shown in the lower-left panel of Figure 4.2. In this calculation, we fix the background cosmological parameters to the values reported by PLANCK COLLABORATION (2020) results, namely  $H_0 = 67.4 \text{ km s}^{-1} \text{ Mpc}^{-1}$  and  $\Omega_m = 0.315$ . This example illustrates how to compute a derived quantity from reconstructed observables, thereby requiring the error propagation procedure described in the previous section.

In this chapter, we have presented the theoretical foundations of Gaussian Processes and detailed their application to the non-parametric reconstruction of cosmological functions and their derivatives. We discussed the role of the covariance function, the optimization of hyperparameters through marginal likelihood maximization, and the consistent propagation of uncertainties, including the treatment of derived observables via Monte Carlo sampling of the full covariance structure. As a practical illustration, we applied this framework to the reconstruction of the Hubble function  $H(z)$  and the derived quantity  $w(z)$  using cosmic chronometer data.

In the next chapter, we present the first original results of this thesis, in which the methodology developed here is applied to current observational data to investigate the cosmological implications of our reconstructions.

# CHAPTER 5

---

## Symbolic regression analysis of dynamical dark energy with DESI-DR2 and SN data

---

*This chapter is based on the published article: SOUSA-NETO et al. (2025).*

### 5.1 Introduction

In this chapter we employ a novel model-independent approach, namely Symbolic Regression using PySR, on DESI-BAO and SN data to reconstruct the dark energy EoS. Hence, we can check whether there is any statistically significant departure of the  $w = -1$  prediction from the standard model, which would help supporting or challenging previous results in the literature.

In what follows, we first review the relevant cosmological equations for our study in Section 5.2, including the key concepts under consideration. Section 5.3 describes the data used in our analysis, detailing the specifics of the DESI, Patheon+, and DESY5 datasets. In Section 5.4, we discuss the theoretical framework of symbolic regression, focusing on its application to cosmological data. Section 5.5 presents the main results of our analysis, while Section 5.6 reports the corresponding results obtained with Gaussian Processes. Finally, Section 5.7 summarizes our main conclusions.

### 5.2 Basic Equations

We can express the equation of state (EoS),  $w(z)$ , in terms of observable distances. To this end, we first consider a spatially flat Universe, for which the normalized Hubble parameter is given by Eq. (2.13),

$$\frac{H(z)^2}{H_0^2} = \Omega_m(1+z)^3 + \Omega_{DE} \exp \left[ 3 \int_0^z \frac{1+w(z')}{1+z'} dz' \right], \quad (5.1)$$

With the assumption of a spatially flat universe, the normalized density parameters satisfy the relation  $\Omega_m + \Omega_{DE} = 1$ .

$$\ln \left( \left[ \frac{H(z)^2}{H_0^2} - \Omega_m(1+z)^3 \right] \frac{1}{(1-\Omega_m)} \right) = 3 \int_0^z \frac{1+w(z')}{1+z'} dz', \quad (5.2)$$

One can invert the above equation to derive an expression for the dark energy EoS,  $w(z)$ , in terms of the  $H(z)$ , and its first derivative,

$$w(z) = -\frac{1}{3} \frac{2(1+z)HH' - 3H^2}{H_0^2(1+z)^3\Omega_m - H^2}, \quad (5.3)$$

which can be rewritten as DINDA and MAARTENS (2024)

$$w(z) = -1 - \frac{2(1+z)\tilde{D}'_H(z) + 3\gamma(1+z)^3\tilde{D}_H^3(z)}{3[\tilde{D}_H(z) - \gamma(1+z)^3\tilde{D}_H^3(z)]}. \quad (5.4)$$

In the above expression, we rewrite Eq. (2.17) as  $\tilde{D}_H(z) = D_H/r_d$ , where  $r_d$  is the sound horizon at the drag epoch. Writing the expression in this form is important because it allows the quantity to be expressed in terms of the observable  $D_H/r_d$  and

$$\gamma = \frac{\Omega_m H_0^2 r_d^2}{c^2}, \quad (5.5)$$

Note that  $D_M$  is related to  $D_H$  by  $D'_M = c/H(z) = D_H$ , allowing us to rewrite Eq. (5.4) as

$$w(z) = -1 - \frac{2(1+z)\tilde{D}''_M(z) + 3\gamma(1+z)^3\tilde{D}_M^3(z)}{3[\tilde{D}'_M(z) - \gamma(1+z)^3\tilde{D}_M^3(z)]}. \quad (5.6)$$

For analyses involving SNe data,  $\tilde{D}_M$  is computed using the second equation above, which relates it to the luminosity distance  $D_L$  through Eq. (2.31), divided by  $r_d$

$$\tilde{D}_M = \frac{D_L}{(1+z)r_d} = \frac{10^{\frac{m-M-25}{5}}}{(1+z)r_d}. \quad (5.7)$$

We can now reconstruct  $w(z)$  from  $H(z)$  using Eq. (5.3), or alternatively from  $\tilde{D}_H$  and  $\tilde{D}_M$  through Eqs. (5.4) and (5.6), respectively. Later on, we return to the latter two approaches and discuss different methods for computing the derivatives of these quantities.

## 5.3 Data

This section describes the observational datasets used in this thesis, including BAO measurements from DESI DR2 (KARIM *et al.*, 2025) and SNe data from the Pantheon+ (SCOLNIC *et al.*, 2022) and DESY5 (ABBOTT *et al.*, 2024) samples.

### 5.3.1 Baryonic acoustic oscillations

The DESI-BAO measurements are expressed by three ratios: the three-dimensional BAO mode ( $D_V/r_d$ ), the transverse mode ( $D_M/r_d$ ), and the radial mode ( $D_H/r_d$ ). In this thesis, we focus on the transverse and radial modes. An important point is that to reconstruct  $w(z)$ , we use only the measurements  $D_H/r_d$  or  $D_M/r_d$  in our analyses, depending on the equation used (Eq.(5.4) or Eq.(5.6)). For this reason, we do not use  $D_V/r_d$  from DESI data. For analyses using only DESI data, we use exclusively  $D_H/r_d$  because the large uncertainties involved in the second derivative in Eq.(5.6) make it implausible to reconstruct  $w(z)$  using only DESI with  $D_M/r_d$  measurements.

The radial comoving distance can be related to the transverse BAO mode measurement by

$$\frac{D_M}{r_d} = \frac{D_A(1+z)}{r_d}, \quad (5.8)$$

where  $D_A$  is the angular diameter distance. Similarly, the radial BAO mode can be written as

$$\frac{D_H(z)}{r_d} = \frac{c}{H(z)r_d}. \quad (5.9)$$

The BAO measurements obtained by the DESI collaboration are summarized in Table 5.1. There, we present the measurement values for different tracers, namely the Luminous Red Galaxy (LRG), divided into two samples (LRG1, LRG2), the combined LRG3 and Emission Line Galaxy (ELG1), ELG2, Lyman- $\alpha$  forest, and the QSO sample. The first column indicates the tracers, the second one the effective redshift,  $z_{eff}$ , while the third and fourth columns list the  $D_H(z)/r_d$  and  $D_M(z)/r_d$  ratios, along with their  $1\sigma$  confidence limits. The last shows the cross-correlation coefficients,  $r_{M,H}$ , between  $D_M/r_d$  and  $D_H/r_d$ .

| tracer      | $z_{eff}$ | $D_H/r_d$          | $D_M/r_d$          | $r_{M,H}$ |
|-------------|-----------|--------------------|--------------------|-----------|
| LRG1        | 0.510     | $21.863 \pm 0.425$ | $13.588 \pm 0.167$ | $-0.459$  |
| LRG2        | 0.706     | $19.455 \pm 0.330$ | $17.351 \pm 0.177$ | $-0.404$  |
| LRG3+ELG1   | 0.934     | $17.641 \pm 0.193$ | $21.576 \pm 0.152$ | $-0.416$  |
| ELG2        | 1.321     | $14.176 \pm 0.221$ | $27.601 \pm 0.318$ | $-0.434$  |
| QSO         | 1.484     | $12.817 \pm 0.516$ | $30.512 \pm 0.760$ | $-0.500$  |
| Ly $\alpha$ | 2.330     | $8.632 \pm 0.101$  | $38.988 \pm 0.531$ | $-0.431$  |

Table 5.1: BAO measurements from the DESI (DR2) collaboration KARIM *et al.* (2025).

### 5.3.2 Type Ia Supernova

#### 5.3.2.1 Pantheon+

The Pantheon+ dataset encompasses 20 supernova compilations spanning a redshift range from 0.00122 to 2.26137 SCOLNIC *et al.* (2022). This dataset includes 1701 light curves corresponding to 1550 spectroscopically confirmed SNe. At low redshifts,  $z < 0.01$ , 111 SNe have been removed in order to prevent a strong peculiar velocity dependence. The 1590 light curves in our final sample span a redshift range from  $0.01016 \leq z \leq 2.26137$ . The data and the covariance matrix used in this thesis are available in the GitHub repository<sup>1</sup>.

#### 5.3.2.2 Dark Energy Survey Supernova Year 5 (DES Y5)

The DES Y5 supernova sample comprises 1635 supernovae observed by the DES collaboration, spanning a redshift range from 0.0596 to 1.12, along with 194 high-quality external supernovae at redshifts below 0.1, for a total of 1829 supernovae. After excluding supernovae with magnitude errors  $\delta m > 1$ , which are unlikely to be classified as Type Ia based on the DES photometric data, our final sample consists of 1754 supernovae. The complete catalog is accessible through the DES collaboration's GitHub repository<sup>2</sup>.

## 5.4 Symbolic regression

In this thesis, SR is employed as a non-parametric tool for the reconstruction of cosmological observables.

As discussed in Chapter 3, we adopt the PySR framework CRANMER (2023a) to perform symbolic regression. In the following, we focus on the specific choices relevant to our analysis, including the set of operators, the loss function, and the model selection criteria.

### 5.4.1 Operators Used in PySR

In symbolic regression with PySR, symbolic expressions are constructed and evolved through combinations of unary and binary operators. As discussed in Chapter 3, the choice of operators plays a central role in defining the functional search space and directly impacts the complexity and performance of the resulting models. In this thesis, we adopt a predefined set of unary and binary operators to explore possible solutions and identify the best-fitting models, as summarized in Table 5.2.

---

<sup>1</sup>[https://github.com/PantheonPlusSHOES/DataRelease/tree/main/Pantheon%2B\\_Data](https://github.com/PantheonPlusSHOES/DataRelease/tree/main/Pantheon%2B_Data)

<sup>2</sup><https://github.com/des-science/DES-SN5YR>.

Table 5.2: List of unary and binary operators used in PySR

| unary  | binary |
|--------|--------|
| log    | —      |
| sqrt   | +      |
| exp    | *      |
| cube   | /      |
| square | pow    |

Additionally, we assessed the performance of three different operator sets while keeping both the dataset (DESI-DR2 + DES Y5) and the loss function (`CustomLoss`) fixed. These include the operator set described above, which we refer to as the *original* set, as well as two alternative operator configurations. The purpose of this analysis is to assess the robustness of our results with respect to the choice of operator set used in this chapter. The two additional operator sets are defined as follows:

- **Simple:** Binary =  $\{+, -, *\}$ , Unary =  $\{\text{sqrt}, \text{square}\}$ .
- **Complex:** Binary =  $\{+, -, *, /, \text{pow}\}$ , Unary =  $\{\text{sqrt}, \text{square}, \text{cube}, \text{exp}, \text{log}, \text{sin}, \text{cos}\}$ .

As shown in Fig. 5.1, the three operator sets achieve comparable performance, with overlapping distributions and similar medians. However, the Original operator set adopted in our chapter yielded the lowest mean MSE, indicating that it provides a good balance between expressiveness and model simplicity. These results suggest that while more complex operator sets introduce additional functional forms, they do not necessarily improve predictive accuracy. In this case, the chosen operator set appears to be the most effective.

### 5.4.2 Loss Function

As shown in Chapter 3, an important hyperparameter in PySR is the loss function. To properly account for the uncertainties in the data and ensure that the model assigns appropriate weights to each observation, we adopt a custom loss function. We adopt the loss function defined as

$$\mathcal{L} = \frac{1}{n} \mathbf{r}^T \mathbf{C}^{-1} \mathbf{r} , \quad (5.10)$$

where  $\mathbf{r} = x - y$  is the residual vector, with  $x$  representing the predicted value and  $y$  denoting the true value;  $\mathbf{C}^{-1}$  is the inverse covariance matrix; and  $n$  is the number of points in the sample, used to normalize the loss. This approach enables the model to

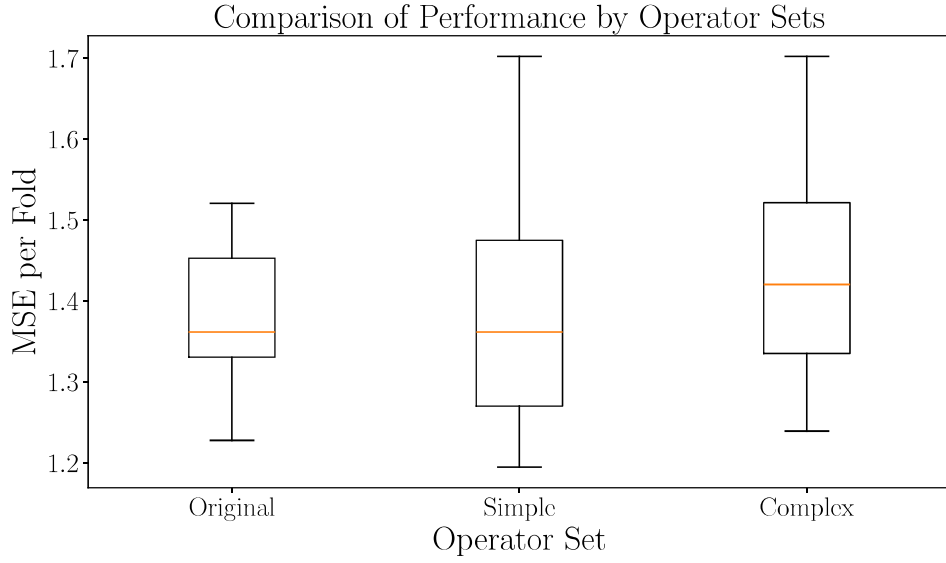


Figure 5.1: Comparison of performance across different operator sets using `CustomLoss`.

more accurately account for data quality and focus on more reliable observations during optimization.

For completeness, we also explore the different loss functions available in PySR. Besides the loss function described in this section, we considered the Huber Loss Eq. (3.1), the L1 Epsilon-Insensitive Loss Eq. (3.2), and the LP Distance Loss Eq. (3.3). Among these, the previously described loss function showed the best overall performance, demonstrating the robustness of our analysis.

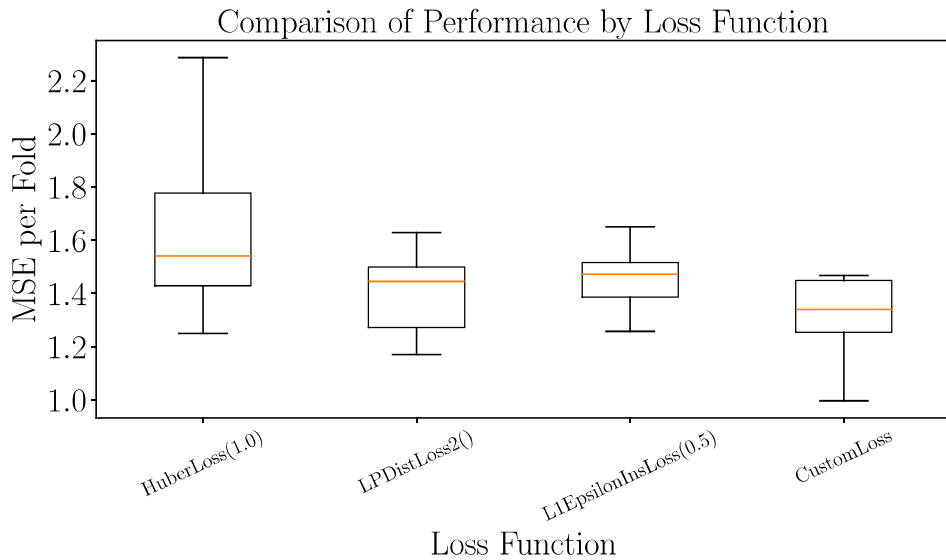


Figure 5.2: Comparison of the Mean Squared Error (MSE) distribution per cross-validation fold for the four tested loss functions.

Figure 5.2 presents a comparative analysis of model performance using four distinct

loss functions, evaluated by the Mean Squared Error (MSE) across cross-validation folds. The proposed CustomLoss function, which uniquely leverages the full data covariance matrix, also demonstrates markedly superior performance. It not only yields the lowest median MSE but also exhibits the least variability, as indicated by its compact interquartile range. This suggests that incorporating the data’s complete covariance structure provides a more robust and accurate model fit compared to the standard loss functions tested.

### 5.4.3 Model Selection

We adopt the *best* selection criterion implemented in PySR, following the discussion presented in Chapter 3. The final expressions selected through this procedure are presented in Tables 5.4, 5.5, and 5.6.

Although these expressions exhibit good statistical performance, they do not inherently guarantee physical meaning, as they are guided solely by data fitting and structural simplicity. For this reason, it is sometimes necessary to impose physical constraints, such as  $D_M = 0$  at  $z = 0$ , to guide the algorithm toward learning representations consistent with physical expectations.

To illustrate the effect of complexity penalties, we chose to fix a given dataset, for instance, DESI-DR2 + DESY5, and varied  $\alpha$ , which represents our complexity penalty. Even with different values of  $\alpha$  (0.001, 0.01, 0.1) and different models, the general form of

Table 5.3: Top 3 PySR results for different values of  $\alpha$ .

| $\alpha$ | Metric | MSE    | Complexity | Equation                 |
|----------|--------|--------|------------|--------------------------|
| 0.001    | 1.285  | 1.2780 | 7          | $30.08 z/1.31^z$         |
|          | 1.287  | 1.2780 | 9          | $z(z(z - 8.04) + 29.99)$ |
|          | 1.289  | 1.2790 | 10         | $z^2 + 30.13 z/1.37^z$   |
| 0.01     | 1.348  | 1.2780 | 7          | $30.08 z/1.31^z$         |
|          | 1.358  | 1.2880 | 7          | $29.48 z/1.28^z$         |
|          | 1.366  | 1.2960 | 7          | $29.41 z/1.28^z$         |
| 0.1      | 1.978  | 1.2780 | 7          | $30.08 z/1.31^z$         |
|          | 1.988  | 1.2880 | 7          | $29.48 z/1.28^z$         |
|          | 1.996  | 1.2960 | 7          | $29.41 z/1.28^z$         |

the top PySR equations remained the same (see Table 5.3). This demonstrates the robustness of symbolic regression in identifying consistent functional forms, while  $\alpha$  primarily



affects the trade-off between metric and complexity.

## 5.5 Results

In what follows, we present the reconstructed results obtained using the method described in the previous section. First, we show the results for the DESI DR2 dataset. Subsequently, we combine the DESI-BAO data with the two supernova datasets described earlier.

