Constraining the assembly history of galaxies with cosmological simulations and deep learning

TRABAJO FIN DE MÁSTER

Submitted by

Maria Delgado Mancheño

Supervised by

Marc Huertas-Company

FACULTAD DE CIENCIAS: MASTER IN ASTROPHYSICS

July 2023

Agradecimientos

En primer lugar quería agradecer a Marc por haberme dado la oportunidad de realizar este trabajo, además de todo ese tiempo que me ha invertido. También a Eirini por haber contribuido en mi aprendizaje tanto astrofísico como de idiomas.

A mi abuela y a mi tía que han estado ahí desde el principio, apoyándome desde el principio de la vida universitaria hasta ahora y haciendo todo lo posible porque pudiera estudiar lo que me gusta.

A Amanda, mi mejor amiga, la cual lleva soportandome mucho tiempo y espero conservar toda la vida.

Por último y, como mención especial, a Matías. Por haber estado siempre ahí, por las correcciones y por todo lo demás. Nunca voy a poder agradecertelo lo suficiente.

Muchas gracias.

Abstract

Las simulaciones cosmológicas hidrodinámicas, como el proyecto IllustrisTNG, desempeñan un papel fundamental en el estudio de la formación y evolución de las galaxias, proporcionándonos un mayor conocimiento sobre el Universo.

Sin embargo, obtener información directa a partir de las observaciones puede resultar difícil, ya que las propiedades observacionales no dejan trazas de cómo han evolucionado con el tiempo. Por lo tanto, resulta interesante desarrollar modelos que nos permitan obtener información acerca de la formación de las galaxias a partir de las observaciones. En la actualidad, se están llevando a cabo estudios que combinan Machine Learning con las simulaciones cosmológicas para investigar la historia de fusión de las galaxias. Sin embargo, muchos de estos modelos carecen de la capacidad de generalizar cuando se aplican a datos observacionales reales.

El estudio realizado por Angeloudi et al. (2023)[1] aborda precisamente este problema mediante la combinación de simulaciones, como TNG100 e EAGLE, evaluándo cómo las diferencias cosmológicas afectan a la determinación de la fracción de masa ex-situ. Los resultados obtenidos en esta investigación indican que es posible desarrollar modelos robustos y precisos de Machine Learning utilizando mapas de masa estelar en dos dimensiones que incluyan información cinemática. Este avance es significativo, ya que permite cuantificar la contribución de la masa estelar ex-situ en la formación y evolución de las galaxias, la cual se define como aquella aportación de masa debida a una galaxia externa.

Con el objetivo de mejorar este modelo y evaluar su capacidad de generalización, se propone mediante el entrenamiento con la simulación TNG100, examinar las predicciones obtenidas de las simulaciones del proyecto "Cosmology and Astrophysics with Machine Learning Simulation"

(CAMELS). Este proyecto varía diferentes parámetros astrofísicos y cosmológicos en comparación con la simulación con la que hemos entrenado el modelo.

Para hacer este estudio y conseguir la mayor presición posible, previamente analizamos la correlación entre TNG100 y el Modelo Fiducial proveniente de CAMELS, que tiene las mismas condiciones iniciales que el primero pero con una menor resolución. Esto nos llevo a la conclusión de que había que eliminar a la Metalicidad y la Half-Stellar-Mass-Ratio como datos de entrada en el proceso de entramiento debido a su mala correlación con la fracción de ex-situ. Posteriormente, mediante una comprobación visual, se verifica que ambas simulaciones tienen la misma tendencia y que la discrepacia de errores, obtenidos como la diferencia entre el valor real y el predecido, se debe a la diferencia de resolución que tienen cada una.

Una vez realizado este paso de comprobación, procedimos a tratar los datos procedentes de CAMELS. Para ello, mostramos gráficamente los errores calculados como se mencionó antes y la diferencia de desviaciones estándar. Comprobamos que para ciertos parámetros, en su mayoría los más extremos, adquieren unos valores muy significativos de estas dos magnitudes estadísticas, en comparación con los otros. Una posible justificación que encontramos fue que en estas simulaciones existe una mala correlación con la fracción de ex-situ, y es por ello por lo que la precisión a la hora de obtener las predicciones no sea muy buena. A pesar de todo, las diferencias obtenidas no son considerablemente grandes, con lo cual se podría afirmar la generalidad del modelo.