### 5.5.1 DESI-only

For completeness, we begin our analysis by reconstructing the quantity  $D_H/r_d$  from the DESI-BAO dataset. The best-fit equation obtained by PySR to describe these data is given by

$$\frac{D_H}{r_d} = 28.4 \exp(-0.519z) , \quad (5.11)$$

which corresponds to the model selected using the “**best**” criterion described in Section 5.4.3. Other candidate equations generated during the training, along with their complexity and loss, are listed in Table 5.4. The results are displayed in the left panel of

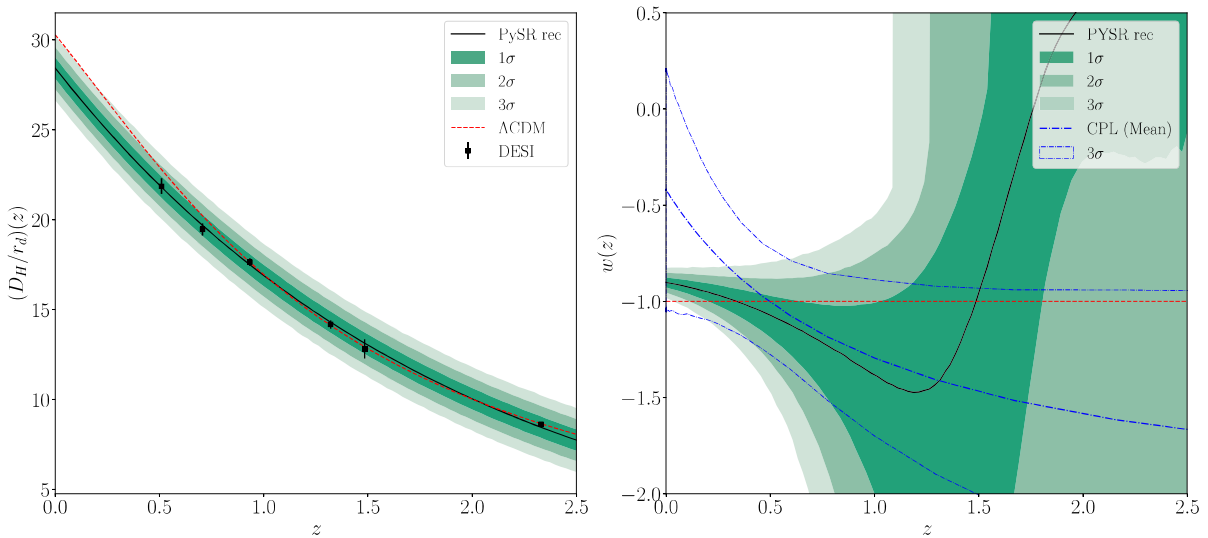


Figure 5.3: The reconstruction results for  $D_H/r_d$  DESI-only. The black line represents the mean for the PySR reconstruction; the green bands indicate the CLs. The red dashed line corresponds to the  $\Lambda$ CDM theoretical curve, while the blue dashed line represents the CPL model with the  $w_0 - w_a$  DESI-BAO+CMB results in  $3\sigma$  CL KARIM *et al.* (2025). **Left:**  $D_H/r_d$  reconstruction with PySR. The black points correspond to the DESI-BAO dataset, with respective uncertainties. **Right:** The  $w(z)$  derived using Eq. (5.4) and the reconstructed  $D_H/r_d$ , assuming  $r_d = 101 \pm 2.3h^{-1}$  Mpc, as reported by VERDE *et al.* (2017), and  $\Omega_m = 0.3$ .

Table 5.4: DESI-only

| Complexity | Loss        | Equation                                                                                        |
|------------|-------------|-------------------------------------------------------------------------------------------------|
| 1          | 2519.06     | $y = 2.8069465$                                                                                 |
| 2          | 120.73      | $y = \sqrt{276.74304}$                                                                          |
| 3          | 91.5        | $y = 17.705526 - z$                                                                             |
| 4          | 1.22        | $y = (5.127707 - z)^2$                                                                          |
| <b>5</b>   | <b>0.30</b> | <b><math>y = 28.42 / (1.68^z)</math></b>                                                        |
| 7          | 0.25        | $y = (3.77 - z)^{2.32} + 6.24$                                                                  |
| 8          | 0.24        | $y = ((z - 3.18)^2 - 3.47) \times 2.06$                                                         |
| 9          | 0.21        | $y = 2.70^z - 11.8 \times z + 26.0$                                                             |
| 11         | 0.17        | $y = 4.19 \times (1 - 0.787 \times z)^6 + (4.94 - z)^{2.1}$                                     |
| 13         | 0.02        | $y = (3.68 - z)^{2.4} + 6.56 - \frac{0.02}{z-0.76}$                                             |
| 15         | 0.016       | $y = (3.68 - z)^{2.37} + 6.56 - \frac{0.0169}{0.862(z-0.763)}$                                  |
| 16         | 0.015       | $y = 21.9 \times (1 - 0.272 \times z)^{2.37} + 6.56 + \frac{0.0169}{\sqrt{z \times (z-0.763)}}$ |
| 17         | 0.013       | $y = (3.68 - z)^{2.37} + 6.56 - \frac{0.0169}{z \times (z-0.763-0.0216)}$                       |

Figure 5.3, along with the  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  confidence intervals. These intervals are estimated using the Monte Carlo method, assuming a Gaussian distribution centered on the best-fit equation provided by PySR, with standard deviations given by the reconstruction uncertainties.

As for the dark energy EoS,  $w(z)$ , we can compute it by plugging the reconstructed  $D_H(z)/r_d$  given by Eq. (5.11) and its derivative into Eq. (5.4), as presented in the right panel of Figure 5.3. The dashed red line represents the  $\Lambda$ CDM scenario,  $w(z) = -1$ , while the blue curve corresponds to the CPL parameterization with the values  $w_0 = -0.42 \pm 0.21$  and  $w_a = -1.75 \pm 0.58$  from the DESI-BAO results KARIM *et al.* (2025). The green-shaded regions correspond to the reconstruction's  $1\sigma$  to  $3\sigma$  CL (from darker to lighter shade, respectively), also obtained using the Monte Carlo method. At  $1\sigma$  level, this analysis shows good agreement with the  $\Lambda$ CDM prediction up to  $z \approx 0.66$ .

In our final analysis using DESI-only data, we examine the sensitivity of  $w(z)$  to variations in the parameters  $\Omega_m$  and  $r_d$ , which enters in the calculation of  $\gamma$  – Eq. (5.5). Specifically, we explore  $\Omega_m$  values of 0.25, 0.30, and 0.35, while testing  $r_d$  at  $90.9 h^{-1}$  Mpc,  $101 h^{-1}$  Mpc, and  $111.1 h^{-1}$  Mpc, based on values reported by VERDE *et al.* (2017) (see also LEMOS *et al.* (2023) for model-independent estimates of  $r_d$  from angular BAO mea-

surements).

### 5.5.2 DESI + Pantheon+

Our first analysis, which combines the DESI dataset with SNe data, focuses on the Pantheon+ dataset. We use the apparent magnitude,  $m$ , from the Pantheon+ compilation to reconstruct the quantity  $D_M/r_d$  fixing  $M = -19.253$  RIESS *et al.* (2022). Then, we compute the normalized comoving distance ( $D_M$ ) of the SNe assuming Eq. (5.7) and  $r_d = 138.28 \pm 3.1$  Mpc, which corresponds to the  $r_d$  value given by VERDE *et al.* (2017) combined with the  $H_0$  value reported in RIESS *et al.* (2022) – which is consistent with the  $M$  value previously assumed. Finally, we compute the corresponding  $D_M/r_d$  values from the SNe apparent magnitude measurements, so that they can be combined with the  $D_M/r_d$  from the DESI-BAO measurements. This scenario is denoted as DESI + Pantheon+.

It is important to note that the SNe data do not constrain  $H_0$  and  $M$  independently, but the combination  $\mathcal{M} = M - 5 \log H_0$ . In this sense, the values of  $M$  and  $H_0$  must satisfy this relation to be compatible for a specific SNe data set. Thus, Eq. (5.7) can be rewritten in terms of  $\mathcal{M}$  and  $H_0$  to obtain:

$$\tilde{D}_M \propto \frac{10^{-\mathcal{M}/5}/H_0}{r_d}. \quad (5.12)$$

Since an uncalibrated sound horizon scale is used (Ref. VERDE *et al.* (2017)), where

$$r_d \propto \frac{1}{H_0}, \quad (5.13)$$

our SNe data expressed in terms of  $\tilde{D}_M$  and the parameter  $\gamma$  in Eq. (5.5) are independent of the  $H_0$  value. Consequently, the reconstruction of the dark energy EoS in Eq. (5.6) does not depend on the value of  $H_0$  and  $M$ .

The best equation derived by PySR for the  $D_M/r_d$  reconstruction, in the first scenario, is given by

$$\frac{D_M}{r_d} = 33.3z \exp(-0.274z). \quad (5.14)$$

The left panel of Figure 5.4 shows the reconstruction of  $D_M(z)/r_d$  for the DESI + Pantheon+ dataset. The reconstructed  $D_M/r_d$  values agree with the  $\Lambda$ CDM model within  $1\sigma$  for  $z \lesssim 1.7$ , but deviate from it at larger  $z$ , reaching the  $3\sigma$  region. A probable reason for this deviation is the inclusion of Pantheon+ data points at  $z > 1$ , where larger uncertainties and possible systematics are more prevalent. In contrast, DESI data alone show a better agreement with the  $\Lambda$ CDM model in high-redshift regions.

The result of the  $w(z)$  reconstruction for the DESI + Pantheon+ is shown in the right panel of Figure 5.4.

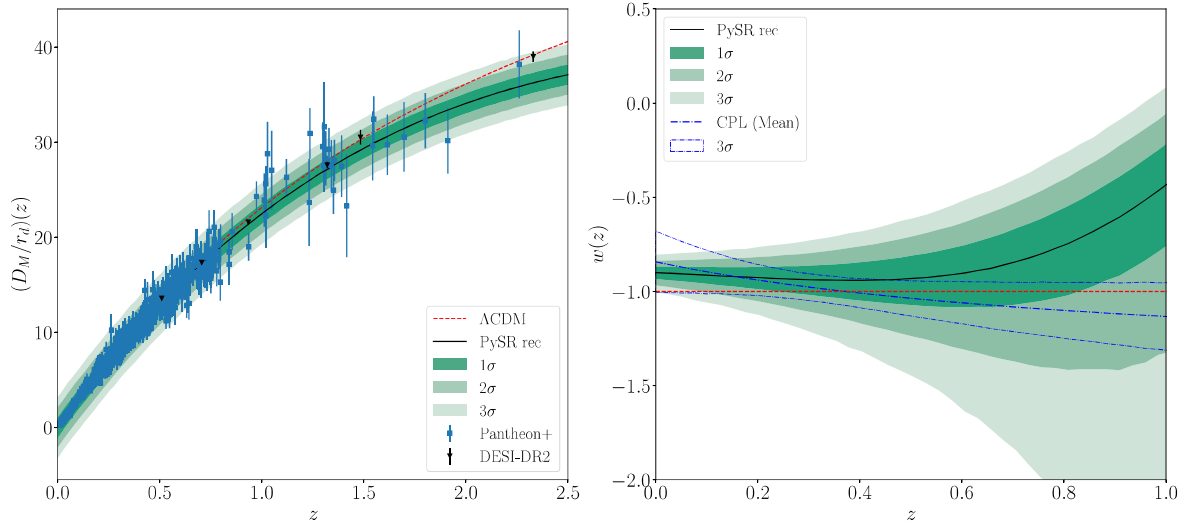


Figure 5.4: The reconstruction results for  $D_M/r_d$  using DESI + Pantheon+ dataset. The black line represents the mean for the PySR reconstruction; the green bands indicate the confidence levels. The red dashed line corresponds to the  $\Lambda$ CDM theoretical curve. **Left:**  $D_M/r_d$  reconstruction with PySR. The black points correspond to the dataset, with respective uncertainties. **Right:** The  $w(z)$  derived using Eq. (5.6) and the reconstructed  $D_M/r_d$ , assuming  $r_d = 101 \pm 2.3h^{-1}$  Mpc, as reported by VERDE *et al.* (2017), and  $\Omega_m = 0.3$ . While the blue dashed region represents the CPL model with the DESI + CMB + Pantheon+ results KARIM *et al.* (2025) in  $3\sigma$  CL.

Table 5.5: DESI + Pantheon+.

| Complexity | Loss       | Equation                                                                   |
|------------|------------|----------------------------------------------------------------------------|
| 1          | 124.5      | $y = 1.24$                                                                 |
| 2          | 116.74     | $y = e^z$                                                                  |
| 3          | 4.65       | $y = 27.1 \times z$                                                        |
| 5          | 2.69       | $y = (z/0.034)^{0.915}$                                                    |
| 6          | 2.39       | $y = z \times 30.20 \times (0.18 \times z - 1)^2$                          |
| <b>7</b>   | <b>1.7</b> | <b><math>y = (0.76^2 \times z)/0.03</math></b>                             |
| 9          | 1.68       | $y = 0.77^z \times (z \times 29.96)^{0.99}$                                |
| 10         | 1.68       | $y = z^3 + z \times (29.8 - z \times 8.24)$                                |
| 13         | 1.68       | $y = (z^3 - z^2 \times 7.60 + z \times 27.1)/0.907$                        |
| 14         | 1.68       | $y = 0.739^z \times (z \times (\sqrt{z} - 0.528) + 30.0)$                  |
| 17         | 1.68       | $y = z \times 30.0 \times 1/(0.196 \times 0.465 + e^{0.266})^z + z^{2.13}$ |
| 19         | 1.68       | $y = (30.0 \times z)/\sqrt{1.95^z} + z^{2.13}$                             |

### 5.5.3 DESI + DESY5

The procedure for reconstructing the ratio  $D_M/r_d$  follows the same methodology as in the analysis presented in the section 5.5.2. However, instead of using the Pantheon+ SNe dataset, we utilize the DESY5 dataset, whose absolute magnitude is consistent with a Hubble constant calibration of  $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ .

The resulting data for DESI + DESY5, is used to derive a new relationship for  $D_M/r_d$ . The best equation obtained from PySR for this dataset is

$$\frac{D_M}{r_d} = 30.1z \exp(-0.262z). \quad (5.15)$$

The reconstruction of  $D_M/r_d$  for DESI + DESY5 is shown in the left panel of Figure 5.5. The reconstruction agrees with  $\Lambda\text{CDM}$  within the  $2\sigma$  CL, as the dashed red line indicates. Additionally, the DESI + DESY5 reconstruction of  $D_M/r_d$  aligns more closely with  $\Lambda\text{CDM}$  than the DESI + Pantheon+ case.

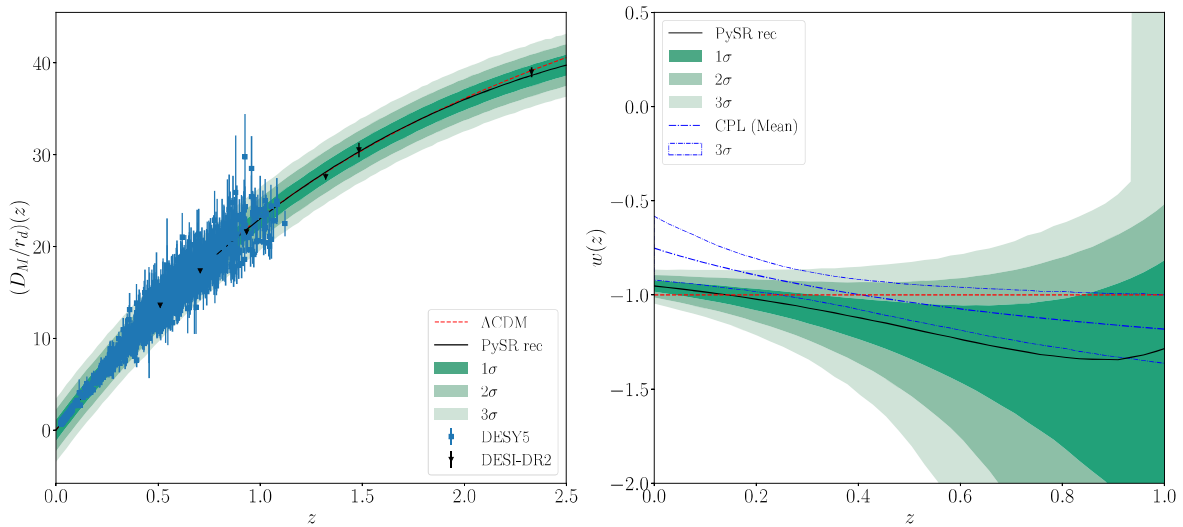


Figure 5.5: The same as Figure 5.4, but using the DESI + DESY5, and a CPL model with the DESI + CMB + DESY5 results KARIM *et al.* (2025).

The resulting  $w(z)$  for the scenario DESI + DESY5 is displayed in the right panel of Figure 5.5, which is in agreement with  $\Lambda\text{CDM}$  within  $2\sigma$ , as indicated by the dashed red line.

The essential point is the structural equivalence of the expressions explained for DESI+Pantheon and DESI+DESY5 in Tables 5.5 and 5.6, respectively, despite their different appearances. In both cases, the reconstructed expressions follow the general form

$$\frac{D_M}{r_d} = a z \exp(bz). \quad (5.16)$$

For DESI+Pantheon, the best-fitting expression is

$$\frac{D_M}{r_d} = \frac{0.76^z z}{0.03} = 33.3 z \exp(-0.274z), \quad (5.17)$$

while for DESI+DESY5 we obtain

$$\frac{D_M}{r_d} = \frac{30.1 z}{1.30^z} = 30.1 z \exp(-0.262z). \quad (5.18)$$

The small differences in the coefficients reflect the use of different SNIa datasets. For this reason, the equations presented in Eqs. (5.14) and (5.15) are not identical to those reported in the corresponding tables.

For the DESI-only analysis, reported in Table 5.4, PySR reconstructs the quantity  $D_H/r_d$  as

$$\frac{D_H}{r_d} = \frac{28.42}{1.68^z} = 28.4 \exp(-0.519z). \quad (5.19)$$

This expression follows the more general form

$$\frac{D_H}{r_d} = a \exp(bz). \quad (5.20)$$

As an instructive exercise, one may ask under which conditions the derivative of Eqs. (5.16) and (5.20) can represent the same quantity. Taking the derivative of Eq. (5.16) with respect to  $z$  and equating it to Eq. (5.20),

$$\frac{d}{dz} \left( \frac{D_M}{r_d} \right) = a e^{bz} (1 + bz) = \frac{D_H}{r_d}, \quad (5.21)$$

one immediately obtains the condition

$$bz = 0. \quad (5.22)$$

Since  $b \neq 0$ , this equality can only be satisfied at  $z = 0$ , implying that the two expressions cannot be simultaneously valid for  $z > 0$ .

This result is nevertheless expected, since  $D_M$  and  $D_H$  were reconstructed independently. In particular, they are related through the geometric consistency relation already introduced in Sec. 2.2 [see Eqs. (2.26) and (2.25)],

$$D_M(z) = \int_0^z D_H(z') dz',$$

or, equivalently,

$$D_H(z) = \frac{dD_M(z)}{dz}. \quad (5.23)$$

was not imposed during the reconstruction. As a consequence, the reconstructed

expressions are not required to satisfy the condition  $bz = 0$ .

In a complementary analysis of DESI + SNIa, we performed a test to explore the dependence of  $w(z)$  on the parameters  $\Omega_m$  and  $r_d$  by varying their values in the calculation of  $\gamma$ . Specifically, we consider  $\Omega_m = 0.25, 0.30$ , and  $0.35$ , as well as  $r_d = 90.9 h^{-1} \text{ Mpc}$ ,  $101 h^{-1} \text{ Mpc}$ , and  $111.1 h^{-1} \text{ Mpc}$  – chosen to lie below and above the fiducial values adopted in this chapter, although still compatible with most of their measurements in the literature.

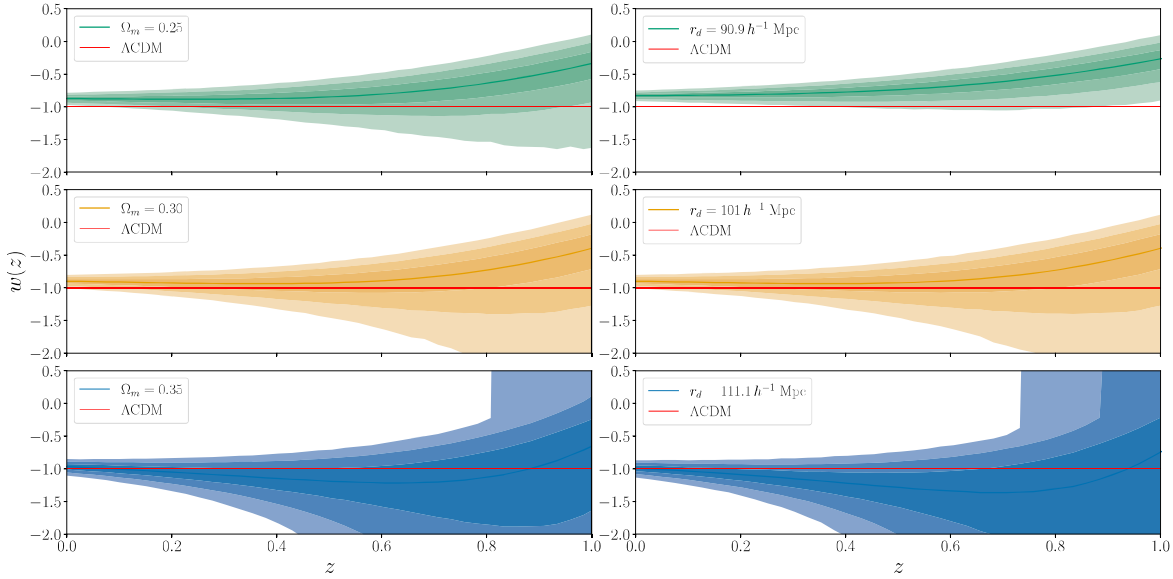


Figure 5.6: Reconstruction ( $3\sigma$ ) of  $w(z)$  for DESI + Pantheon+ considering variations in  $\Omega_m$  and  $r_d$  for different scenarios. **Left:**  $w(z)$  reconstruction with fixed  $r_d$ . **Right:**  $w(z)$  reconstruction with fixed  $\Omega_m$ , for the same scenario.