Por otro lado, se propone utilizar la biblioteca de High-Performance Symbolic Regression (PySR), que permite obtener una expresión analítica para calcular la fracción de ex-situ a partir de los parámetros observacionales. Para ello, desarrollamos un código mediante los comandos que nos ofrece, el cual genera diferentes ecuaciones con distintos niveles de complejidad. Evaluamos la diferencia entre el valor esperado y el obtenido mediante un visualizado de cajas, en el que mostramos los errores que se obtienen según distintos porcentajes de datos, dónde se encuentra la mediana y los valores átipicos. De todas las ecuaciones obtenidas, seleccionamos como candidata la que aporta menos error al cálculo de la fracción de masa ex-situ. Esta expresión matemática es acorde a lo esperado, ya que depende solo de la masa estelar de la galaxia.

Esto es justificable, ya que cuánto más masiva son las galaxias, mayor es la probabilidad de que se hayan formado mediante fusiones, contribuyendo a una parte importante de su masa.

Posteriormente, comparamos lo obtenido con PySR con las predicciones obtenidas mediante

Machine Learning de forma visual a través de un gráfico masa estelar - fracción de ex-situ. Comprobamos que la ecuación no era capaz de representar correctamente la dispersión debido a su forma lineal, a diferencia del modelo. En cuanto al error, con Machine Leaning se obtienen valores más cercanos a los que buscamos que con Symbolic Regression, que a pesar de ello no tiene errores muy significativos.

Como conclusión final, encontramos ciertos valores que presentan una mayor diferencia con los esperados de fracción de ex-situ, en comparación con los otros parámetros astrofísicos y cosmológicos. A pesar de ello, estas diferencias no tienen mucho peso, por lo que se puede asegurar la generalidad del modelo. Además, aunque High-Performance Symbolic Regression es una herramienta útil que permite encontrar expresiones sencillas que relacionan diferentes magnitudes, el modelo utilizado mediante Machine Learning es mejor para calcular la fracción de masa ex-situ. Por otro lado, la expresión obtenida a través de Symbolic Regression no presenta errores muy significativos, pudiendo proporcionar una primera estimación fácil y sencilla de esta magnitud

Como trabajo futuro, se propone estudiar en detalle el por qué de esta falta de correlación con algunas simulaciones, con el fin de conseguir la menor diferencia entre las predicciones y el valor real. Además, se podría combinar Symbolic Regression con este modelo para poder obtener ecuaciones más complejas y con mayor precisión para el cálculo de la fracción de ex-situ.

Contents

1 Introduction

Galaxies are gravitationally bound systems composed of stars, interstellar matter, stellar remnants, and large amounts of dark matter. They are various systems with a wide range of morphologies and properties[2]. Their study is related to cosmology, which is a branch from which the evolution of the universe is studied.

The Big Bang theory, one of the prominent models explaining the origin and development of the universe, receives substantial support from the discovery of the cosmic microwave background (CMB). This gives rise to a vision of the universe based on this and the ΛCDM cosmological framework, which is known as the standard model of cosmology. It is based on the assumption that the universe is homogeneous and isotropic on large scales, made of baryonic matter, photons, neutrinos, and cold matter, which are composed of non-relativistic massive particles that interact with each other and with ordinary matter.

A detailed study of galaxies can be useful to test this model because it provides us with information about how baryonic and dark matter are related as a function of cosmic time. For this purpose, cosmological hydrodynamical galaxy simulations are valuable tools, which allow us to simulate the formation and evolution of galaxies on a large scale, taking into account both gravitational dynamics and hydrodynamic processes involving baryonic gas. These simulations enable us to accurately reproduce realistic galaxies, from which we can establish the relationship between their physical and observational properties, thereby enhancing our understanding of the mechanisms driving their evolution.

In the standard model of cosmology, the mergers of galaxies play a fundamental role to assemble the galaxies we see today. This model indicates that the growth of the structure occurs bottom-up. From smaller systems, larger systems are formed through subsequent mergers, which is known as hierarchical merging. The growth history is characterized by its merger tree, which is a graph plotting the progenitors in cosmic time[2]. The merger of galaxies provides a galaxy with new stars. The fraction of stellar mass in a given halo that was formed in another halo before the merger is referred to as the ex-situ stellar mass or to abbreviate, ex-situ fraction (as opposed to in-situ mass which refers to stars formed from gas in the halo).

Having information about these processes can contribute to a better understanding of the galaxy's history, particularly in its early stages in the universe. However, obtaining direct information about the stellar assembly histories from observational data is challenging, because observations do not provide clear traces of how they have evolved and because we cannot obviously follow a given halo across cosmic time.

To bridge this gap, cosmological simulations become crucial as they provide distinct snapshots of galaxy evolution across cosmic time. Notably, the most recent simulations, like the IllustrisTNG project, have the capability to generate a diverse and realistic population of galaxies[3]. These simulations capture both observational aspects and physical processes, enabling us to establish connections between different aspects of galaxy evolution.

A concrete example that showcases the potential contributions of simulations is demonstrated in studies like Oser et al. (2010)[4]. In this research, they employ a technique known as zoom-in simulations, which focus specifically on regions of interest. Through these targeted simulations, they unveil a significant finding that the most massive galaxies primarily consist of ex-situ stars, which are acquired through mergers with satellite galaxies. To complement and enhance these investigations, additional tools such as neural networks can be employed.

Currently, there are numerous machine learning studies trained on cosmological simulations aimed at investigating the merging history of galaxies like Lovell C. C.(2019)[5]. However, most of these studies do not account for the effects of different physics in each simulation, resulting in models trained with specific sub-grid recipes (baryonic physics) that lack generalization when applied to real observational data. In this regard, Angeloudi et al. (2023)[1] combines two cosmological simulations, TNG100 and EAGLE, to assess how differences in cosmology affect the determination of the ex-situ fraction. Their findings suggest that robust and accurate models can be achieved when using spatially resolved 2D stellar mass maps with kinematic information, which provides insights into the motion and kinematics of galaxies. However, when including age and metallicity maps, the estimation of the ex-situ fraction is affected by a phenomenon known as domain drift, which occurs when the accuracy is diminished due to the use of data from different domains.

The purpose of this Master's Thesis is to investigate further the robsutness of this estmation by quantifying the impact of changing cosmological or astrophysical parameters on the calculation of the ex-situ mass fraction of galaxies using neural networks. To accomplish this, we will employ the model used by Angeloudi et al. (2023)[1], which will be trained with the TNG100 simulation and attempt to predict the ex-situ stellar mass fraction from simulations provided by the Cosmology and Astrophysics with Machine Learning Simulation (CAMELS) project, encompassing various physical and cosmological parameters, along with the fiducial TNG100 simulation. Additionally, we propose using High-Performance Symbolic Regression in Python and Julia (PySR), which allows us to obtain a simple analytical equation that relates the ex-situ fraction to observational parameters. Both calculation methods will be discussed and compared at the end of this study.

2 Objectives

The main objective of this work is to study how robust is the estimate of the ex-situ fraction from a neural network model using only observable information[1] when cosmology and baryonic physics are changed. In addition, we use High-Performance Symbolic Regression in Python and Julia (PySR) to obtain an equation that allows us to estimate this parameter in a simple way. The following objectives were accomplished:

- Obtain the prediction of the ex-situ mass fraction from the neural network model, which is trained on TNG100 from the IllustrisTNG project.
- Explore the CAMELS dataset to quantify the robustness of the estimator when cosmological and astrophysical parameters change.
- Use Symbolic Regression to obtain an analytic expression for the ex-situ fraction based on observational properties.
- Perform a comparison between the results from the Neural Network model and the High-Performance Symbolic Regression.

3 Simulations and Data

3.1 TNG100

The IllustrisTNG[6] project performs a series of large, cosmological magnetohydrodynamical simulations of galaxy formation. Simulations come in three physical box volumes (TNG50, TNG100 and TNG300) corresponding to comoving volumes of 51.7^3, 110^3 and 302.6^3, and spatial resolutions around 0.1 kpc. Each TNG simulation starts from a redshift z=127 to z=0[7]. In this work, we used TNG100, which we used to train the Neural Network model. Its characteristics are a volume of 110^3 Mpc^3 which contains 1820^3 dark matter particles and initial gas cells (elements used to represent the gas distribution).[1].