The results are presented in Figures 5.6 and 5.7. In the upper panel of Figure 5.6, we analyze the scenario of DESI + Pantheon, fixing  $r_d = 101 h^{-1} \text{ Mpc}$  while varying different values of  $\Omega_m$ . In the bottom panel of Figure 5.6, we fix  $\Omega_m = 0.30$  and allow  $r_d$  to vary. Figure 5.7 also contains two panels: the upper panel shows the reconstruction of  $w(z)$  with  $r_d$  fixed for the DESI + DESY5 scenario, allowing  $\Omega_m$  to vary. The bottom panel presents  $w(z)$  with  $\Omega_m$  fixed while varying  $r_d$ .

In all panels, the colored regions indicate the  $3\sigma$  confidence intervals obtained from MCM, and the black dashed line represents the theoretical  $\Lambda\text{CDM}$  curve ( $w = -1$ ). These results demonstrate the sensitivity of  $w(z)$  to the cosmological parameters in Eq. (5.4). It is notable that mild deviations from the  $3\sigma$  (C.L.) occur when the smallest values of  $\Omega_m$  and  $r_d$  are assumed. These lower values represent more extreme choices compared to the higher values, which are better supported by current data, indicating that our results remain robust under reasonable variations in parameters.

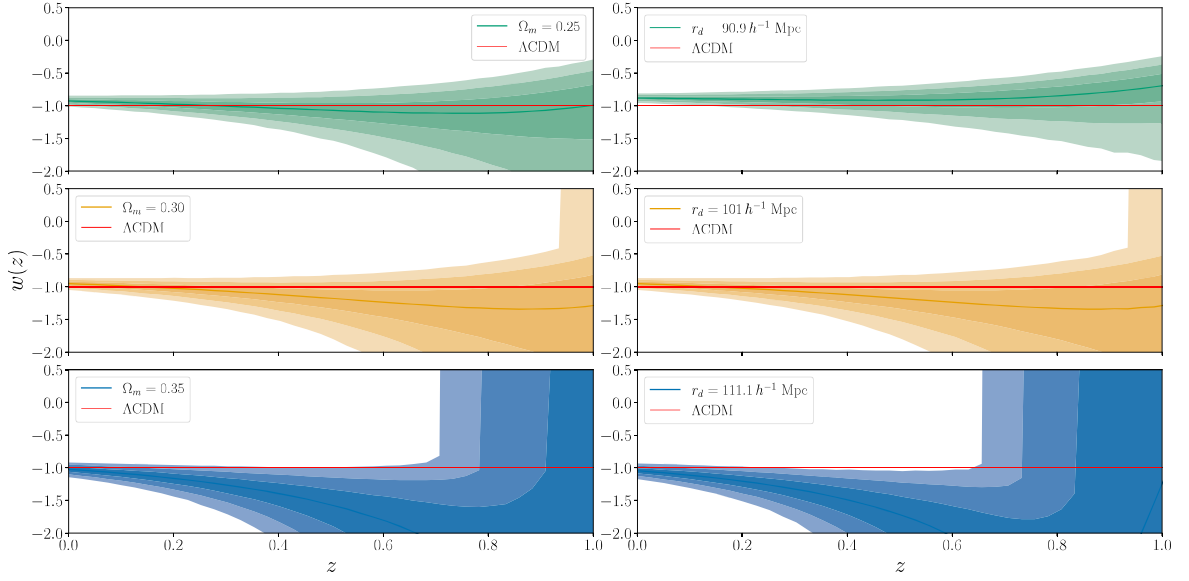


Figure 5.7: Same as Figure 5.6, but using DESI + DESY5.

Table 5.6: DESI + DESY5.

| Complexity | Loss        | Equation                                                            |
|------------|-------------|---------------------------------------------------------------------|
| 1          | 796.        | $y = z$                                                             |
| 2          | 499.        | $y = 8.65$                                                          |
| 3          | 22.5        | $y = 21.5 \times z$                                                 |
| 5          | 4.82        | $y = (z \times 54.8)^{0.77}$                                        |
| 6          | 4.82        | $y = 22.1 \times z^{0.774}$                                         |
| <b>7</b>   | <b>2.23</b> | <b><math>y = (30.1 \times z)/(1.31^z)</math></b>                    |
| 8          | 2.10        | $y = 29.44^z \times z$                                              |
| 9          | 2.05        | $y = z \times (z + 30.1/1.37^z)$                                    |
| 11         | 2.05        | $y = z \times (z + 30.0/1.37^z) - 0.0361$                           |
| 13         | 2.05        | $y = z \times (z^2 + 31.7)/\sqrt{1.88^z} - z$                       |
| 14         | 2.04        | $y = 0.0129 + z \times ((z^2)^{0.923} + 30.5)/1.38^z$               |
| 18         | 2.04        | $y = 31.1 \times z \times \sqrt[4]{z^{0.2 \times z}}/\sqrt{1.85^z}$ |
| 20         | 2.04        | $y = z \times (z + 30^{-z}) - 0.0218/(z - 1.61)$                    |



## 5.6 Gaussian processes

As a complementary analysis, we repeat the reconstruction procedure described above by replacing the reconstruction technique. Instead of using PySR, we adopt Gaussian Processes, allowing for a direct comparison with a method that is more widely used in cosmological analyses. In this approach, a different treatment of BAO data is adopted, in which  $D_H/r_d$  and  $D_M/r_d$  are reconstructed jointly, and their covariance is taken into account. In addition, the first and second derivatives of the reconstructed quantities, together with their associated covariance matrices, are also obtained, with further details described in Chapter 4.

As a first application, this reconstruction is performed using the DESI + Pantheon+ dataset. The results are presented in Fig. 5.8. The left panel shows the reconstruction of  $D_M/r_r$ , where a clear reduction in the uncertainties is observed when compared to the SR results described in Section 5.5.2. The right panel displays the reconstructed equation-of-state parameter  $w(z)$ . For comparison, the blue dashed curves represent the CPL parametrization constrained by DESI + CMB + Pantheon+ data, shown at the  $3\sigma$  confidence level KARIM *et al.* (2025).

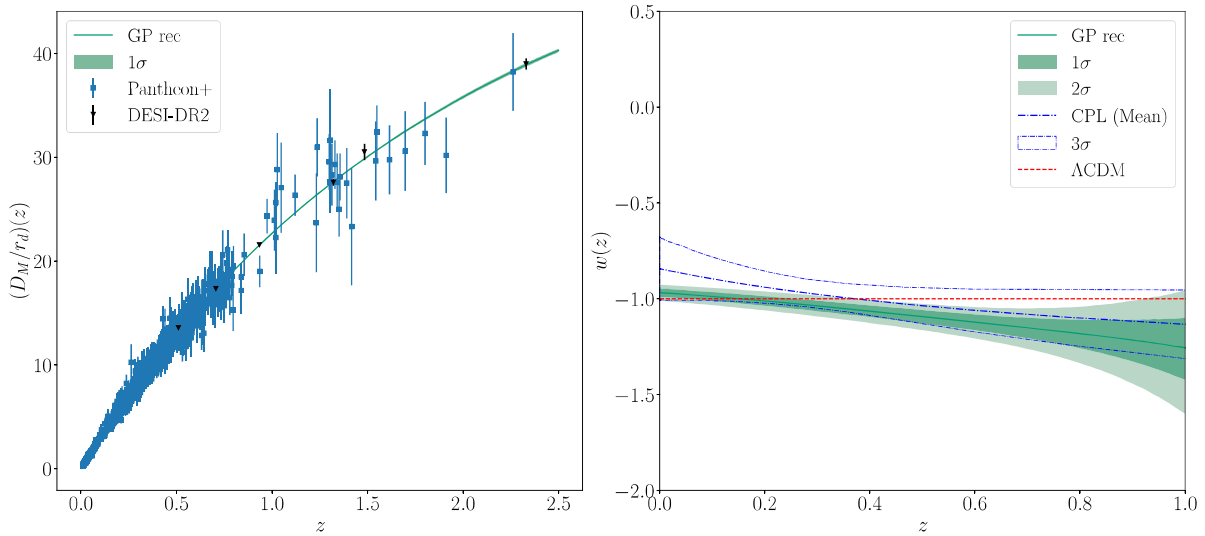


Figure 5.8: The same as Figure 5.4, but using GP.

Following the previous analysis, we apply the reconstruction to the DESI+DESY5 dataset, with the results shown in Figure 5.9. As in the SR results discussed in Subsection 5.5.3, the left panel shows that the reconstruction of  $D_M/r_d$  is primarily driven by the DESY5 measurements at low redshifts, leading to well-constrained uncertainties in the region  $z \lesssim 1$ , while the DESI-DR2 data anchor the reconstruction at higher redshifts. The right panel shows the reconstructed equation-of-state parameter  $w(z)$ , which remains consistent with a cosmological constant within the uncertainties, with increasing uncertainties toward higher redshifts. The reconstruction is in agreement with the CPL

parametrization constrained by DESI + CMB + DESY5 data at the  $3\sigma$  confidence level KARIM *et al.* (2025).

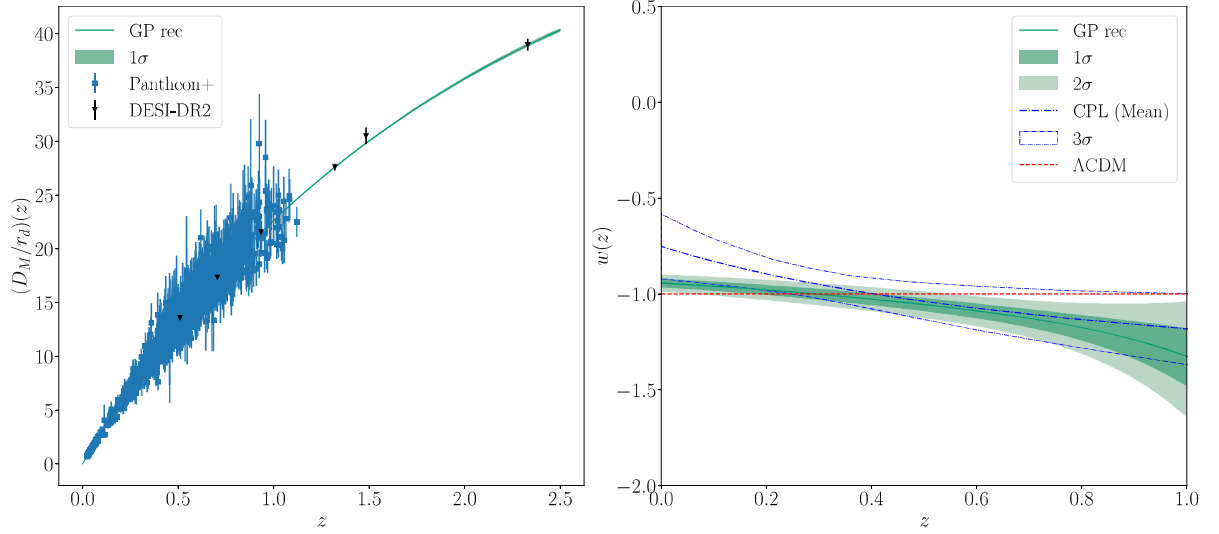


Figure 5.9: The same as Figure 5.4, but using GP and DESI + DESY5.

The results obtained for  $w(z)$  from the GP reconstruction are consistent with the findings of DINDA and MAARTENS (2024), who also reported no statistically significant deviation from the cosmological constant when using GP-based methods. In both analyses, the reconstructed equation of state remains compatible with  $w = -1$  within the  $2\sigma$  confidence level over the redshift range probed by the data

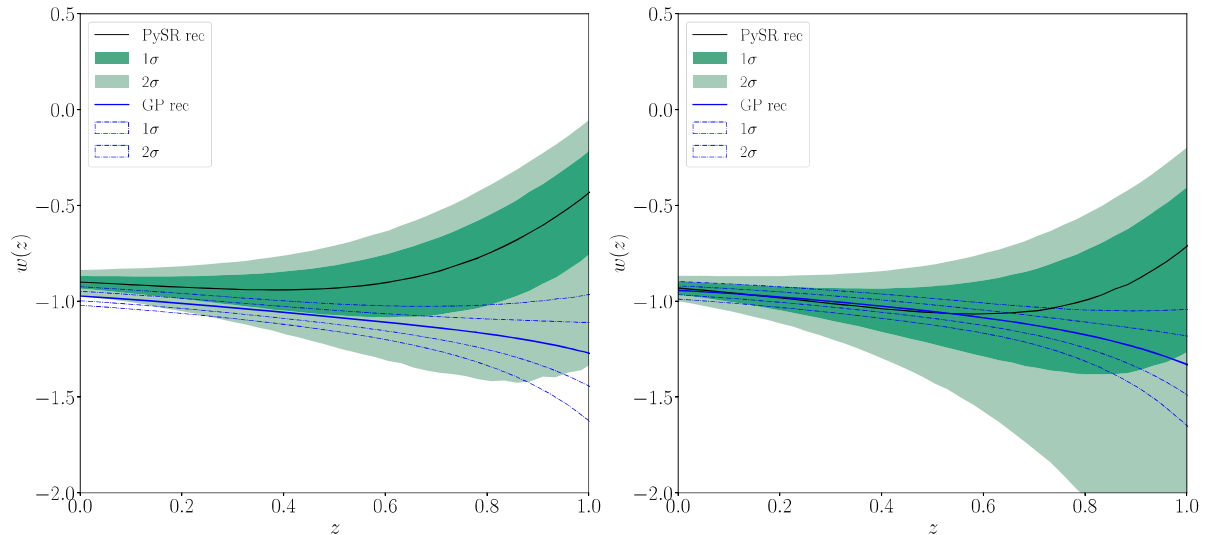


Figure 5.10: Reconstruction of  $w(z)$  using PySR and GP for two combined data sets. The left panel shows the results for DESI + Pantheon+, whereas the right panel shows the results for DESI + DESY5. The shaded regions indicate the  $1\sigma$  and  $2\sigma$  confidence intervals.

A comparison between the GP and SR reconstructions (see Fig. 5.10) shows that the

GP reconstruction is fully contained within the  $2\sigma$  confidence region of the PySR reconstruction over the redshift range up to  $z \simeq 1$ . In particular, for the DESI + DESY5, both methods show good agreement around  $z \leq 0.6$ . This divergence reflects the fundamentally different extrapolation behaviors of the two approaches in data-scarce regions. The GP reconstruction becomes increasingly dominated by the kernel assumptions, leading to smoother mean functions and rapidly expanding uncertainties.

At higher redshifts, where the observational data become increasingly sparse, the divergence between the GP and SR mean reconstructions indicates that the inferred behavior of  $w(z)$  is strongly method-dependent. In this regime, the widening confidence regions of the GP and the model-dependent extrapolation of the SR preclude a robust statistical preference for any specific dark energy behavior. In particular, the current data do not allow us to distinguish between phantom-like ( $w < -1$ ), quintessence-like ( $w > -1$ ), or thawing/freezing scenarios at high redshift.

Therefore, any apparent trend in the reconstructed mean of  $w(z)$  at large  $z$  should be interpreted with caution, as it primarily reflects the extrapolation properties of the reconstruction method rather than a data-driven physical signal.

## 5.7 Conclusion

Given its potential consequences for cosmology, DESI's result showing a preference for a time-dependent dark energy EoS has motivated several analyses using parametric and non-parametric methods. In this chapter, we investigated this issue from a PySR reconstruction of  $w(z)$ . The reconstruction is first performed using the DESI  $D_H/r_d$  measurements and then refined by combining DESI  $D_M/r_d$  with SNe data. Our primary result shows that asserting that they favor any particular cosmological model is premature. Such a conclusion is due to the large uncertainties derived from these data, which limit the precision needed to support a specific model definitively.

The EoS,  $w(z)$ , shows significant sensitivity to the cosmological parameters entering the  $\gamma$  expression – Eq. (5.5). Variations in both  $\Omega_m$  and  $r_d$  lead to noticeable changes in the reconstruction of  $w(z)$ , highlighting the dependence of the equation of state on these key cosmological parameters. However, within the explored range of  $\Omega_m$  and  $r_d$ , the reconstructed EoS remains consistent with  $\Lambda$ CDM, except for the lower values of  $\Omega_m$  and  $r_d$ , which show mild deviations.

The  $w(z)$  for this scenario shows no statistically significant departure of  $\Lambda$ CDM within  $3\sigma$  confidence level; however, the large error bars prevent the exclusion of alternative models.

Concerning the DESI + DESY5, the reconstructed  $D_M/r_d$  values also agree with the  $\Lambda$ CDM model within  $1\sigma$ . Since the DESY5 dataset only extends up to  $z = 1.12$ , the following points in the reconstruction come from DESI-BAO data, which agree well with

$\Lambda$ CDM. This alignment likely caused the reconstruction to fit more closely to  $\Lambda$ CDM than the previous scenario with Pantheon+ data. The  $w(z)$  reconstruction from the DESY5 dataset agrees with  $\Lambda$ CDM within  $2\sigma$ .

Finally, our PySR reconstruction is consistent with the corresponding Gaussian Process results, with the GP mean lying entirely within the  $2\sigma$  confidence region of the PySR reconstruction over the redshift range up to  $z \simeq 1$ . In turn, the GP-based reconstruction is in agreement with previous model-agnostic analyses in the literature, such as the study of DINDA and MAARTENS (2024). While differences in the detailed redshift evolution of  $w(z)$  arise at higher redshifts due to the distinct extrapolation behaviors of the methods, the overall conclusions remain robust across these independent approaches.

# CHAPTER 6

---

## Cosmological constraints on galaxy formation time

---

*This chapter is based on the published article: SOUSA-NETO et al. (2026).*

### 6.1 Introduction

In this chapter, we investigate the  $t_f$  redshift dependence and characterize its behavior throughout cosmic evolution. Here, we specifically focus on a particular class of galaxies, namely, passively evolving galaxies, like early-type galaxies (ETG). These objects exhibit little to no ongoing star formation, are dominated by older stellar populations that evolve slowly, and possess well-characterized colors, spectra, and luminosities. Their age can be calculated by modeling the stellar populations using either simple or composite stellar population models. Although our calculations rely on observational data of ETGs, it is important to acknowledge that the method employed is not entirely model-independent; the formation time can only be inferred from galaxy ages by assuming a cosmological model.

We organize this chapter as follows. In Section 6.2, we present our definition of the formation time. Section 6.3 describes the methodology, including the data selection and reconstruction techniques used to constrain  $t_f(z)$ . Our results are presented in Section 6.4. In Section 6.5, we perform robustness tests of the formation time  $t_f$ . Finally, Section 6.6 summarizes our main conclusions.

### 6.2 Galaxy formation time

In this thesis, we shall adopt the following definition for the galaxy formation time:  $t_f$  represents the cosmic time, measured from  $t = 0$  (corresponding to the beginning of the Universe), at which a given fraction of a galaxy's final stellar mass had formed. This is clearly a simplification that does not introduce significant error in the estimates  $t_f$ . As is well known, the process of galaxy formation commences only after a sufficient cosmic structure has emerged, and the universe has cooled adequately. The gravitational seeds

of galaxy formation are large-scale dark matter halos ( $> 10^8 M_\odot$ ) that typically begin to collapse around 100–300 million years after the Big Bang, roughly corresponding to redshifts  $z \sim 15 - 20$ .

If the age of a galaxy at a given redshift  $z$  is denoted by  $t_g$ , representing the time elapsed since its stellar population was largely assembled (the galaxy’s current age), then  $t_g$  reflects the duration available for stellar evolution and other post-formation processes. The formation time can thus be expressed as

$$t_f = t_u - t_g, \quad (6.1)$$

where  $t_u$  is the age of the Universe at redshift  $z$ . In the literature, related terms such as ‘formation epoch’, ‘delay factor’, or ‘incubation time’ may be used interchangeably with  $t_f$ , depending on the specific modeling or observational context.