3.2 CAMELS

CAMELS[8] is a comprehensive dataset designed to establish connections between cosmology and astrophysics through the use of machine learning techniques. It comprises a total of 6325 simulations, including 3049 N-body simulations and 3276 state-of-the-art (magneto) hydrodynamic simulations within a periodic comoving volume of $(25h^{-1}\text{Mpc})^3$ from redshift z=127 to z=0[8]. These simulations are categorized into different suites or folders, namely IllustrisTNG, SIMBA, Astrid, and N-body. Among these, we specifically focus on the IllustrisTNG suite, which consists of 1092 hydrodynamic simulations. These simulations employ the same subgrid physics, encompassing identical models and algorithms as the original IllustrisTNG simulations explained in the Section(3.1). It has a series of sets or collections of simulations and, in our case, we utilize the 1P set consisting of 61 simulations. Additionally, it includes a simulation known as the Fiducial Model, which has the same physical parameters as the TNG100 but in a smaller volume and a spatial resolution roughly two times smaller. In each simulation, the cosmological and astrophysical parameters listed in Table (1) vary individually. The following provides a more detailed explanation of the concepts[9]:

- **Fraction of energy density in matter**: This refers to the ratio between the energy density of matter (both baryonic and dark matter), which represents the amount of energy stored in the form of matter within a given volume of space, and the critical energy density.

The critical energy density is the average density of matter and energy required for the universe to be spatially flat.

- **Variance of the linear field**: This concept quantifies the fluctuations or variations in the density or temperature field of the universe on a specific spatial scale. It measures the degree of deviation from the average value of the field on that scale.

- **Galactic winds: Energy per unit SFR**: Galactic winds refer to the outflows of gas and energy from galaxies, typically driven by supernovae explosions. This concept specifically represents the amount of energy released by these galactic winds per unit of star formation rate (SFR) of the galaxy. It indicates how much energy is being expelled relative to the rate at which new stars are being formed.

- **Galactic winds: Wind speed**: This term describes the velocity at which galactic winds travel relative to the galaxy itself. It represents the speed at which the outflowing gas and energy are moving away from the galaxy due to the influence of processes such as supernovae.

- **Energy per unit Black Hole (BH) accretion rate**: This concept refers to the amount of energy released per unit of time as a black hole accretes matter from its surroundings.

- **Ejection speed/burstiness**: This term describes the velocity at which material is expelled or ejected from the vicinity of a black hole.

Parameter	Meaning	Range
Ω_m	Fraction of energy density in matter.	0.1-0.5
σ_s	Variance of the linear field on 8h^{-1} M at z=0.	0.6-0.1
$\mathrm{A}_{\mathrm{SN}_1}$	Galactic winds: Energy per unit SFR.	0.25-4.0
$\mathrm{A}_{\mathrm{SN}_2}$	Galactic winds: winds speed.	0.5-2.0
$\mathrm{A}_{\mathrm{AGN}_1}$	Energy per unit BH accretion rate.	0.25-4.0
$\mathrm{A}_{\mathrm{AGN}_2}$	Ejection speed/burstiness.	0.5-2.0

Table 1: Summary table with the meanings of the different parameters of CAMELS[8].

4 Methodology

4.1 Calculation of the ex-situ stellar mass fraction

For the determination of the ex-situ stellar mass fraction, Rodriguez-Gomez et al (2016)[10] classified all the stellar particles based on their formation and accretion histories. For this purpose, they use the merger trees calculated with the SubLink algorithm[10], that tracks individual star particles and star-forming gas elements. In Figure (1), we can see an example of a merger tree of TNG100.

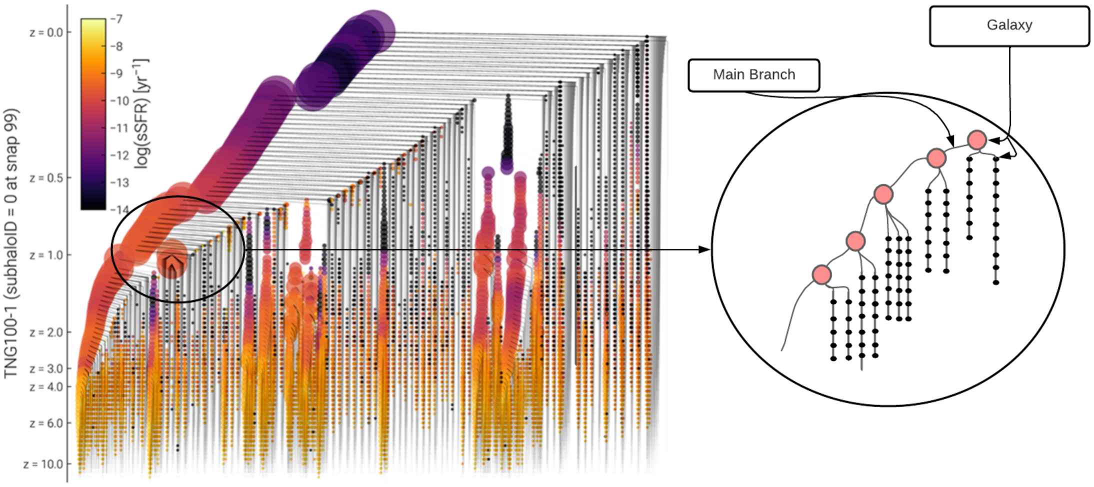

Figure 1: Merger tree from TNG100. The black dots represent different galaxies that merged to form larger structures, which are represented by the pink dots. These pink dots located in the main progenitor branch, in turn, represent the progenitor branch with the most massive structures in the history, also known as the main progenitors[6].

In cosmological simulations, galaxies form through the gradual accumulation of matter into larger structures as the universe evolves. These larger structures, known as halos, are high-density regions composed of dark matter and gas. Within these halos, smaller substructures can form, referred to as subhalos or galaxies. These structures play a crucial role in the creation of a merger tree.

For each snapshot, TNG100 identifies halos with the friends-of-friends (FoF) algorithm[11]. Then, the particles are analyzed, and those that are gravitationally bound together form galaxies, identified as subhalos with the Subfind algorithm([12], [13]). Finally, we utilize the algorithm mentioned at the beginning of the section, namely the SubLink algorithm[14], where the possible descendants are searched in each subhalo for their construction.

The criteria for identifying them rely on the particle coincidence. It is checked whether there are common particles between the subhalo and the descendant candidates in the next timestep. The unique descendant is assigned based on the highest number of shared particles and thus the merger tree is constructed in this manner. The main progenitor, located in the main branch, is determined as the most massive in the history.

Once this information has been obtained, Rodriguez-Gomez et al. (2016)[10] defined ex-situ stellar particles as those stellar particles formed outside of the main progenitor branch, while in-situ stellar particles were defined as those formed through internal star formation. Based on these definitions, TNG created a catalog containing the amount of ex-situ and in-situ stellar mass for all subhalos in TNG100, which has been calculated by Rodriguez et al.([14] ,[10],[15])

Finally, we calculated the ex-situ stellar mass fraction as the division of the ex-situ stellar mass and the total stellar mass of the galaxy. For each CAMELS simulation, we follow the same procedure for calculating the ex-situ mass fraction. This ensures that both the CAMELS simulations and the TNG100 simulation adhere to the same definition and calculation of this quantity.

4.2 The model

A Neural Network is a computational system inspired by the functioning of neurons. It is based on the connection between input values and output values through a non-linear function (activation function), which adjusts weights during the learning process. Multiple layers, also known as hidden layers, can be added, allowing the signal to travel from the first (the input layer) to the last (the output layer)[16].