It is crucial to recognize that the above equations establish a direct equivalence between  $t_f$  and the formation redshift ( $z_f$ ). Consequently, for any given  $z_f$  and a specified cosmological model, one can easily calculate  $t_f$ . This relationship is particularly important because numerous studies report observational measurements and estimates primarily in terms of formation redshift rather than formation time.

## 6.3 Methodology and Data

One of our primary goals is to demonstrate that it is possible to constrain the galaxy formation time using current cosmological data sets. To achieve this, we aim to develop a model-independent method to reconstruct galaxy ages and, consequently,  $t_f$ .

We use GP and PySR to reconstruct galaxy ages. The GP method has been widely employed in cosmology for various applications, including null tests of the FLRW metric (SEIKEL *et al.*, 2012; SHAFIELOO and CLARKSON, 2010), reconstruction of cosmological matter perturbations (GONZALEZ, 2017; GONZÁLEZ *et al.*, 2016; GONZALEZ *et al.*, 2017), estimates of the Hubble constant BUSTI *et al.* (2014), among others. On the other hand, symbolic regression has recently been applied to investigate a potential dynamical behavior of dark energy, particularly in light of recent BAO data from the DESI collaboration (SOUSA-NETO *et al.*, 2025).

We follow two approaches to determine  $t_f$ . The first one calculates  $t_f$  as the difference between the theoretical curve of the age of the Universe, obtained from Eq. (2.51), and the galaxy age reconstruction via GP and PySR, where the theoretical curve is based on the  $\Lambda$ CDM model for two different values of  $H_0$  reported by the PLANCK COLLABORATION (2020) and SH0ES RIESS *et al.* (2019) collaborations. The second approach uses a reconstruction of the lookback-time redshift relation from SNe data combined with galaxy ages.

### 6.3.1 Gaussian Processes

We use the code GP in Python **GaPP** to derive our GP results SEIKEL *et al.* (2012). As discussed in the Chapter 4, the distribution of functions provided by GP is suitable for describing the observed data. At each redshift  $z$ , the GP reconstruction provides a mean value and an associated uncertainty, assuming a Gaussian distribution. Correlations between different points are encoded in a covariance function  $k(z, \tilde{z})$ , which depends on a small set of hyperparameters, namely the characteristic correlation length  $l$  and the signal variance  $\sigma_f$ . These hyperparameters are optimized directly from the data by maximizing the marginal likelihood within the **GaPP** framework. As a non-parametric approach, GP does not assume a specific functional form for the reconstructed quantity.

| Kernel              | $\log \mathcal{L}$ |
|---------------------|--------------------|
| Squared Exponential | −23.1176           |
| Matern 3/2          | −24.1007           |
| Matern 5/2          | −23.4050           |
| Matern 9/2          | −23.1718           |

Table 6.1: Log-likelihood values for different Gaussian Process kernels.

The different choices of the covariance function may affect the final reconstruction. To assess the impact of this choice on our results, we perform Gaussian Process reconstructions using the squared exponential and Matérn covariance functions with  $\nu = 3/2, 5/2$ , and  $9/2$  (see SEIKEL and CLARKSON (2013)). The corresponding log-likelihood values for each kernel are reported in Table 6.1. We find that all kernels lead to very similar reconstructions, with only marginal differences in the log-likelihood, indicating that the results are robust against the specific choice of covariance function. Among the tested kernels, the squared exponential presents one of the highest log-likelihood values and yields a smooth reconstruction. Therefore, in what follows, we adopt the squared exponential form SEIKEL *et al.* (2012),

$$k(z, \tilde{z}) = \sigma_f^2 \exp \left( -\frac{(z - \tilde{z})^2}{2l^2} \right) . \quad (6.2)$$

### 6.3.2 High-performance Symbolic Regression

As discussed in the previous chapters, one of the key steps in symbolic regression is the choice of the operator set used in the reconstruction. Since this aspect has already been thoroughly addressed, we simply state here that we adopt a set of unary operators —

`{log, sqrt, exp, cube, square, inv}` — and binary operators — `{+, -, *, /, ^}` — as implemented in the PySR library.

Another crucial component of SR is the definition of the loss function, which is a key factor in optimizing the model’s performance by measuring the difference between predicted and input values. We adopt the  $P$ -th power absolute distance loss, defined as

$$\mathcal{L} = |x - y|^P, \quad (6.3)$$

where  $x$  represents the predicted value,  $y$  denotes the true value, and  $P$  is a positive real number that controls the sensitivity of the loss to errors. The methodology used with PySR follows the same procedure as in the previous chapter, where further details, including the choice of the best-fit equation, can be found.

### 6.3.3 Age Data

In this chapter, we use age estimates of 32 old passive galaxies distributed over the redshift interval  $0.11 \leq z \leq 1.84$ , as analyzed by SIMON *et al.* (2005). The total sample is composed by three sub-samples: 10 field early-type galaxies (TREU *et al.*, 1999, 2001, 2002), whose ages were obtained by using the SPEED models of JIMENEZ *et al.* (2004); 20 red galaxies from the publicly released Gemini Deep Deep Survey (GDDS) whose integrated light is fully dominated by evolved stars (ABRAHAM *et al.*, 2004; MCCARTHY *et al.*, 2004). Simon *et al.* re-analyzed the GDDS old sample by using a different stellar population model and obtained ages within 0.1 Gyr of the GDDS collaboration estimates and the two radio galaxies LBDS 53W091 and LBDS 53W069 (DUNLOP *et al.*, 1996; SPINRAD *et al.*, 1997). The age data set used in this analysis is shown in Fig. 6.1.

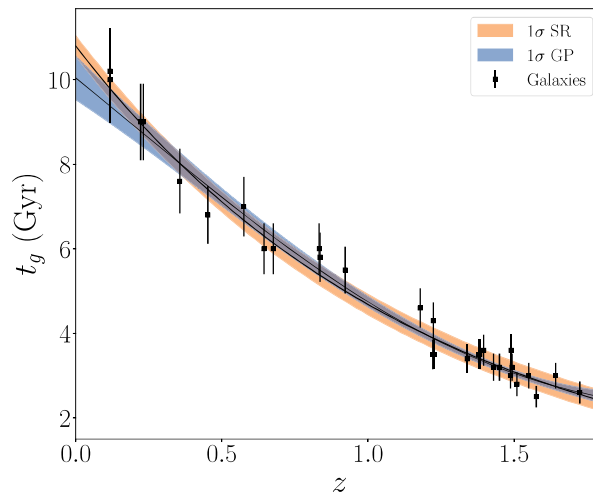


Figure 6.1: GP and SR reconstructions of the age-redshift relation from age estimates of 32 old passive galaxies (black points) SIMON *et al.* (2005). The contours show the reconstruction at  $1\sigma$  level while the solid black line represents the mean of the reconstructions.



### 6.3.4 Type Ia Supernova Data

We use the Pantheon+ catalog BROUT and ET AL. (2022) for the SNe observations, which comprises 1550 SNe within the redshift range  $0.001 \leq z \leq 2.26$ . From the definition of the distance modulus relation, we obtain the luminosity distance by Eq. (2.30). We assume two different values for this latter quantity:  $-19.42$ , corresponding to the  $H_0$  value of the PLANCK COLLABORATION (2020) and  $-19.253$ , corresponding to the measurements of  $H_0$  reported by the SH0ES Collaboration (RIESS *et al.*, 2019).

## 6.4 Results

From Eq. (6.1),  $t_f$  can be estimated as the difference between the theoretical curve of the age of the Universe and the ages of galaxies. Fig. 6.1 shows the GP and SR reconstruction of the ages of 32 galaxies as a function of redshift. The blue contours indicate  $1\sigma$  confidence level (C.L.) while the solid black line represents the mean of the reconstruction. The black points correspond to the original data from SIMON *et al.* (2005) with their respective uncertainties. The total age of Universe is obtained from Eq. (2.51) assuming a flat,  $\Lambda$ CDM model with the following parameters: (i)  $H_0 = 73.04 \pm 1.04 \text{ km s}^{-1} \text{ Mpc}^{-1}$  and  $\Omega_m = 0.33$ , corresponding to the values from the SH0ES Collaboration – see the upper curve in left panel of Fig. 6.2; (ii)  $H_0 = 67.4 \pm 0.5 \text{ km s}^{-1} \text{ Mpc}^{-1}$  and  $\Omega_m = 0.315$ , corresponding to Planck’s best-fit scenario – see the upper curve in the middle panel of Fig. 6.2.

The right panel of Fig. 6.2 shows the evolution of  $t_f$  with redshift. In that panel, the upper light red curve corresponds to the Planck’s  $H_0$  and  $\Omega_m$  values, while the lower blue curve corresponds to the SH0ES values for the same parameters. These curves represent  $1\sigma$  C.L. intervals for the formation time and show noticeable variations at lower redshifts ( $z \lesssim 0.5$ ).

When we replace the GP reconstruction method with the SR approach, we follow the same procedure as before. The best-fitting equation for the galaxy age,  $t_g$ , is

$$t_g = 2.31^{2.85-z}. \quad (6.4)$$

The full set of symbolic regression equations, along with their corresponding complexities and losses, can be found in Table 6.2. The reconstruction results, as well as the obtained  $t_f$  curve, are shown in Fig. 6.3. Although the uncertainty estimates provided by PySR are less precise than those obtained with the GP method, the overall reconstruction behavior remains consistent. Evaluating Eq. (6.4) at  $z = 0$  yields a galaxy age of  $t_g = 10.87 \text{ Gyr}$ , which is larger than the value obtained with the GP reconstruction,  $t_g = 10.03 \text{ Gyr}$ . This difference in  $t_g$  at  $z = 0$  directly impacts the inferred formation time, leading to smaller

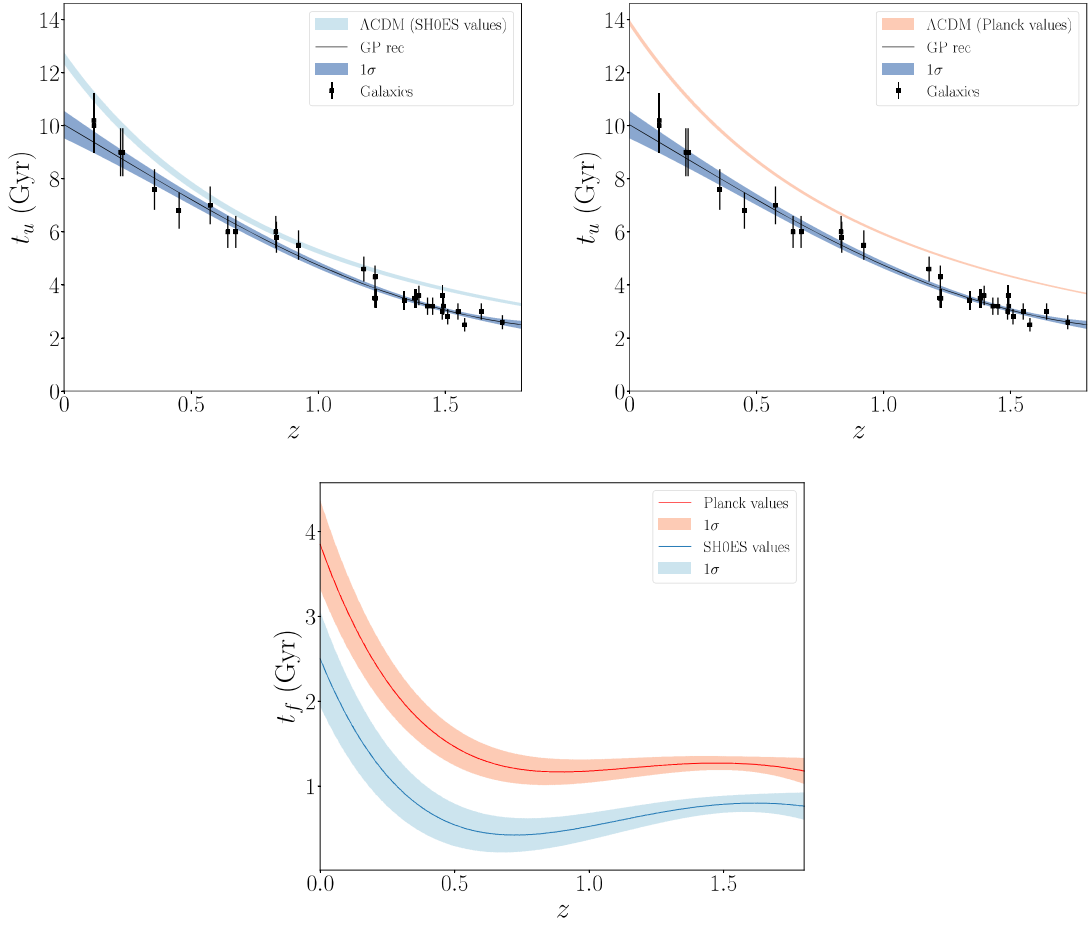


Figure 6.2: **Upper left:** The age-redshift relation assuming a flat,  $\Lambda$ CDM model and the values of  $H_0$  and  $\Omega_m$  provided by the SH0ES collaboration. **Upper right:** The same as in the previous panel for the values of  $H_0$  and  $\Omega_m$  provided by the Planck collaboration. In both panels, the dark blue region corresponds to the reconstructed age-redshift relation shown in Fig. 6.1. **Lower panel:** Evolution of the galaxy formation time  $t_f(z)$  with redshift obtained from the difference between  $t_u$  and  $t_g$  – Eq. (6.1). The red and blue regions represent the reconstructed  $1\sigma$  intervals.

values of  $t_f$  at  $z = 0$  in the SR case compared to the GP case.

For completeness, we also perform the above analysis assuming a  $w_0 w_a$ CDM model characterized by a CPL parameterization for the dark energy equation-of-state,  $w(z) = w_0 + w_a z / (1 + z)$ , with  $w_0 = -0.838 \pm 0.055$  and  $w_a = -0.62^{+0.22}_{-0.19}$ , corresponding to de DESI+Pantheon+as reported by the DESI collaboration ABDUL KARIM *et al.* (2025).

Fig. 6.4 presents the cosmic age  $t_u(z)$  with  $1\sigma$  uncertainties for the  $\Lambda$ CDM and CPL dark-energy parametrizations. As shown in the upper panel, both models yield nearly indistinguishable age-redshift relations. The lower panel displays the fractional difference  $\Delta t(z)/t_{\Lambda\text{CDM}}(z)$ . The small relative difference between the CPL and  $\Lambda$ CDM predictions indicates that variations in the dark-energy parametrization have a negligible impact on  $t_u(z)$  when compared to changes in the underlying cosmological parameters.

Rather than the specific form of the dark-energy parametrization, the uncertainty

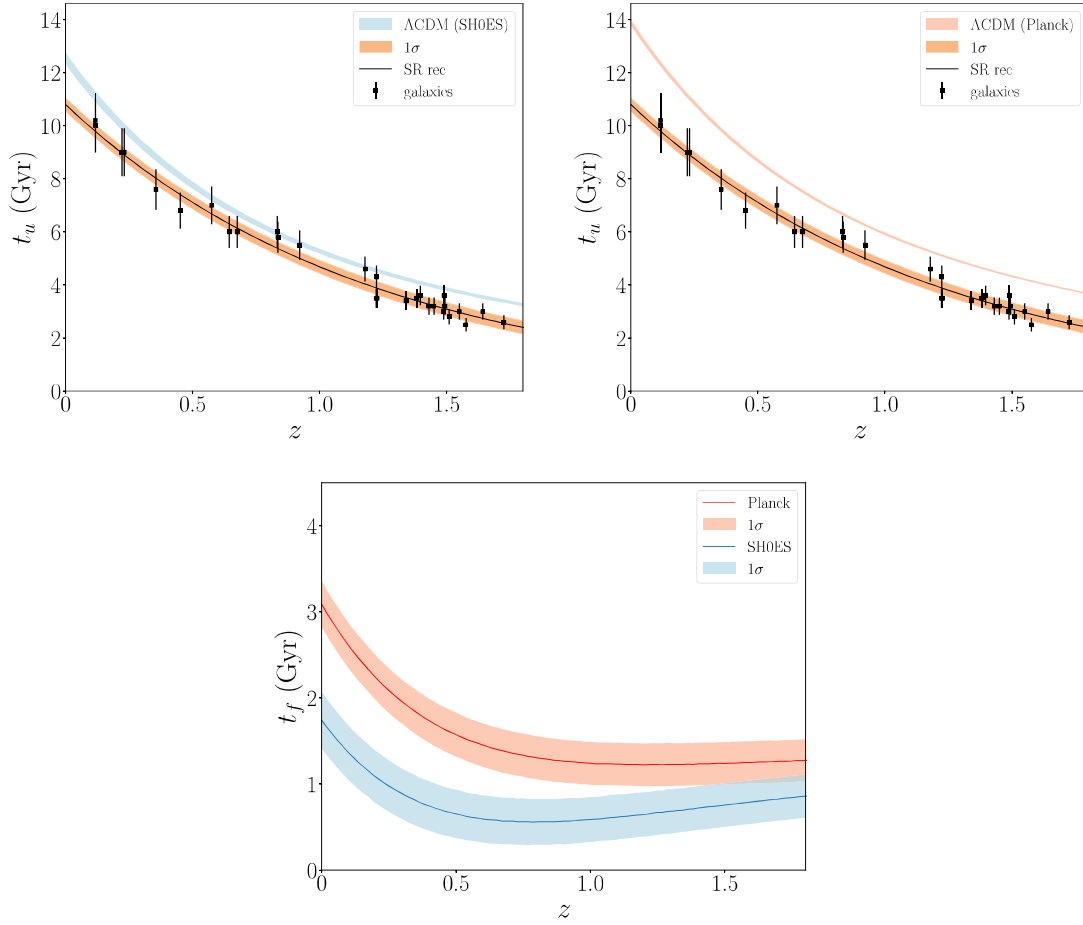


Figure 6.3: Same as Fig. 6.2, but using PySR to reconstruct galaxy ages.

in the cosmic age  $t_u(z)$  is primarily determined by the uncertainties of the underlying cosmological parameters. We compute the partial derivatives of  $t_u$  with respect to each model parameter in the  $\Lambda$ CDM model, listed below, to assess their individual contributions to the total uncertainty.

$$\begin{aligned}
 \frac{\delta t_u}{t_u(z)} &= -\frac{\delta H_0}{H_0}, \\
 \frac{\delta t_u}{\delta \Omega_M} &= -\frac{1}{2H_0} \int_{1+z}^{\infty} \frac{x^3 - 1}{x \mathcal{H}(x)^3} dx, \\
 \frac{\delta t_u}{\delta \Omega_k} &= \frac{1}{2H_0} \int_{1+z}^{\infty} \frac{x^2 - 1}{x \mathcal{H}(x)^3} dx.
 \end{aligned} \tag{6.5}$$

Here we have defined  $x \equiv 1 + z$ . The total uncertainty in the age–redshift relation can be characterized by the individual contributions from each cosmological parameter and written as

$$\frac{\delta t_u}{t_u(z)} = \left\{ \frac{\delta H_0}{H_0}, \alpha_M \delta \Omega_M, \alpha_k \delta \Omega_k \right\}, \tag{6.6}$$

| Complexity | Loss   | Equation                                                                |
|------------|--------|-------------------------------------------------------------------------|
| 1          | 1.85   | $y = 3.23$                                                              |
| 2          | 1.75   | $y = e^{1.27}$                                                          |
| 3          | 1.02   | $y = 4.94 - z$                                                          |
| 4          | 0.126  | $y = 10.5 (1 - 0.309z)^2$                                               |
| 5          | 0.0805 | $\mathbf{y} = \mathbf{2.31}^{2.85-z}$                                   |
| 6          | 0.0805 | $y = e^{0.835(2.85-z)}$                                                 |
| 8          | 0.0782 | $y = z + 5.93 (1 - 0.411z)^2 \times 1.79$                               |
| 10         | 0.0779 | $y = 2.95 \left( \frac{0.582}{z^{0.119}} - 0.582z + 1 \right)^2 + 1.71$ |
| 11         | 0.0778 | $y = 10.1 (1 - 0.315z)^2 +  0.915 - z ^3$                               |
| 12         | 0.0696 | $y = (z^3 - 0.591)^{-0.1008} + e^{2.27-z}$                              |
| 13         | 0.0692 | $y = e^{2.34-z} + 0.755 + \frac{0.0422}{z^2 - 0.637}$                   |
| 14         | 0.0689 | $y = e^{2.27-z} + (z^3 - 0.593)^{-0.1008/z}$                            |
| 15         | 0.0671 | $y = 0.795(z^3 - 0.524)^9 + e^{2.11-z} + 1.22$                          |
| 16         | 0.0623 | $y = (2.17 - z)^{z^{18}} + 14.4e^{-1.26\sqrt{z}}$                       |
| 17         | 0.0611 | $y = [10.2 (1 - 0.461z)^3]^{z^{12}} + 14.3e^{-1.26\sqrt{z}}$            |
| 18         | 0.0605 | $y = (2.18 - z)^{4.13z^{12}} + 14.3e^{-1.25\sqrt{z}}$                   |
| 20         | 0.0600 | $y = [10.3 (1 - 0.459z)^3]^{(z^4)^{e^z}} + 14.3e^{-1.25\sqrt{z}}$       |

Table 6.2: Symbolic regression results for  $t_g$ .

where

$$\begin{aligned}
\alpha_M &= -\frac{1}{2\alpha_0} \int_{1+z}^{\infty} \frac{x^3 - 1}{x \mathcal{H}(x)^3} dx, \\
\alpha_k &= \frac{1}{2\alpha_0} \int_{1+z}^{\infty} \frac{x^2 - 1}{x \mathcal{H}(x)^3} dx. \\
\alpha_0 &= \int_{1+z}^{\infty} \frac{dx}{x \mathcal{H}(x)}.
\end{aligned} \tag{6.7}$$

For clarity, we do not combine the parameter uncertainties in quadrature. Such a procedure would require a specific choice for the uncertainty in  $H_0$  and a proper treatment of correlations among the cosmological parameters. Instead, we present the individual contributions separately to highlight their relative importance. As a practical example, we adopt PLANCK COLLABORATION (2020) (TT,TE,EE+lowE+lensing) best-fit values

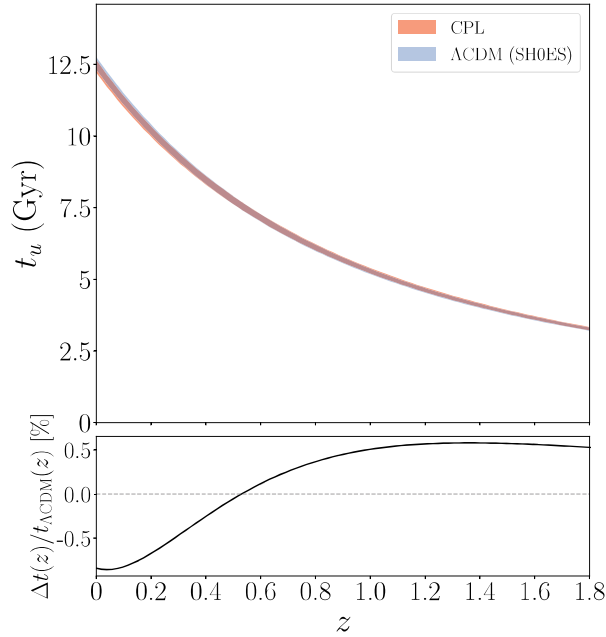


Figure 6.4: Cosmic age as a function of redshift for the CPL model and  $\Lambda$ CDM (SH0ES). **Top panel:**  $t_u(z)$  with  $1\sigma$  uncertainties. **Bottom panel:** relative difference  $\Delta t(z)/t_{\Lambda\text{CDM}}(z)$  in percentage.

for the cosmological parameters, assuming  $H_0 = 67.36 \pm 0.54 \text{ km s}^{-1} \text{ Mpc}^{-1}$ ,  $\Omega_M = 0.315 \pm 0.007$ ,  $\Omega_k = 0.0007 \pm 0.0019$ <sup>1</sup>.

The fractional contributions of the cosmological parameters to the uncertainty in the cosmic age–redshift relation are displayed in Fig. 6.5. The uncertainty associated with the Hubble constant dominates at all redshifts, yielding a constant fractional contribution of 5%, while the contributions from  $\Omega_M$  and  $\Omega_k$  remain below the percent level over the entire range.

From the analysis above, we have identified the relative importance of the cosmological parameters in determining the uncertainty of the cosmic age  $t_u(z)$ . However, our primary quantity of interest is the formation time  $t_f(z)$ , defined in Eq. 6.1, which depends not only on  $t_u(z)$  but also on the galaxy age  $t_g(z)$  reconstructed via GP.

When propagating the uncertainties to  $t_f(z)$ , we find that the total error budget is dominated by the uncertainty associated with the Hubble constant. Adopting the PLANCK COLLABORATION (2020) constraints on  $H_0$ , the contribution from  $t_u(z)$  exceeds that from the GP reconstruction of  $t_g(z)$  over most of the explored redshift range. This behavior reflects the strong sensitivity of the cosmic age to  $H_0$ , while the contribution from the GP reconstruction becomes subdominant once a realistic Planck-level uncertainty is assumed. Consequently, the uncertainty in  $t_f(z)$  is primarily driven by the error in  $H_0$ , and only weakly dependent on the specific dark-energy parametrization adopted for  $t_u(z)$ .

<sup>1</sup>here we adopt (TT,TE,EE+lowE+lensing+BAO)

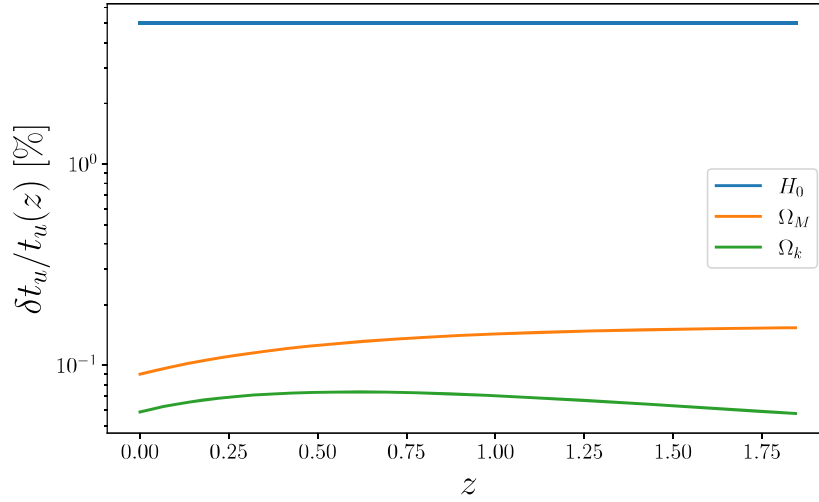


Figure 6.5: Contributions to the uncertainty in the age of the Universe at redshift  $z$  from the cosmological parameters  $H_0$ ,  $w$ ,  $\Omega_M$  and  $\Omega_k$ .

The second approach uses measurements of the SN luminosity distance to reconstruct the lookback time-redshift relation  $t_L(z)$ , which is defined as the difference between the total age of the Universe,  $t_0$ , and its age at a given redshift  $z$ , i.e.,  $t_u$ . As we have seen in the Eq. (2.52) the integral form,  $t_L(z)$  can be expressed as

$$t_L(z) = H_0^{-1} \int_0^z \frac{dz'}{(1+z')\mathcal{H}(z', \mathbf{p})} ,$$

which can be integrated by parts, yielding

$$t_L(z) = \frac{1}{1+z} \left[ \int_0^z \frac{dz'}{H(z')} \right] + \int_0^z \frac{dz'}{(1+z')^2} \left[ \int_0^{z'} \frac{dz''}{H(z'')} \right] . \quad (6.8)$$

Using the definition of luminosity distance given by Eq. (2.32), we obtain

$$t_L(z) = \frac{1}{c} \frac{d_L(z)}{(1+z)^2} + \frac{1}{c} \int_0^z \frac{d_L(z') dz'}{(1+z')^3} . \quad (6.9)$$

An important feature of Eq. (6.9) is that it allows the lookback time to be expressed directly in terms of the luminosity distance, thereby eliminating its explicit dependence on a specific cosmological model through  $H(z)$ . This reformulation removes the need to assume a particular expansion history when reconstructing  $t_L(z)$  from observational data. Nevertheless, a residual dependence on  $H_0$  remains due to the calibration of Type Ia supernovae, which is tied to the choice of the absolute magnitude. This property highlights the main advantage of the new expression: it enables a largely model-independent determination of the lookback time, with minimal assumptions about the underlying cosmology.

From Eq. (2.30), we perform a GP reconstruction of  $d_L(z)$  assuming  $M_B = -19.42$

and  $M_B = -19.253$ , which correspond to the values of  $H_0$  reported by the Planck and SH0ES collaborations, respectively VON MARTTENS *et al.* (2025) – see left panel of Fig. 6.6. Along with the values of  $\Omega_m$  presented earlier, we also calculate the total age of the Universe as  $t_0 = 13.8$  Gyr (Planck) and  $t_0 = 12.43$  Gyr (SH0ES) at  $1\sigma$  level. Using the

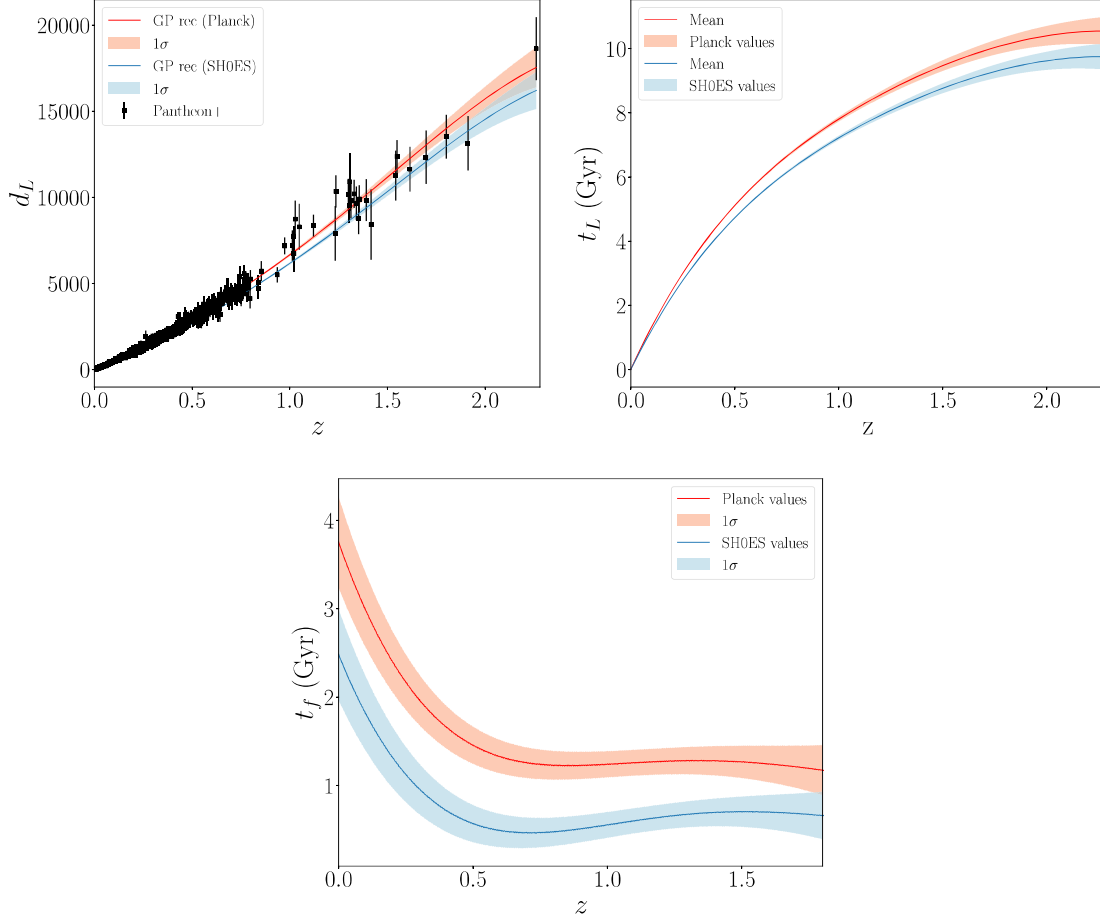


Figure 6.6: **Upper left:** GP reconstruction of the luminosity distance as a function of redshift using the SH0ES and Planck values discussed in the text. **Upper right:** The lookback-time redshift relation reconstructed from  $d_L(z)$  curves, as expressed by Eq. (6.9). **Lower panel:** Evolution of the galaxy formation time  $t_f(z)$  with redshift obtained from Eq. (6.10). In all panels, the blue and red regions represent the reconstructed  $1\sigma$  intervals assuming SH0ES and Planck parameter values, respectively.

definitions above, the formation time can be expressed in terms of the observed lookback time as DANTAS *et al.* (2007a); PIRES *et al.* (2006); SANTOS *et al.* (2007)

$$t_f = (t_0 - t_{L,R}) - t_{g,R} , \quad (6.10)$$

where  $t_0$  is the age of the universe at  $z = 0$  and  $t_{L,R}$  and  $t_{g,R}$  are the reconstructed values of the lookback time and galaxy age, respectively. The right panel of Fig. 6.6 shows the result of this analysis, where the upper light red contour corresponds to the Planck values ( $t_0 = 13.8$  Gyr) and the lower blue contour corresponds to the SH0ES

values ( $t_0 = 12.43$  Gyr). Similarly to the results obtained using the first approach, we also find that  $t_f$  varies significantly at lower redshifts. It is worth emphasizing that other independent determinations of the cosmic age, such as those obtained by VALCIN *et al.* (2025), can be straightforwardly incorporated within this framework. Moreover, the methodology is not restricted to luminosity-distance data and can be consistently extended to other cosmological observables.

We also perform the same analysis by replacing the GP reconstruction of  $t_L$  with the SR method. Although not shown in Fig. 6.6, both methods agree with each other and provide the same qualitative behavior for  $t_f$ . However, the SR reconstruction produces somewhat broader uncertainty bands, suggesting a more conservative estimate of the propagated errors relative to the GP method.

Based on Eq. (6.10), we propagate the uncertainties of the cosmic age  $t_0$ , the lookback time  $t_L$ , and the galaxy age  $t_g$  to determine how each term contributes to the final uncertainty of the formation time  $t_f$ . In this analysis, we assume that the uncertainties associated with  $t_0$ ,  $t_L$ , and  $t_g$  are statistically independent, such that

$$f_i(z) \equiv \frac{\text{Var}(t_i(z))}{\text{Var}(t_f(z))}, \quad i \in \{t_0, t_L, t_g\}. \quad (6.11)$$

Figure 6.7 shows the relative contributions of the cosmic age  $t_0$ , the lookback time  $t_L$ , and the galaxy age  $t_g$  to the total variance of the formation time  $t_f$ . At low  $z$ , the uncertainty budget is clearly dominated by the galaxy age term, indicating that the reconstruction of  $t_g$  is the main source of uncertainty in  $t_f$ . At intermediate redshifts ( $z \sim 1$ ), the contributions from  $t_0$  and  $t_g$  become comparable, while at higher redshifts, the uncertainty is increasingly dominated by the lookback time. For the second approach, the uncertainty budget of the  $t_f$  reconstruction exhibits a qualitatively similar behavior, highlighting the critical role of galaxy age measurements in constraining formation times at low redshift.

To better assess the consistency of the estimated values of  $t_f$  among the different approaches, we evaluate  $t_f$  at three redshifts, namely  $z = 0, 0.5$ , and  $1.5$ . The corresponding values are reported in Table 6.3. We find that the results obtained with the GP-based implementations are in good agreement at all redshifts. At  $z = 0$ , the PySR reconstruction shows a mild but noticeable deviation with respect to the GP-based results, corresponding to a discrepancy at the  $\sim 1$ – $1.3\sigma$ , depending on the GP approach considered. This behavior is not observed at higher redshifts, where all methods show full agreement. A more extensive consistency analysis, based on a larger redshift sample and a quantitative comparison, is presented in Section 6.5.

Finally, it should be mentioned that our results show significant differences (within  $\sim 3\sigma$ ) in the formation time estimates obtained from SH0ES and Planck values, which is a direct consequence of the well-known tension in the measurements  $H_0$  of both col-



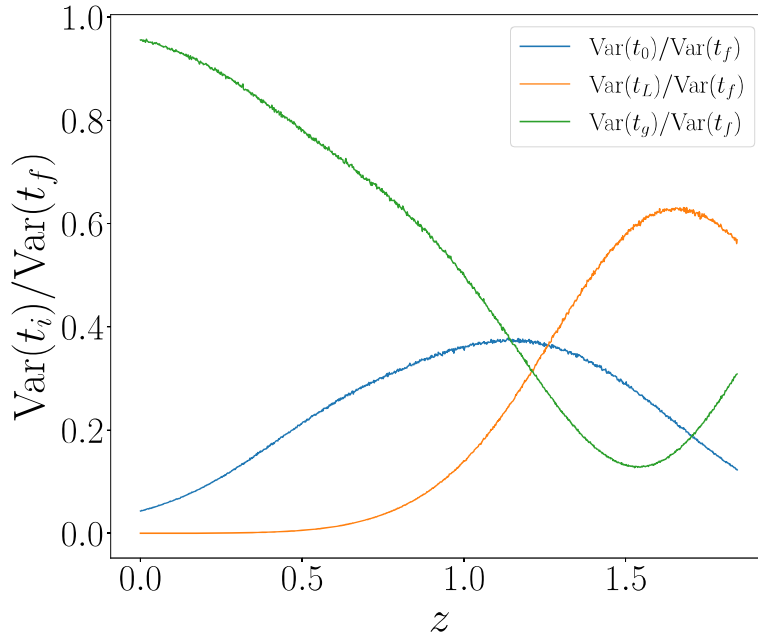


Figure 6.7: Relative contributions of the cosmic age  $t_0$ , lookback time  $t_L$ , and galaxy age  $t_g$  to the total variance of the formation time  $t_f$  as a function of redshift.

laborations (see e.g. FREEDMAN (2025) for a recent discussion). In particular, the larger (smaller) the value of the Hubble constant, the smaller (larger) the total age of the universe and the estimated galaxy formation time.

Table 6.3: Values of  $t_f$  at three redshifts obtained from the different methods discussed in the text.

| $z$ | SR – Approach 1 |                 | GP – Approach 1 |                 | GP–Approach 2   |                 |
|-----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|     | Planck          | SH0ES           | Planck          | SH0ES           | Planck          | SH0ES           |
| 0.0 | $3.09 \pm 0.27$ | $1.74 \pm 0.33$ | $3.75 \pm 0.53$ | $2.49 \pm 0.52$ | $3.85 \pm 0.53$ | $2.50 \pm 0.56$ |
| 0.5 | $1.58 \pm 0.25$ | $0.65 \pm 0.28$ | $1.46 \pm 0.24$ | $0.57 \pm 0.21$ | $1.46 \pm 0.22$ | $0.54 \pm 0.25$ |
| 1.5 | $1.24 \pm 0.25$ | $0.76 \pm 0.25$ | $1.27 \pm 0.21$ | $0.71 \pm 0.17$ | $1.27 \pm 0.08$ | $0.79 \pm 0.10$ |

## 6.5 Robustness tests of the formation time $t_f$

To assess the robustness of our results against methodological choices, we perform a series of consistency tests comparing different reconstruction approaches and techniques. In particular, we analyze the differences in the reconstructed formation time  $t_f$  obtained using GP under two distinct approaches, as well as the comparison between GP and PySR

reconstructions.

We randomly select ten redshift points and compare the corresponding estimates of the formation time  $t_f$  evaluated at the same redshift values for each approach. This extended sample complements the three representative redshifts reported in Table 6.3. We note that none of the redshift points used in Table 1 are included in this random selection. Here, we recall that Approach 1 corresponds to the determination of  $t_f$  from Eq. (6.1), while Approach 2 is based on Eq. (6.10). For each selected redshift, we compute the difference

$$\Delta t_f = t_f^{(i)} - t_f^{(j)}, \quad (6.12)$$

where  $\Delta t_f = 0$  corresponds to perfect agreement between the compared estimates.

Figure 6.8 illustrates these consistency tests. The left panel shows the comparison between GP reconstructions obtained using Approach 1 and Approach 2, while the right panel compares GP and PySR reconstructions within Approach 1. In both cases, the differences are consistent with zero within the uncertainties, reinforcing our result that the inferred formation time  $t_f$  is robust against both the choice of approach and the reconstruction technique.