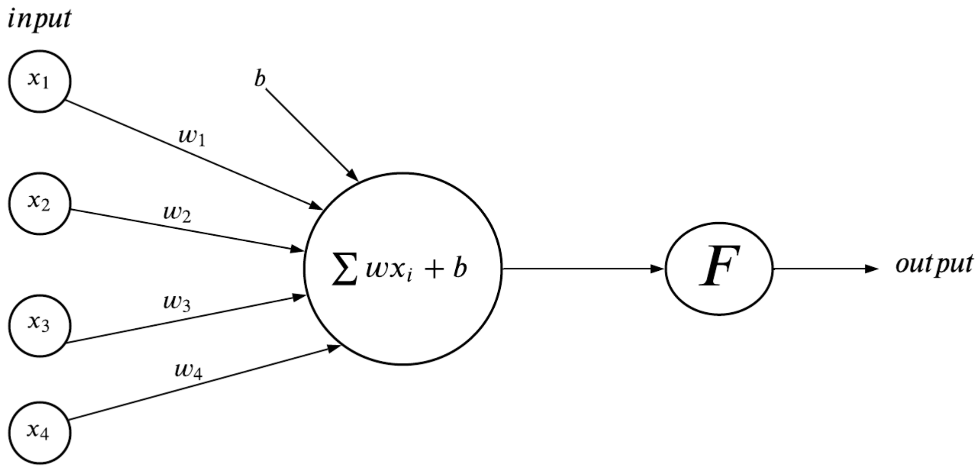

Figure 2: Basic scheme of a Neural Network, where input values are combined with weights and a bias to perform an operation represented in the large circle. The result of this operation is then passed through an activation function F, which generates an output.

Through the training process, the model adjusts and improves the weights and biases. As a result, more accurate predictions are obtained, meaning that the differences between the input and output are minimized. In our case, the Backpropagation algorithm is utilized, which operates as follows:

1. **Forward propagation**: this step consists of randomly initializing the weights (w) and biases (b). Each input is associated with a weight, and they are multiplied by their respective weights. The sum over is then calculated and added to the bias as follows

$$z(\vec{x}) = \sum \vec{w} \cdot \vec{x} + b, \tag{1}$$

where $z(\vec{x})$ is the pre-activation. This result is passed through the activation function as described below

$$f(\vec{x}) = g(z(\vec{x})) = g(\vec{w} \cdot \vec{x} + b), \tag{2}$$

where $f(\vec{x})$ is the output and $g(\vec{w} \cdot \vec{x} + b)$ the activation function. In Figure(2), a diagram illustrating this step of the algorithm can be observed. The process is repeated for each subsequent layer, with the results flowing to the next layers until reaching the output layer along the direction shown in Figure(3)[17].

2. **Backpropagation**: when obtaining the final prediction, the loss function is applied to evaluate the difference between the obtained and true value. In our case, we assume a Gaussian posterior distribution for the ex-situ fraction parametrized with the mean and the standard deviation. Therefore, we maximize the likelihood using the following equation since our distribution is Gaussian

$$L(w, g_w(x_i)^2; y, x) = (2\pi g_w(x_i)^2)^{-N/2} exp(-\frac{1}{2\pi g_w(x_i)^2}\sum_{i=1}^{N}(y_i - f_w(x_i))^2))[17], \tag{3}$$

where w represents the weights and the biases of the Neural Network which are learned, and f_w and g_w design 2 outputs of the Neural Network parametrizing the mean and standard deviation of the posterior. y_i is the true value, x_i is the input and N is the total size of the sample. The objective of the training is to maximize the likelihood (or to minimize the negative likelihood). In other words, we aim at finding the weights and the bias that produce the minimum value of the loss function. The minimization is performed with standard gradient descent

$$w_{t+1} = w_t - \alpha\frac{\partial L}{\partial w_t}, \tag{4}$$

$$b_{t+1} = b_t - \alpha\frac{\partial L}{\partial b_t},$$

where α is the learning rate, which in our case is equal to 10^{-3}, and $\frac{\partial L}{\partial}$ is the derivate of the loss function with respect the weights or the bias.

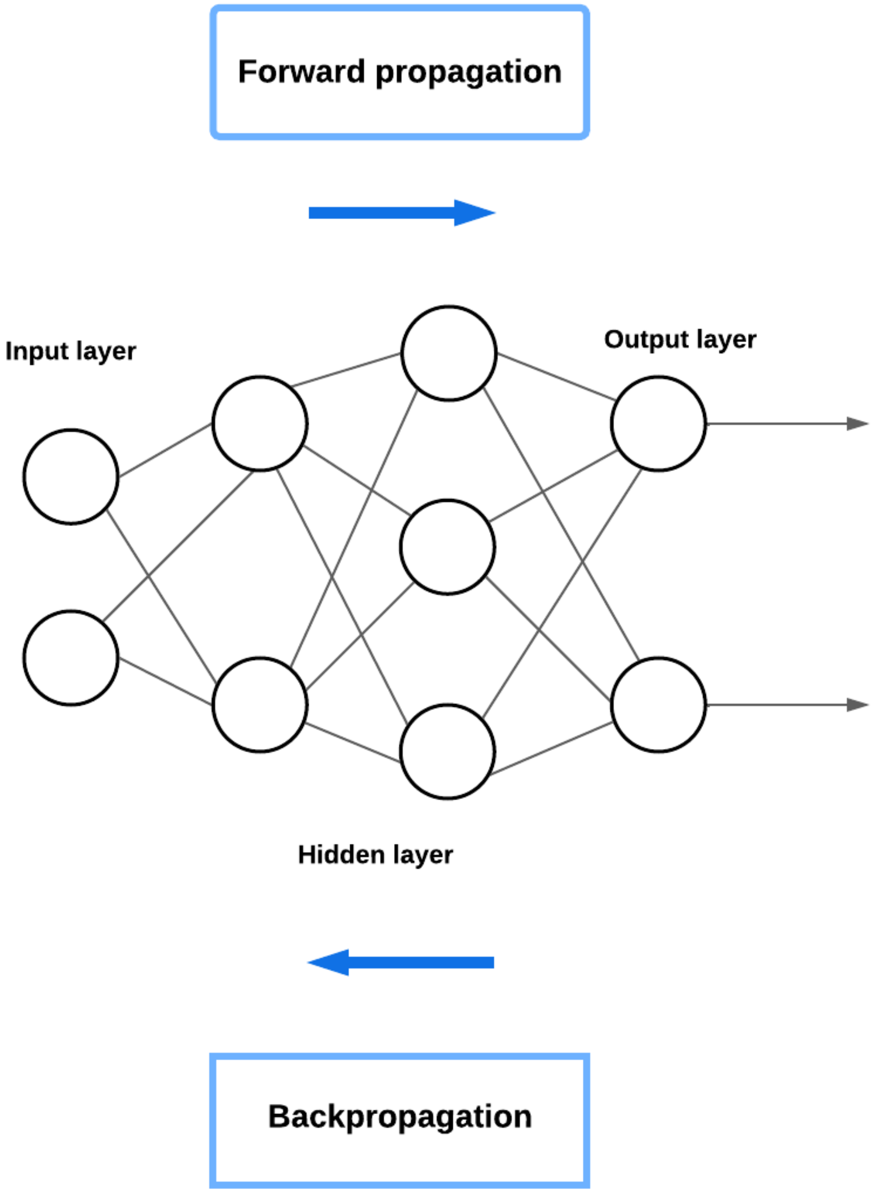

Figure 3: Scheme of the Neural Network training where the different layers are represented and the arrows indicate the direction followed by the steps of the Backpropagation algorithm.

Hence, the process involves moving backward following the direction indicated in Figure (3). To achieve this, we start by modifying the weights and the biases of the output layer using Equation (4). First of all, the algorithm has to calculate the derivatives, and for this purpose, the chain rule is employed as follows[17],

$$\frac{\partial L}{\partial w_t} = \frac{\partial L}{\partial f(\vec{x})} \cdot \frac{\partial f(\vec{x})}{\partial z(x)} \cdot \frac{\partial z(x)}{\partial w_t}, \tag{5}$$

$$\frac{\partial L}{\partial b_t} = \frac{\partial L}{\partial f(\vec{x})} \cdot \frac{\partial f(\vec{x})}{\partial z(x)} \cdot \frac{\partial z(x)}{\partial b_t}.$$

Subsequently, by carrying out these operations, the various weights are iteratively adjusted in a backward propagation manner, progressing towards the input layer. At the input layer, the process is reiterated starting anew. This algorithm follows a cyclic pattern and continues until a certain error criterion is achieved, determined either by convergence or user-defined threshold, or until a predetermined number of iterations[16].