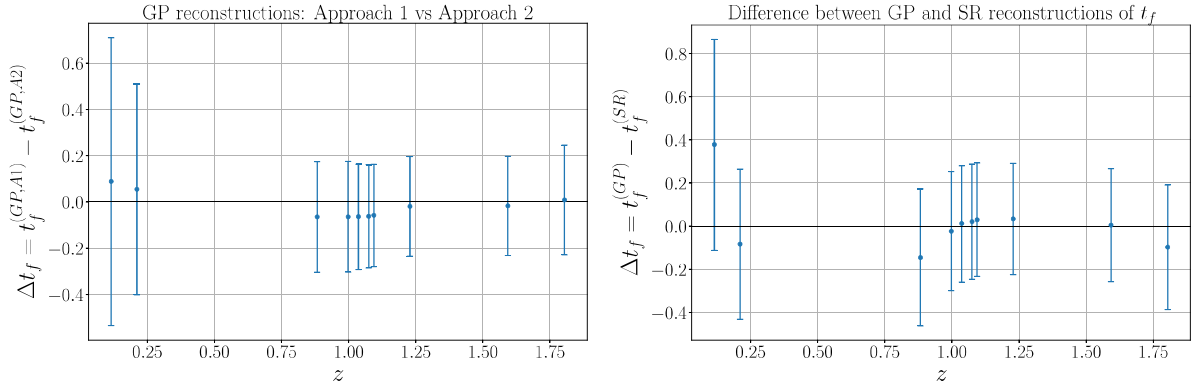


Figure 6.8: Comparison of the reconstructed formation time  $t_f$  obtained with different approaches and techniques. Left panel: difference between GP-based reconstructions using Approach 1 and Approach 2. Right panel: difference between GP and PySR reconstructions for the Approach 1. Error bars represent the combined  $1\sigma$  uncertainties, and the horizontal line denotes  $\Delta t_f = 0$ .

To further quantify the level of agreement between different reconstructions, we compute a tension metric, defined as

$$T_{t_f} = \frac{|t_{f,A1} - t_{f,A2}|}{\sqrt{\sigma_{t_f,A1}^2 + \sigma_{t_f,A2}^2}} \quad (6.13)$$

which measures the discrepancy between two estimates in units of their combined  $1\sigma$  uncertainty. Values of  $T_{t_f} \lesssim 1$  indicate statistical consistency between the reconstructions.

Applying this metric to the selected redshift sample, we find that the level of agreement

between the different reconstructions is very high. For the comparison between GP and PySR within Approach 1, we obtain a maximum value  $T_{t_f}^{\max} = 0.77$ , with the largest deviation occurring at the lowest redshift in the sample. Similarly, when comparing GP reconstructions obtained using Approach 1 and Approach 2, we find  $T_{t_f}^{\max} = 0.28$ . In both cases, all values remain well below unity, indicating strong statistical consistency between the different approaches and reconstruction techniques.

All consistency tests are performed within the same cosmological scenario (Planck or SH0ES). Cross-comparisons between these scenarios are not considered as their well-known tension leads to large discrepancies. For example, a GP-based comparison between Planck and SH0ES reconstructions obtained with different techniques yields  $T_{t_f}^{\max} \simeq 3.7$ , confirming their incompatibility.

## 6.6 Conclusion

In this chapter, we employed two approaches to investigate the evolution of  $t_f(z)$  from age estimates of 32 high- $z$  galaxies and the Pantheon+ type Ia supernova data. To reconstruct the redshift evolution of galaxy ages and the SN luminosity distance, we used the GP and SR reconstruction techniques.

The first approach is based on Eq. (6.1), assuming a  $\Lambda$ CDM background, and allows the determination of  $t_f(z)$  for different values of the Hubble constant reported by the Planck and SH0ES collaborations. We find that the inferred values of  $t_f(z)$  obtained with GP are in good agreement across redshifts and remain consistent with the PySR reconstructions. Although a mild discrepancy is observed at very low redshift in the PySR case, the overall redshift dependence of  $t_f(z)$  is compatible between the two reconstruction techniques within the uncertainties.

Additionally, from a model-independent approach, we estimate  $t_f(z)$  from the reconstruction of the lookback-time–redshift relation using supernova data, as described in Eq. (6.10), fixing only the present age of the Universe. Despite relying on different assumptions, the results obtained with this approach are consistent with those from the first method and confirm that the formation time exhibits a marked redshift dependence, particularly at low redshift ( $z \lesssim 0.5$ ).

# CHAPTER 7

---

## Conclusions

---

In this thesis, we applied non-parametric reconstruction techniques to cosmological observables, with particular emphasis on SR as a supervised machine learning approach. By allowing the data to determine the functional form of the reconstructed quantities, this method provides a flexible and largely model-independent framework for cosmological analysis. Our results show that SR is capable of recovering key cosmological observables directly from observational data while maintaining physical interpretability.

Regarding the results presented in Chapter 5, we performed an SR reconstruction of the dark energy EoS parameter  $w(z)$  using recent DESI measurements. The reconstruction was first carried out using  $D_H/r_d$  data and subsequently refined by combining  $D_M/r_d$  measurements with supernova datasets. This constitutes, to our knowledge, the first application of symbolic regression to directly reconstruct  $w(z)$  from DESI data, adding a new non-parametric tool to the current discussion on the possible time dependence of dark energy.

Our analysis shows that, given the present uncertainties, it is premature to claim evidence in favor of any specific dynamical dark energy model. The reconstructed  $w(z)$  exhibits a strong sensitivity to the cosmological parameters entering the  $\gamma$  expression, particularly  $\Omega_m$  and  $r_d$ , and variations in these parameters lead to noticeable changes in the inferred evolution of the equation of state. Nevertheless, within the explored parameter ranges, the results remain statistically consistent with  $\Lambda$ CDM at the  $3\sigma$  confidence level. The large uncertainties associated with the current DESI data prevent the exclusion of alternative models and limit the statistical significance of possible deviations.

Furthermore, the SR reconstruction is consistent with GP results over the data-dominated redshift range, and both approaches agree with previous model-independent analyses in the literature. Although differences arise at higher redshifts due to distinct extrapolation properties, the overall conclusions remain robust: current data do not provide decisive evidence for departures from  $\Lambda$ CDM.

In Chapter 6, we investigated the cosmic formation time of galaxies using two independent approaches and distinct reconstruction techniques. Both strategies consistently

indicate an evolution in the inferred formation time with redshift.

A key result concerns the comparison between analyses performed adopting values of the absolute magnitude  $M$  consistent with the  $H_0$  determinations from Planck and SH0ES. Since  $M$  and  $H_0$  are intrinsically related through the supernova distance calibration, choosing values of  $M$  corresponding to different  $H_0$  determinations effectively modifies the overall distance scale. This change naturally propagates into the estimation of the lookback time and is consequently reflected in the reconstructed formation time  $t_f(z)$ .

When a specific cosmological model is assumed, the inferred formation time shows only a mild sensitivity to the choice of model itself. Instead, the dominant source of variation arises from the values of the cosmological parameters — particularly  $H_0$ . In fact, the uncertainties in  $t_f(z)$  are largely driven by the uncertainties associated with these parameters.

By applying SR to different observables — including  $D_H/r_d$ ,  $D_M/r_d$ ,  $D_L$ , and galaxy ages — we showed that the method consistently recovers physically meaningful behaviors across independent datasets. To the best of our knowledge, this systematic application of SR to multiple cosmological observables represents a novel contribution, highlighting its potential as a viable framework for cosmological data analysis.

A direct comparison with Gaussian Processes shows that SR is able to recover a qualitatively similar redshift dependence of the reconstructed quantities within the data-dominated regime. When uncertainties are propagated through Monte Carlo sampling, the SR reconstructions typically exhibit broader confidence regions than those obtained with GP, reflecting the intrinsically probabilistic nature of the latter. Nevertheless, the overall functional behavior inferred by both techniques remains consistent within uncertainties. This agreement in the reconstructed trends, despite differences in uncertainty estimation, strengthens the robustness of the main conclusions.

An important outcome of this thesis concerns the derivation of secondary physical quantities, such as the dark energy equation-of-state parameter  $w(z)$  and the galaxy formation time  $t_f(z)$ . Although both SR and GP yield compatible behaviors within uncertainties, differences emerge in poorly sampled redshift regions due to their distinct extrapolation properties. This highlights a crucial point: apparent dynamical features in reconstructed quantities may reflect limitations of the data rather than genuine physical evolution. Careful interpretation is therefore essential, particularly in redshift intervals where the observational coverage is sparse and the available information is limited.

Model selection remains an open and subtle issue. Different functional forms can describe the current data with comparable statistical performance, raising the question of which guiding principle should be adopted. In this thesis, we favored a balance between goodness-of-fit and model simplicity, following the spirit of Occam’s razor. Within this perspective, when multiple models provide statistically equivalent descriptions of the data, preference should be given to the simplest one. Nonetheless, the discussion on op-

timal model selection criteria remains open and may evolve with future, higher-precision datasets.

Finally, one of the main challenges identified throughout this thesis concerns the propagation of uncertainties from reconstructed observables to derived quantities. This limitation is not unique to SR but is common to most machine-learning-based approaches. A rigorous treatment of parameter correlations and covariance structures is essential for improving the reliability of non-parametric cosmological reconstructions in the era of precision cosmology.

A natural next step of this thesis is to improve the treatment of uncertainties in the reconstructed quantities. Although SR provides uncertainty estimates, a more rigorous characterization of parameter correlations is necessary, particularly when deriving secondary quantities such as  $w(z)$  and  $t_f(z)$ . A promising direction is to combine the analytic functional forms obtained via Symbolic Regression with Bayesian sampling techniques, such as Markov Chain Monte Carlo methods (e.g., **emcee** (FOREMAN-MACKEY *et al.*, 2013)). By sampling the parameter space of the reconstructed functions, it becomes possible to estimate parameter correlations explicitly and construct the corresponding covariance matrices, leading to a more robust propagation of uncertainties.

Another relevant extension concerns the use of independent observational datasets to obtain alternative estimates of the formation time. For instance, additional luminosity-distance measurements from different supernova compilations or future surveys may be employed. Through the relation given by Eq. (2.28), one can independently reconstruct the lookback time and derive  $t_f$  from a complementary observable. Such cross-validation would provide an important consistency test of the reconstruction framework and help assess possible systematic effects.

Finally, further developments in symbolic regression techniques may enhance both the stability and interpretability of the reconstructions. Exploring alternative implementations, such as ESR (BARTLETT *et al.*, 2022), may offer improved control over model complexity and uncertainty quantification through the Minimum Description Length (MDL) principle GRÜNWALD (2007), which balances goodness-of-fit and model simplicity. These methodological advances could consolidate symbolic regression as a competitive and reliable tool for cosmological data analysis in the era of high-precision surveys.

---

## Bibliography

---

- ABBOTT, T. M. C., OTHERS, 2024. “The Dark Energy Survey: Cosmology Results With 1500 New High-redshift Type Ia Supernovae Using The Full 5-year Dataset”. Disponível em: <<https://arxiv.org/abs/2401.02929>>.
- ABDUL KARIM, M., OTHERS, 2025, “DESI DR2 results. II. Measurements of baryon acoustic oscillations and cosmological constraints”, *Phys. Rev. D*, v. 112 (Oct), pp. 083515. doi: 10.1103/tr6y-kpc6. Disponível em: <<https://link.aps.org/doi/10.1103/tr6y-kpc6>>.
- ABRAHAM, R. G., GLAZEBROOK, K., MCCARTHY, P. J., et al., 2004, “The Gemini Deep Deep Survey. I. Introduction to the survey, catalogs, and composite spectra”, *Astrophysical Journal*, v. 127, n. 5, pp. 2455.
- ADAME, A. G., OTHERS, 2024a. “DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations”. a. Disponível em: <<https://arxiv.org/abs/2404.03002>>.
- ADAME, A. G., OTHERS, 2024b. “DESI 2024 VII: Cosmological Constraints from the Full-Shape Modeling of Clustering Measurements”. b. Disponível em: <<https://arxiv.org/abs/2411.12022>>.
- AGHAMOUSA, A., OTHERS, 2016. “The DESI Experiment Part I: Science, Targeting, and Survey Design”. Disponível em: <<https://arxiv.org/abs/1611.00036>>.
- AGHANIM, N., OTHERS, 2020, “Planck2018 results: VI. Cosmological parameters”, *Astronomy & Astrophysics*, v. 641 (set.), pp. A6. ISSN: 1432-0746. doi: 10.1051/0004-6361/201833910. Disponível em: <<http://dx.doi.org/10.1051/0004-6361/201833910>>.
- ALBRECHT, A., BERNSTEIN, G., CAHN, R., et al., 2006. “Report of the Dark Energy Task Force”. Disponível em: <<https://arxiv.org/abs/astro-ph/0609591>>.
- ALCANIZ, J. S., LIMA, J. A. S., 1999, “New Limits on  $\Omega_\Lambda$  and  $\Omega_M$  from Old Galaxies at High Redshift”, *Astrophysical Journal Letters*, v. 521, n. 2 (ago.), pp. L87–L90. doi: 10.1086/312191.

- ALCANIZ, J. S., LIMA, J. A. S., CUNHA, J. V., 2003, “Cosmological implications of APM 08279+5255, an old quasar at  $z = 3.91$ ”, *Monthly Notices of the Royal Astronomical Society*, v. 340, n. 4 (abr.), pp. L39–L42. doi: 10.1046/j.1365-8711.2003.06553.x.
- ARJONA, R., NESSERIS, S., 2020, “What can machine learning tell us about the background expansion of the Universe?” *Physical Review D*, v. 101, n. 12 (jun.). ISSN: 2470-0029. doi: 10.1103/physrevd.101.123525. Disponível em: <<http://dx.doi.org/10.1103/PhysRevD.101.123525>>.
- BARBOZA, JR., E. M., ALCANIZ, J. S., 2008, “A parametric model for dark energy”, *Phys. Lett. B*, v. 666, pp. 415–419. doi: 10.1016/j.physletb.2008.08.012.
- BARTLETT, D. J., DESMOND, H., FERREIRA, P. G., 2022, “Exhaustive Symbolic Regression”, *arXiv e-prints*, art. arXiv:2211.11461. doi: 10.48550/arXiv.2211.11461.
- BARTLETT, D. J., DESMOND, H., FERREIRA, P. G., 2024a, “Exhaustive Symbolic Regression”, *IEEE Transactions on Evolutionary Computation*, v. 28, n. 4 (ago.), pp. 950–964. ISSN: 1941-0026. doi: 10.1109/tevc.2023.3280250. Disponível em: <<http://dx.doi.org/10.1109/TEVC.2023.3280250>>.
- BARTLETT, D. J., DESMOND, H., FERREIRA, P. G., 2024b, “Exhaustive Symbolic Regression”, *IEEE Transactions on Evolutionary Computation*, v. 28, n. 4, pp. 950–964. doi: 10.1109/TEVC.2023.3280250.
- BASSETT, B. A., HLOZEK, R., 2009. “Baryon Acoustic Oscillations”. Disponível em: <<https://arxiv.org/abs/0910.5224>>.
- BAUMANN, D., 2022, *Cosmology*. 2 ed. Cambridge, Cambridge University Press. ISBN: 9781108838061.
- BENGALY, C., DANTAS, M. A., CASARINI, L., et al., 2023, “Measuring the Hubble constant with cosmic chronometers: a machine learning approach”, *Eur. Phys. J. C*, v. 83, n. 6, pp. 548. doi: 10.1140/epjc/s10052-023-11734-1.
- BHATTACHARYA, S., BORGHETTO, G., MALHOTRA, A., et al., 2024. “Cosmological constraints on curved quintessence”. Disponível em: <<https://arxiv.org/abs/2405.17396>>.
- BISHOP, C. M., 2006, *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York, Springer. ISBN: 978-0-387-31073-2. See Introduction and Chapter 1 for general definitions of machine learning and pattern recognition.



- BLUMENTHAL, G. R., FABER, S. M., PRIMACK, J. R., et al., 1984, “Formation of galaxies and large-scale structure with cold dark matter.” *Nature*, v. 311 (out.), pp. 517–525. doi: 10.1038/311517a0.
- BOGDANOS, C., NESSERIS, S., 2009, “Genetic algorithms and supernovae type Ia analysis”, *Journal of Cosmology and Astroparticle Physics*, v. 2009, n. 05 (maio), pp. 006–006. ISSN: 1475-7516. doi: 10.1088/1475-7516/2009/05/006. Disponível em: <<http://dx.doi.org/10.1088/1475-7516/2009/05/006>>.
- BOGGESESS, N. W., MATHER, J. C., WEISS, R., et al., 1992, “The COBE Mission: Its Design and Performance Two Years after Launch”, *Astrophysical Journal*, v. 397 (out.), pp. 420. doi: 10.1086/171797.
- BORGHETTO, G., MALHOTRA, A., TASINATO, G., et al., 2025, “Bounded dark energy”, *Phys. Rev. D*, v. 112, n. 2, pp. 023521. doi: 10.1103/tscs-h4xv.
- BOUSIS, D., PERIVOLAROPOULOS, L., 2024. “Hubble tension tomography: BAO vs SnIa distance tension”. Disponível em: <<https://arxiv.org/abs/2405.07039>>.
- BRAUN, R., BOURKE, T., GREEN, J. A., et al., 2015, “Advancing Astrophysics with the Square Kilometre Array”. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 174, abr. doi: 10.22323/1.215.0174.
- BROUT, D., ET AL., 2022, “The Pantheon+ Analysis: Cosmological Constraints”, *The Astrophysical Journal*, v. 938, n. 2 (out.), pp. 110. ISSN: 1538-4357. doi: 10.3847/1538-4357/ac8e04. Disponível em: <<http://dx.doi.org/10.3847/1538-4357/ac8e04>>.
- BROYDEN, C. G., 1970, “The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations”, *Ima J. Appl. Math.*, v. 6, n. 1, pp. 76–90. doi: 10.1093/imamat/6.1.76.
- BUSTI, V. C., CLARKSON, C., SEIKEL, M., 2014, “Evidence for a lower value for  $H_0$  from cosmic chronometers data?” *Monthly Notices of the Royal Astronomical Society*, v. 441 (jun.), pp. L11–L15. doi: 10.1093/mnrasl/slu035.
- CAGLIARI, M. S., OTHERS, 2024, “Euclid: Testing photometric selection of emission-line galaxy targets”, *Astron. Astrophys.*, v. 689, pp. A166. doi: 10.1051/0004-6361/202449970.
- CALDEIRA, J., WU, W., NORD, B., et al., 2019, “DeepCMB: Lensing reconstruction of the cosmic microwave background with deep neural networks”, *Astronomy and Computing*, v. 28 (jul.), pp. 100307. ISSN: 2213-1337. doi: 10.1016/j.ascom.

- 2019.100307. Disponível em: <<http://dx.doi.org/10.1016/j.ascom.2019.100307>>.
- CAPOZZIELLO, S., CARDONE, V. F., FUNARO, M., et al., 2004, “Constraining dark energy models using the lookback time to galaxy clusters and the age of the Universe”, *Phys. Rev. D*, v. 70, pp. 123501. doi: 10.1103/PhysRevD.70.123501.
- CARLEO, G., CIRAC, I., CRANMER, K., et al., 2019, “Machine learning and the physical sciences”, *Reviews of Modern Physics*, v. 91, n. 4 (dez.). ISSN: 1539-0756. doi: 10.1103/revmodphys.91.045002. Disponível em: <<http://dx.doi.org/10.1103/RevModPhys.91.045002>>.
- CHABOYER, B., 1998, “The age of the universe”, *Physics Reports*, v. 307, n. 1–4 (dez.), pp. 23–30. ISSN: 0370-1573. doi: 10.1016/s0370-1573(98)00054-4. Disponível em: <[http://dx.doi.org/10.1016/S0370-1573\(98\)00054-4](http://dx.doi.org/10.1016/S0370-1573(98)00054-4)>.
- CHEVALLIER, M., POLARSKI, D., 2001, “Accelerating Universes with Scaling Dark Matter”, *International Journal of Modern Physics D*, v. 10, n. 2 (jan.), pp. 213–223. doi: 10.1142/S0218271801000822.
- CONCEIÇÃO, M., KRONE-MARTINS, A., DA SILVA, A., et al., 2024, “Fast emulation of cosmological density fields based on dimensionality reduction and supervised machine learning”, *Astronomy and Astrophysics*, 681:A123. doi: 10.1051/0004-6361/202346734.
- CORTES, C., VAPNIK, V., 1995, “Support-vector networks”, *Machine Learning*, v. 20, n. 3, pp. 273–297. ISSN: 1573-0565. doi: 10.1007/BF00994018. Disponível em: <<https://doi.org/10.1007/BF00994018>>.
- CORTÊS, M., LIDDLE, A. R., 2024. “Interpreting DESI’s evidence for evolving dark energy”. Disponível em: <<https://arxiv.org/abs/2404.08056>>.
- CRANMER, M., 2023a, “Interpretable machine learning for science with PySR and SymbolicRegression.jl”, *arXiv preprint arXiv:2305.01582*.
- CRANMER, M., 2023b. “Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl”. b. Disponível em: <<https://arxiv.org/abs/2305.01582>>.
- DANTAS, M. A., ALCANIZ, J. S., JAIN, D., et al., 2007a, “Dark Energy Constraints from Gemini Deep Deep Survey”, *Astron. Astrophys.*, v. 467, pp. 421. doi: 10.1051/0004-6361:20066632.
- DANTAS, M., ALCANIZ, J., JAIN, D., et al., 2007b, “Age constraints on the cosmic equation of state”, *Astronomy & Astrophysics*, v. 467, n. 2, pp. 421–426.

- DANTAS, M., ALCANIZ, J., PIRES, N., 2009, “Current lookback time-redshift bounds on dark energy”, *Physics Letters B*, v. 679, n. 5 (Sep), pp. 423–427. ISSN: 0370-2693. doi: 10.1016/j.physletb.2009.08.008. Disponível em: <<http://dx.doi.org/10.1016/j.physletb.2009.08.008>>.
- DANTAS, M., ALCANIZ, J., MANIA, D., et al., 2011, “Time and distance constraints on accelerating cosmological models”, *Physics Letters B*, v. 699, n. 4, pp. 239–245.
- DE SOUZA, R., SOUSA-NETO, A., GONZÁLEZ, J. E., et al., 2025, “A model-independent assessment of the late-time dark energy density evolution”, (11).
- DESMOND, H., 2025, “(Exhaustive) Symbolic Regression and model selection by minimum description length”, *arXiv e-prints*, art. arXiv:2507.13033. doi: 10.48550/arXiv.2507.13033.
- DIALEKTOPOULOS, K., SAID, J. L., MIFSUD, J., et al., 2022, “Neural network reconstruction of late-time cosmology and null tests”, *Journal of Cosmology and Astroparticle Physics*, v. 2022, n. 02 (fev.), pp. 023. ISSN: 1475-7516. doi: 10.1088/1475-7516/2022/02/023. Disponível em: <<http://dx.doi.org/10.1088/1475-7516/2022/02/023>>.
- DINDA, B. R., MAARTENS, R., 2024. “Model-agnostic assessment of dark energy after DESI DR1 BAO”. Disponível em: <<https://arxiv.org/abs/2407.17252>>.
- DUNLOP, J., PEACOCK, J., SPINRAD, H., et al., 1996, “A 3.5-Gyr-old galaxy at redshift 1.55”, *Nature*, v. 381, n. 6583 (jun.), pp. 581–584. doi: 10.1038/381581a0.
- DVORKIN, C., MISHRA-SHARMA, S., NORD, B., et al., 2022. “Machine Learning and Cosmology”. Disponível em: <<https://arxiv.org/abs/2203.08056>>.
- EFSTATHIOU, G., 2025. “Evolving Dark Energy or Supernovae Systematics?” Disponível em: <<https://arxiv.org/abs/2408.07175>>.
- EGGEN, O. J., LYNDEN-BELL, D., SANDAGE, A. R., 1962, “Evidence from the motions of old stars that the Galaxy collapsed.” *Astrophysical Journal*, v. 136 (nov.), pp. 748. doi: 10.1086/147433.
- EINSTEIN, A., 1915, “Die feldgleichungen der gravitation”, *Sitzung der physikalisch-mathematischen Klasse*, v. 25, pp. 844–847.
- EISENSTEIN, D. J., ZEHAVI, I., HOGG, D. W., et al., 2005, “Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies”, *The Astrophysical Journal*, v. 633, n. 2 (nov.), pp. 560–574. ISSN: 1538-4357. doi: 10.1086/466512. Disponível em: <<http://dx.doi.org/10.1086/466512>>.

- ELBERS, W., AVILES, A., NORIEGA, H. E., et al., 2025. “Constraints on Neutrino Physics from DESI DR2 BAO and DR1 Full Shape”. Disponível em: <<https://arxiv.org/abs/2503.14744>>.
- ESCAMILLA-RIVERA, C., CARVAJAL QUINTERO, M. A., CAPOZZIELLO, S., 2020, “A deep learning approach to cosmological dark energy models”, *Journal of Cosmology and Astroparticle Physics*, 2020(3):008. doi: 10.1088/1475-7516/2020/03/008.
- FOREMAN-MACKEY, D., HOGG, D. W., LANG, D., et al., 2013, “emcee: The MCMC Hammer”, *Publ. Astron. Soc. Pacific*, v. 125, n. 925 (mar.), pp. 306. doi: 10.1086/670067.
- FREEDMAN, W., 2025, “The expanding Universe — do ongoing tensions leave room for new physics?” *Nature*, v. 639, n. 8056, pp. 858–860. doi: 10.1038/d41586-025-00896-5.
- GAMOW, G., 1946, “Expanding universe and the origin of elements”, *Physical review*, v. 70, n. 7-8, pp. 572.
- GHOSH, B., BENGALY, C., 2024a. “Consistency tests between SDSS and DESI BAO measurements”. a. Disponível em: <<https://arxiv.org/abs/2408.04432>>.
- GHOSH, B., BENGALY, C., 2024b, “Consistency tests between SDSS and DESI BAO measurements”, *Phys. Dark Univ.*, v. 46, pp. 101699. doi: 10.1016/j.dark.2024.101699.
- GIARÈ, W., NAJAFI, M., PAN, S., et al., 2024, “Robust preference for Dynamical Dark Energy in DESI BAO and SN measurements”, *JCAP*, v. 10, n. 10, pp. 035. doi: 10.1088/1475-7516/2024/10/035.
- GILLET, N., MESINGER, A., GREIG, B., et al., 2019, “Deep learning from 21-cm tomography of the cosmic dawn and reionization”, *Monthly Notices of the Royal Astronomical Society*, v. 484, n. 1 (mar.), pp. 282–293. doi: 10.1093/mnras/stz010.
- GONZALEZ, J. E., 2017, “Reconstruction of cosmological matter perturbations in modified gravity”, *Phys. Rev. D*, 96(12):123501. doi: 10.1103/PhysRevD.96.123501.
- GONZÁLEZ, J. E., ALCANIZ, J. S., CARVALHO, J. C., 2016, “Non-parametric reconstruction of cosmological matter perturbations”, *Journal of Cosmology and Astroparticle Physics*, 2016(4):016. doi: 10.1088/1475-7516/2016/04/016.

- GONZALEZ, J. E., ALCANIZ, J. S., CARVALHO, J. C., 2017, “Smoothing expansion rate data to reconstruct cosmological matter perturbations”, *Journal of Cosmology and Astroparticle Physics*, 2017(8):008. doi: 10.1088/1475-7516/2017/08/008.
- GONZALEZ, J., 2018, *Reconstrucao nao parametrica das perturbacoes cosmologicas*. Tese de Doutorado, National Astronomy Observatory, Brazil, mar.
- GRÜNWALD, P. D., 2007, *The Minimum Description Length Principle*. Cambridge, MA, MIT Press.
- GUY, J., ASTIER, P., BAUMONT, S., et al., 2007, “SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators \*\*\*”, *A&A*, v. 466, n. 1, pp. 11–21. doi: 10.1051/0004-6361:20066930. Disponível em: <<https://doi.org/10.1051/0004-6361:20066930>>.
- HE, S., LI, Y., FENG, Y., et al., 2019, “Learning to predict the cosmological structure formation”, *Proceedings of the National Academy of Sciences*, v. 116, n. 28 (jun.), pp. 13825–13832. ISSN: 1091-6490. doi: 10.1073/pnas.1821458116. Disponível em: <<http://dx.doi.org/10.1073/pnas.1821458116>>.
- HILDEBRANDT, H., OTHERS, 2017, “KiDS-450: Cosmological parameter constraints from tomographic weak gravitational lensing”, *Mon. Not. Roy. Astron. Soc.*, v. 465, pp. 1454. doi: 10.1093/mnras/stw2805.
- HLOŽEK, R., 2023, “How do we design data sets for Machine Learning astronomy?” *Proceedings of the International Astronomical Union*, v. 19, n. S368, pp. 11–27. doi: 10.1017/S1743921323001394.
- HOGG, D. W., 1999, “Distance measures in cosmology”, *arXiv e-prints*, art. astro-ph/9905116. doi: 10.48550/arXiv.astro-ph/9905116.
- HOGG, N. B., MARTINELLI, M., NESSERIS, S., 2020, “Constraints on the distance duality relation with standard sirens”, *Journal of Cosmology and Astroparticle Physics*, 2020(12):019. doi: 10.1088/1475-7516/2020/12/019.
- HUBBLE, E., 1929, “A relation between distance and radial velocity among extragalactic nebulae”, *Proceedings of the National Academy of Sciences*, v. 15, n. 3, pp. 168–173. doi: 10.1073/pnas.15.3.168. Disponível em: <<https://www.pnas.org/doi/abs/10.1073/pnas.15.3.168>>.
- JACOBS, C., COLLETT, T., GLAZEBROOK, K., et al., 2019, “Finding high-redshift strong lenses in DES using convolutional neural networks”, *Monthly Notices of*

- the Royal Astronomical Society*, v. 484, n. 4 (jan.), pp. 5330–5349. ISSN: 1365-2966. doi: 10.1093/mnras/stz272. Disponível em: <<http://dx.doi.org/10.1093/mnras/stz272>>.
- JIANG, J.-Q., PEDROTTI, D., DA COSTA, S. S., et al., 2024. “Non-parametric late-time expansion history reconstruction and implications for the Hubble tension in light of DESI”. Disponível em: <<https://arxiv.org/abs/2408.02365>>.
- JIMENEZ, R., 1997. “The Age of the Universe”. Disponível em: <<https://arxiv.org/abs/astro-ph/9701222>>.
- JIMENEZ, R., LOEB, A., 2002, “Constraining Cosmological Parameters Based on Relative Galaxy Ages”, *The Astrophysical Journal*, v. 573, n. 1 (jul.), pp. 37–42. ISSN: 1538-4357. doi: 10.1086/340549. Disponível em: <<http://dx.doi.org/10.1086/340549>>.
- JIMENEZ, R., MACDONALD, J., DUNLOP, J. S., et al., 2004, “Synthetic stellar populations: single stellar populations, stellar interior models and primordial protogalaxies”, *Monthly Notices of the Royal Astronomical Society*, v. 349, n. 1 (mar.), pp. 240–254. doi: 10.1111/j.1365-2966.2004.07492.x.
- KAMIONKOWSKI, M., RIESS, A. G., 2022. “The Hubble Tension and Early Dark Energy”. Disponível em: <<https://arxiv.org/abs/2211.04492>>.
- KARIM, M. A., OTHERS, 2025. “DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints”. Disponível em: <<https://arxiv.org/abs/2503.14738>>.
- KOZA, J. R., 1994, “Genetic programming as a means for programming computers by natural selection”, *Statistics and computing*, v. 4, pp. 87–112.
- KRAUSS, L. M., CHABOYER, B., 2003, “Age Estimates of Globular Clusters in the Milky Way: Constraints on Cosmology”, *Science*, v. 299, n. 5603 (jan.), pp. 65–70. doi: 10.1126/science.1075631.
- KRAUSS, L. M., TURNER, M. S., 1995, “The cosmological constant is back”, *General Relativity and Gravitation*, v. 27, n. 11 (nov.), pp. 1137–1144. ISSN: 1572-9532. doi: 10.1007/bf02108229. Disponível em: <<http://dx.doi.org/10.1007/BF02108229>>.
- LANDAU, L. D., LIFSHITZ, E. M., 1987, *Fluid Mechanics*, v. 6. Oxford, Pergamon Press.



- LANUSSE, F., MA, Q., LI, N., et al., 2018, “CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding”, *Monthly Notices of the Royal Astronomical Society*, v. 473, n. 3 (jan.), pp. 3895–3906. doi: 10.1093/mnras/stx1665.
- LEMOS, T., RUCHIKA, CARVALHO, J. C., et al., 2023, “Low-redshift estimates of the absolute scale of baryon acoustic oscillations”, *Eur. Phys. J. C*, v. 83, n. 6, pp. 495. doi: 10.1140/epjc/s10052-023-11651-3.
- LIMA, J. A. S., ALCANIZ, J. S., 2000, “Constraining the cosmic equation of state from old galaxies at high redshift”, *Monthly Notices of the Royal Astronomical Society*, v. 317, n. 4 (out.), pp. 893–896. doi: 10.1046/j.1365-8711.2000.03695.x.
- LINDER, E. V., 2003, “Exploring the Expansion History of the Universe”, *Phys. Rev. Lett.*, 90(9):091301. doi: 10.1103/PhysRevLett.90.091301.
- LOCHNER, M., MCEWEN, J. D., PEIRIS, H. V., et al., 2016, “PHOTOMETRIC SUPERNOVA CLASSIFICATION WITH MACHINE LEARNING”, *The Astrophysical Journal Supplement Series*, v. 225, n. 2 (ago.), pp. 31. ISSN: 1538-4365. doi: 10.3847/0067-0049/225/2/31. Disponível em: <<http://dx.doi.org/10.3847/0067-0049/225/2/31>>.
- LSST DARK ENERGY SCIENCE COLLABORATION, 2012, “Large Synoptic Survey Telescope: Dark Energy Science Collaboration”, *arXiv e-prints*, art. arXiv:1211.0310. doi: 10.48550/arXiv.1211.0310.
- LUONGO, O., MUCCINO, M., 2024. “Model independent cosmographic constraints from DESI 2024”. Disponível em: <<https://arxiv.org/abs/2404.07070>>.
- MARTÍN, A., YASIN, T., BARTLETT, D. J., et al., 2025. “Constraining dark matter halo profiles with symbolic regression”. Disponível em: <<https://arxiv.org/abs/2511.23073>>.
- MCCARTHY, P. J., BORGNE, D. L., CRAMPTON, D., et al., 2004, “Evolved Galaxies at  $z > 1.5$  from the Gemini Deep Deep Survey: The Formation Epoch of Massive Stellar Systems”, *The Astrophysical Journal*, v. 614, n. 1 (sep), pp. L9. doi: 10.1086/425306. Disponível em: <<https://dx.doi.org/10.1086/425306>>.
- MERTEN, J., GIOCOLI, C., BALDI, M., et al., 2019, “On the dissection of degenerate cosmologies with machine learning”, *Monthly Notices of the Royal Astronomical Society*, v. 487, n. 1 (jul.), pp. 104–122. doi: 10.1093/mnras/stz972.
- MORIWAKI, K., NISHIMICHI, T., YOSHIDA, N., 2023, “Machine learning for observational cosmology”, *Reports on Progress in Physics*, v. 86, n. 7 (maio),

- pp. 076901. ISSN: 1361-6633. doi: 10.1088/1361-6633/acd2ea. Disponível em: <<http://dx.doi.org/10.1088/1361-6633/acd2ea>>.
- MUKHERJEE, P., SEN, A. A., 2024. “Model-independent cosmological inference post DESI DR1 BAO measurements”. Disponível em: <<https://arxiv.org/abs/2405.19178>>.
- MURPHY, K. P., 2012, *Machine Learning: A Probabilistic Perspective*. 1 ed. Cambridge, MA, MIT Press. ISBN: 978-0262018029.
- NEAL, R. M., 1996, *Bayesian Learning for Neural Networks*, v. 118, *Lecture Notes in Statistics*. New York, Springer. ISBN: 978-0-387-94724-2. doi: 10.1007/978-1-4612-0745-0.
- NESSERIS, S., GARCÍA-BELLIDO, J., 2012, “A new perspective on dark energy modeling via genetic algorithms”, *Journal of Cosmology and Astroparticle Physics*, 2012(11):033. doi: 10.1088/1475-7516/2012/11/033.
- NESSERIS, S., SHAFIELOO, A., 2010, “A model-independent null test on the cosmological constant”, *Monthly Notices of the Royal Astronomical Society*, v. 408, n. 3 (nov.), pp. 1879–1885. doi: 10.1111/j.1365-2966.2010.17254.x.
- NOLAN, L. A., DUNLOP, J. S., JIMENEZ, R., 2001, “The Sun, stellar-population models and the age estimation of high-redshift galaxies”, *Monthly Notices of the Royal Astronomical Society*, v. 323, n. 2, pp. 385–390.
- OSTRIKER, J. P., STEINHARDT, P. J., 1995. “Cosmic Concordance”. Disponível em: <<https://arxiv.org/abs/astro-ph/9505066>>.
- PAL, S., SAHA, R., 2024, “ParamANN: a neural network to estimate cosmological parameters for  $\Lambda$ CDM Universe using Hubble measurements”, *Physica Scripta*, v. 99, n. 11 (out.), pp. 115007. ISSN: 1402-4896. doi: 10.1088/1402-4896/ad804d. Disponível em: <<http://dx.doi.org/10.1088/1402-4896/ad804d>>.
- PARK, C.-G., DE CRUZ PEREZ, J., RATRA, B., 2024. “Using non-DESI data to confirm and strengthen the DESI 2024 spatially-flat  $w_0w_a$ CDM cosmological parameterization result”. Disponível em: <<https://arxiv.org/abs/2405.00502>>.
- PEARSON, K., 1901, “On lines and planes of closest fit to systems of points in space”, *Philosophical Magazine*, v. 2, n. 11, pp. 559–572.



- PEEBLES, P. J. E., 1965, “The Black-Body Radiation Content of the Universe and the Formation of Galaxies.” *Astrophysical Journal*, v. 142 (nov.), pp. 1317. doi: 10.1086/148417.
- PEEBLES, P. J. E., 1984, “The Origin of Galaxies and Clusters of Galaxies”, *Science*, v. 224, n. 4656, pp. 1385–1391. doi: 10.1126/science.224.4656.1385. Disponível em: <<https://www.science.org/doi/abs/10.1126/science.224.4656.1385>>.
- PENZIAS, A. A., WILSON, R. W., 1965, “A measurement of excess antenna temperature at 4080 Mc/s.” *The Astrophysical Journal*, v. 142, pp. 419–421.
- PERLMUTTER, S., ALDERING, G., GOLDBABER, G., et al., 1999, “Measurements of  $\Omega$  and  $\Lambda$  from 42 High-Redshift Supernovae”, *The Astrophysical Journal*, v. 517, n. 2 (jun.), pp. 565–586. ISSN: 1538-4357. doi: 10.1086/307221. Disponível em: <<http://dx.doi.org/10.1086/307221>>.
- PIRES, N., ZHU, Z.-H., ALCANIZ, J. S., 2006, “Lookback time as a test for brane cosmology”, *Physical Review D*, v. 73, n. 12 (jun.). ISSN: 1550-2368. doi: 10.1103/physrevd.73.123530. Disponível em: <<http://dx.doi.org/10.1103/PhysRevD.73.123530>>.
- PLANCK COLLABORATION, 2020, “Planck2018 results: VI. Cosmological parameters”, *Astronomy & Astrophysics*, v. 641 (set.), pp. A6. ISSN: 1432-0746. doi: 10.1051/0004-6361/201833910. Disponível em: <<http://dx.doi.org/10.1051/0004-6361/201833910>>.
- PRESS, W. H., SCHECHTER, P., 1974, “Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation”, *Astrophysical Journal*, v. 187 (fev.), pp. 425–438. doi: 10.1086/152650.
- RAMADAN, O. F., SAKSTEIN, J., RUBIN, D., 2024. “DESI Constraints on Exponential Quintessence”. Disponível em: <<https://arxiv.org/abs/2405.18747>>.
- RASMUSSEN, C. E., WILLIAMS, C. K. I., 2006, *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, MIT Press. ISBN: 0-262-18253-X.
- REBOUÇAS, J., DE SOUZA, D. H. F., ZHONG, K., et al., 2024. “Investigating Late-Time Dark Energy and Massive Neutrinos in Light of DESI Y1 BAO”. Disponível em: <<https://arxiv.org/abs/2408.14628>>.
- RIESS, A. G., FILIPPENKO, A. V., CHALLIS, P., et al., 1998, “Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant”,

- The Astronomical Journal*, v. 116, n. 3 (set.), pp. 1009–1038. ISSN: 0004-6256. doi: 10.1086/300499. Disponível em: <<http://dx.doi.org/10.1086/300499>>.
- RIESS, A. G., CASERTANO, S., YUAN, W., et al., 2019, “Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond  $\Lambda$ CDM”, *The Astrophysical Journal*, v. 876, n. 1 (may), pp. 85. doi: 10.3847/1538-4357/ab1422. Disponível em: <<https://dx.doi.org/10.3847/1538-4357/ab1422>>.
- RIESS, A. G., OTHERS, 2022, “A Comprehensive Measurement of the Local Value of the Hubble Constant with  $1 \text{ km s}^{-1} \text{ Mpc}^{-1}$  Uncertainty from the Hubble Space Telescope and the SH0ES Team”, *Astrophys. J. Lett.*, v. 934, n. 1, pp. L7. doi: 10.3847/2041-8213/ac5c5b.
- RODRIGUES, G., DE SOUZA, R., ALCANIZ, J., 2025, “Cosmography with DESI DR2 and SN data”, *Phys. Rev. D*, v. 112, n. 10, pp. 103519. doi: 10.1103/fh98-1b3d.
- ROY CHOUDHURY, S., OKUMURA, T., 2024, “Updated Cosmological Constraints in Extended Parameter Space with Planck PR4, DESI Baryon Acoustic Oscillations, and Supernovae: Dynamical Dark Energy, Neutrino Masses, Lensing Anomaly, and the Hubble Tension”, *The Astrophysical Journal Letters*, v. 976, n. 1 (nov.), pp. L11. ISSN: 2041-8213. doi: 10.3847/2041-8213/ad8c26. Disponível em: <<http://dx.doi.org/10.3847/2041-8213/ad8c26>>.
- RUBIN, D., ALDERING, G., BETOULE, M., et al., 2024. “Union Through UNITY: Cosmology with 2,000 SNe Using a Unified Bayesian Framework”. Disponível em: <<https://arxiv.org/abs/2311.12098>>.
- SANTOS, J., ALCANIZ, J. S., REBOUCAS, M. J., et al., 2007, “Lookback time bounds from energy conditions”, *Phys. Rev. D*, v. 76, pp. 043519. doi: 10.1103/PhysRevD.76.043519.
- SCARAMELLA, R., OTHERS, 2022, “Euclid preparation. I. The Euclid Wide Survey”, *Astronomy and Astrophysics*, 662:A112. doi: 10.1051/0004-6361/202141938.
- SCHAYE, J., OTHERS, 2023, “The FLAMINGO project: cosmological hydrodynamical simulations for large-scale structure and galaxy cluster surveys”, *Mon. Not. Roy. Astron. Soc.*, v. 526, n. 4, pp. 4978–5020. doi: 10.1093/mnras/stad2419.
- SCHMELZLE, J., LUCCHI, A., KACPRZAK, T., et al., 2017, “Cosmological model discrimination with Deep Learning”, *arXiv e-prints*, art. arXiv:1707.05167. doi: 10.48550/arXiv.1707.05167.

- SCHMIDT, M., LIPSON, H., 2009, “Distilling free-form natural laws from experimental data”, *science*, v. 324, n. 5923, pp. 81–85.
- SCOLNIC, D. M., JONES, D. O., REST, A., et al., 2018, “The Complete Light-curve Sample of Spectroscopically Confirmed SNe Ia from Pan-STARRS1 and Cosmological Constraints from the Combined Pantheon Sample”, *The Astrophysical Journal*, v. 859, n. 2 (maio), pp. 101. ISSN: 1538-4357. doi: 10.3847/1538-4357/aab9bb. Disponível em: <<http://dx.doi.org/10.3847/1538-4357/aab9bb>>.
- SCOLNIC, D., BROUT, D., CARR, A., et al., 2022, “The Pantheon+ Analysis: The Full Data Set and Light-curve Release”, *The Astrophysical Journal*, v. 938, n. 2 (out.), pp. 113. ISSN: 1538-4357. doi: 10.3847/1538-4357/ac8b7a. Disponível em: <<http://dx.doi.org/10.3847/1538-4357/ac8b7a>>.
- SEIKEL, M., CLARKSON, C., 2013, “Optimising Gaussian processes for reconstructing dark energy dynamics from supernovae”, *arXiv e-prints*, art. arXiv:1311.6678. doi: 10.48550/arXiv.1311.6678.
- SEIKEL, M., CLARKSON, C., SMITH, M., 2012, “Reconstruction of dark energy and expansion dynamics using Gaussian processes”, *Journal of Cosmology and Astroparticle Physics*, v. 2012, n. 06 (jun.), pp. 036–036. ISSN: 1475-7516. doi: 10.1088/1475-7516/2012/06/036. Disponível em: <<http://dx.doi.org/10.1088/1475-7516/2012/06/036>>.
- SEIKEL, M., YAHYA, S., MAARTENS, R., et al., 2012, “Using  $H(z)$  data as a probe of the concordance model”, *Phys. Rev. D*, 86(8):083001. doi: 10.1103/PhysRevD.86.083001.
- SHAFIELOO, A., CLARKSON, C., 2010, “Model independent tests of the standard cosmological model”, *Phys. Rev. D*, 81(8):083537. doi: 10.1103/PhysRevD.81.083537.
- SHAH, R., SAHA, S., MUKHERJEE, P., et al., 2024, “LADDER: Revisiting the Cosmic Distance Ladder with Deep Learning Approaches and Exploring Its Applications”, *The Astrophysical Journal Supplement Series*, v. 273, n. 2 (jul.), pp. 27. ISSN: 1538-4365. doi: 10.3847/1538-4365/ad5558. Disponível em: <<http://dx.doi.org/10.3847/1538-4365/ad5558>>.
- SIMON, J., VERDE, L., JIMENEZ, R., 2005, “Constraints on the redshift dependence of the dark energy potential”, *Physical Review D*, v. 71, n. 12, pp. 123001.
- SIMONOFF, J. S., 2012, *Smoothing Methods in Statistics*. New York, Springer Science and Business Media.

- SMOOT, GORENSTEIN, MULLER, 1977, “Detection of Anisotropy in the Cosmic Blackbody Radiation”, *Phys. Rev. Lett.*, v. 39 (Oct), pp. 898–901. doi: 10.1103/PhysRevLett.39.898. Disponível em: <<https://link.aps.org/doi/10.1103/PhysRevLett.39.898>>.
- SMOOT, G. F., BENNETT, C. L., KOGUT, A., et al., 1992, “Structure in the COBE Differential Microwave Radiometer First-Year Maps”, *ApJL*, v. 396 (set.), pp. L1. doi: 10.1086/186504.
- SMOOT, G. F., 1999, “COBE observations and results”. In: Maiani, L., Melchiorri, F., Vittorio, N. (Eds.), *3K cosmology*, v. 476, *American Institute of Physics Conference Series*, pp. 1–10. AIP, maio. doi: 10.1063/1.59326.
- SMOOT ET AL., 1992, “Structure in the COBE Differential Microwave Radiometer First-Year Maps”, *ApJL*, v. 396 (set.), pp. L1. doi: 10.1086/186504.
- SOUSA, T., BARTLETT, D. J., DESMOND, H., et al., 2024, “Optimal inflationary potentials”, *Phys. Rev. D*, 109(8):083524. doi: 10.1103/PhysRevD.109.083524.
- SOUSA-NETO, A., DANTAS, M. A., 2025, “Reconstructing cosmic history with machine learning: a study using CART, MLPR, and SVR”, *Eur. Phys. J. C*, v. 85, n. 11, pp. 1320. doi: 10.1140/epjc/s10052-025-14884-6.
- SOUSA-NETO, A., BENGALY, C., GONZALEZ, J. E., et al., 2025, “Symbolic regression analysis of dynamical dark energy with DESI-DR2 and SN data”, *Phys. Dark Univ.*, v. 50, pp. 102108. doi: 10.1016/j.dark.2025.102108.
- SOUSA-NETO, A., DANTAS, M. A., GONZÁLEZ, J. E., et al., 2026, “Exploring cosmological constraints on galaxy formation time”, *Phys. Dark Univ.*, v. 51, pp. 102226. doi: 10.1016/j.dark.2026.102226.
- SPINRAD, H., DEY, A., STERN, D., et al., 1997, “LBDS 53W091: An Old, Red Galaxy at  $z = 1.552$ ”, *Astrophysical Journal*, v. 484, n. 2 (jul.), pp. 581–601. doi: 10.1086/304381.
- THUMMEL, L., BOSE, B., POURTSIDOU, A., et al., 2024, “Classifying modified gravity and dark energy theories with Bayesian neural networks: massive neutrinos, baryonic feedback, and the theoretical error”, *Monthly Notices of the Royal Astronomical Society*, v. 535, n. 4 (dez.), pp. 3141–3161. doi: 10.1093/mnras/stae2500.
- TREU, T., STIAVELLI, M., CASERTANO, S., et al., 1999, “The properties of field elliptical galaxies at intermediate redshift - I. Empirical scaling laws”, *Monthly*

- Notices of the Royal Astronomical Society*, v. 308, n. 4 (out.), pp. 1037–1052. doi: 10.1046/j.1365-8711.1999.02794.x.
- TREU, T., STIAVELLI, M., MØLLER, P., et al., 2001, “The properties of field elliptical galaxies at intermediate redshift - II. Photometry and spectroscopy of an HST-selected sample”, *Monthly Notices of the Royal Astronomical Society*, v. 326, n. 1 (set.), pp. 221–236. doi: 10.1046/j.1365-8711.2001.04593.x.
- TREU, T., STIAVELLI, M., CASERTANO, S., et al., 2002, “The Evolution of Field Early-Type Galaxies to  $z \sim 0.7$ ”, *Astrophysical Journal Letters*, v. 564, n. 1 (jan.), pp. L13–L16. doi: 10.1086/338790.
- TRIPP, R., 1998, “A two-parameter luminosity correction for Type IA supernovae”, *Astronomy and Astrophysics*, v. 331 (mar.), pp. 815–820.
- UDRESCU, S.-M., TEGMARK, M., 2020, “AI Feynman: A physics-inspired method for symbolic regression”, *Science Advances*, v. 6, n. 16 (abr.), pp. eaay2631. doi: 10.1126/sciadv.aay2631.
- VAGNOZZI, S., PACUCCI, F., LOEB, A., 2022, “Implications for the Hubble tension from the ages of the oldest astrophysical objects”, *Journal of High Energy Astrophysics*, v. 36 (nov.), pp. 27–35. ISSN: 2214-4048.
- VALCIN, D., JIMENEZ, R., SELJAK, U., et al., 2025. “The Age of the Universe with Globular Clusters III: Gaia distances and hierarchical modeling”. Disponível em: <<https://arxiv.org/abs/2503.19481>>.
- VERDE, L., BERNAL, J. L., HEAVENS, A. F., et al., 2017, “The length of the low-redshift standard ruler”, *Monthly Notices of the Royal Astronomical Society*, (jan.), pp. stx116. ISSN: 1365-2966. doi: 10.1093/mnras/stx116. Disponível em: <<http://dx.doi.org/10.1093/mnras/stx116>>.
- VERDE, L., SCHÖNEBERG, N., GIL-MARÍN, H., 2023. “A tale of many  $H_0$ ”. Disponível em: <<https://arxiv.org/abs/2311.13305>>.
- VILLAESCUSA-NAVARRO, F., DING, J., GENEL, S., et al., 2022, “Cosmology with One Galaxy?” *The Astrophysical Journal*, v. 929, n. 2 (abr.), pp. 132. ISSN: 1538-4357. doi: 10.3847/1538-4357/ac5d3f. Disponível em: <<http://dx.doi.org/10.3847/1538-4357/ac5d3f>>.
- VON MARTTENS, R., GONZALEZ, J., ALCANIZ, J., 2025. “Reconstructing the redshift evolution of Type Ia supernovae absolute magnitude”. Disponível em: <<https://arxiv.org/abs/2504.15127>>.

- WAN, Z. Y., SAPSIS, T. P., 2017, “Reduced-space Gaussian Process Regression for data-driven probabilistic forecast of chaotic dynamical systems”, *Physica D Nonlinear Phenomena*, v. 345 (abr.), pp. 40–55. doi: 10.1016/j.physd.2016.12.005.
- WANG, G.-J., MA, X.-J., LI, S.-Y., et al., 2020, “Reconstructing Functions and Estimating Parameters with Artificial Neural Networks: A Test with a Hubble Parameter and SNe Ia”, *The Astrophysical Journal Supplement Series*, v. 246, n. 1 (jan.), pp. 13. ISSN: 1538-4365. doi: 10.3847/1538-4365/ab620b. Disponível em: <<http://dx.doi.org/10.3847/1538-4365/ab620b>>.
- WEI, J.-J., MELIA, F., 2022, “Exploring the Hubble Tension and Spatial Curvature from the Ages of Old Astrophysical Objects”, *The Astrophysical Journal*, v. 928, n. 2 (apr), pp. 165.
- WEINBERG, D. H., MORTONSON, M. J., EISENSTEIN, D. J., et al., 2013, “Observational probes of cosmic acceleration”, *Physics Reports*, v. 530, n. 2 (set.), pp. 87–255. ISSN: 0370-1573. doi: 10.1016/j.physrep.2013.05.001. Disponível em: <<http://dx.doi.org/10.1016/j.physrep.2013.05.001>>.
- WEINBERG, S., 1972, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York, John Wiley & Sons. ISBN: 0-471-92567-5.
- WHITE, S. D. M., REES, M. J., 1978, “Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering.” *Monthly Notices of the Royal Astronomical Society*, v. 183 (maio), pp. 341–358. doi: 10.1093/mnras/183.3.341.
- WOLF, W. J., GARCÍA-GARCÍA, C., BARTLETT, D. J., et al., 2024. “Scant evidence for thawing quintessence”. Disponível em: <<https://arxiv.org/abs/2408.17318>>.
- WRIGHT, A. H., OTHERS, 2025, “KiDS-Legacy: Cosmological constraints from cosmic shear with the complete Kilo-Degree Survey”, *Astron. Astrophys.*, v. 703, pp. A158. doi: 10.1051/0004-6361/202554908.
- YIN, W., 2024. “Cosmic Clues: DESI, Dark Energy, and the Cosmological Constant Problem”. Disponível em: <<https://arxiv.org/abs/2404.06444>>.
- YUAN, G.-W., CALZÀ, M., PEDROTTI, D., 2025, “Machine learning-based analytical expressions for Gray-Body Factors and application to Primordial Black Holes”, *Physics of the Dark Universe*, 50:102078. doi: 10.1016/j.dark.2025.102078.

- ZHANG, C., WANG, C., HOBBS, G., et al., 2020, “Applying saliency-map analysis in searches for pulsars and fast radio bursts”, *Astronomy and Astrophysics*, v. 642 (set.), pp. A26. ISSN: 1432-0746. doi: 10.1051/0004-6361/201937234. Disponível em: <<http://dx.doi.org/10.1051/0004-6361/201937234>>.



# APPENDIX A

---

## Analyses with DESI DR1

---

In what follows, we present the reconstructed results obtained using the method described in Section 5.2, based on DESI DR1 data. These results are shown here for completeness and comparison, as the main analysis in this work has been updated to use DESI DR2.

First, we combine DESI DR1 with Pantheon+, and subsequently with DESY5. The best equation derived by PySR for the  $D_M/r_d$  reconstruction, in the first scenario, is given by

$$\frac{D_M}{r_d} = 29.8 \times 0.760^z z . \quad (\text{A.1})$$

Overall, the reconstruction using DR1 is consistent with that obtained from DR2. No significant changes are observed in the reconstructed trends, with both versions showing good agreement with the  $\Lambda$ CDM model.

Related to the  $w(z)$  reconstruction for DESI DR1 + Pantheon+, no significant deviations are observed when compared to the DR2-based reconstruction, indicating that the transition from DR1 to DR2 did not introduce notable changes in the inferred behavior of  $w(z)$ .

The resulting data for DESI DR1 + DESY5 is used to derive a new relationship for  $D_M/r_d$ . The best equation obtained from PySR for this dataset is

$$\frac{D_M}{r_d} = 29.8 \times 0.784^z z . \quad (\text{A.2})$$

In this case, we observe good agreement with the  $\Lambda$ CDM model across most of the redshift range for the  $D_M/r_d$  reconstruction. Notably, for  $z > 1.5$ , the reconstruction with DESI DR1 + DESY5 remains closer to  $\Lambda$ CDM compared to the DESI DR2 results, which show a mild deviation in that region (see Figure A.2, top). The reconstructed  $w(z)$  shows a distinct evolution compared to the result from DESI DR2. At high redshifts,  $w(z) < -1$ , increasing with time and crossing the phantom behaviours around  $z \sim 0.2$ , reaching  $w(z) > -1$  today ( $z = 0$ ) (see Figure A.2, bottom).



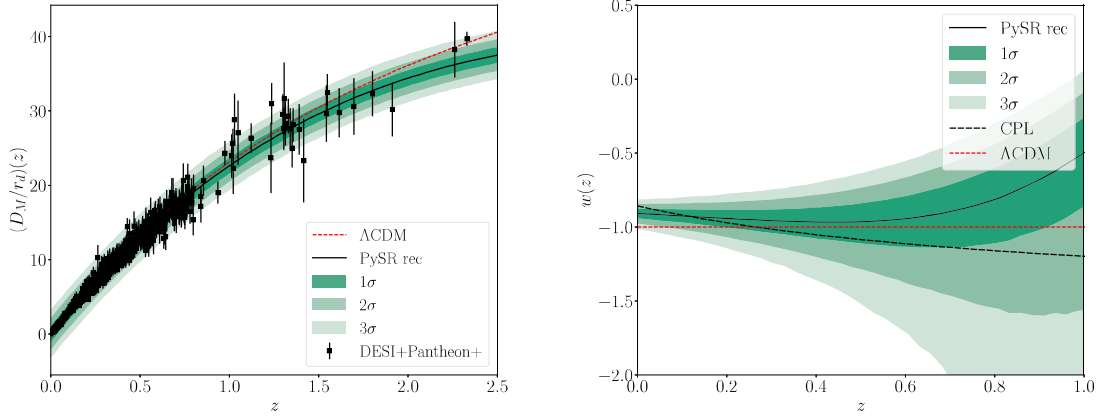


Figure A.1: The reconstruction results for  $D_M/r_d$  using DESI + Pantheon+ dataset. The black line represents the mean for the PySR reconstruction; the green bands indicate the confidence levels. The red dashed line corresponds to the  $\Lambda$ CDM theoretical curve. **Left:**  $D_M/r_d$  reconstruction with PySR. The black points correspond to the dataset, with respective uncertainties. **Right:** The  $w(z)$  derived using Eq. (5.6) and the reconstructed  $D_M/r_d$ , assuming  $r_d = 101 \pm 2.3h^{-1}$  Mpc, as reported by VERDE *et al.* (2017), and  $\Omega_m = 0.3$ . While the black dashed line represents the CPL model with the DESI (FS + BAO) + CMB + Pantheon+ results (ADAME *et al.*, 2024b).

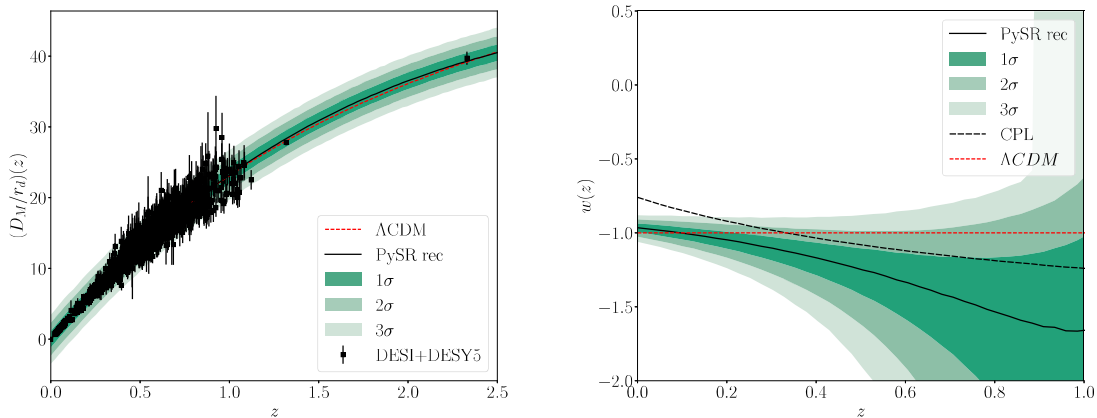


Figure A.2: The same as Figure A.1, but using the DESI + DESY5, and a CPL model with the DESI (FS + BAO) + CMB + DESY5 results (ADAME *et al.*, 2024b).