4.3 Preprocessing

Before conducting any training with the model, we need to perform a series of checks and adjustments to achieve the highest possible accuracy. The first thing we did was to select the range of galaxies we were going to work with. As explained in Angeloudi et al. (2023)[1], in the case of TNG100 at redshift z=0, they found that the galaxy sample is highly unbalanced, with a large number of galaxies having a low ex-situ mass fraction and very few with high values of this magnitude (>0.7). To avoid any impact on the predictions, we updated our TNG100 dataset by including galaxies that have an ex-situ fraction value greater than 0.2 at z=0.1 and z=0.2.

Regarding the case of CAMELS, it also needs to be adjusted for this purpose. Therefore, we utilize galaxies with redshifts z=0, z=0.05, z=0.1, z=0.15, and z=0.2. Since smaller boxes generally contain fewer galaxies, we decide to include intermediate range redshifts in this sample. Additionally, as shown in Figure (4), the mass range of the TNG100 sample is greater than 10^{10} $M_\odot$, which is why we ensure that CAMELS has the same range.

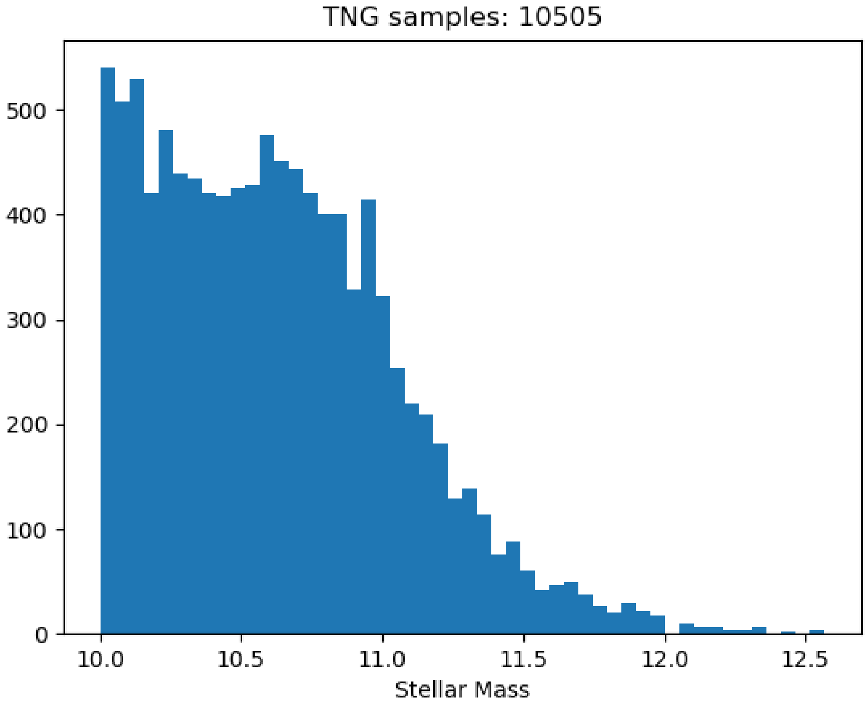

Figure 4: Histogram of the stellar mass of TNG100, which contains 10505 galaxies at redshifts z=0, z=0.1 and z=0.2.

CAMELS offers us, as we mentioned in Section(3.2), a dataset called IllustrisTNG, which modifies some cosmological and astrophysical parameters from the original . The Fiducial Model has the same cosmological and astrophysical parameters as TNG100, as we can see in Table(2). It provides us with a sample of 2175 galaxies in the mass and redshift range that we defined above. We use this simulation to check that the model trained in TNG100 generalizes well to CAMELS fiducial with lower spatial resolution, before trying it out with the other data set[8].

Name	Ω_m	σ_s	A_{SN_1}	A_{SN_2}	A_{AGN_1}	A_{AGN_2}
Fiducial Model	0.30000	0.80000	1.00000	1.00000	1.00000	1.00000

Table 2: Values of the parameters for the Fiducial Model.

An important aspect to consider is the values of the ex-situ fraction contained in TNG100 and in the Fiducial Model. This is because if they had a different range, for example, one of them had higher values and the other lower values, it could affect the predictions. In Figure(5), a histogram normalized of the ex-situ stellar mass fractions from these simulations is shown. As we can see, they have an imbalanced distribution due to the presence of many low ex-situ values, mainly in the Fiducial Model, and very few high ones in both distributions. In addition, they cover the same ex-situ fraction range. For this reason no further changes are made.

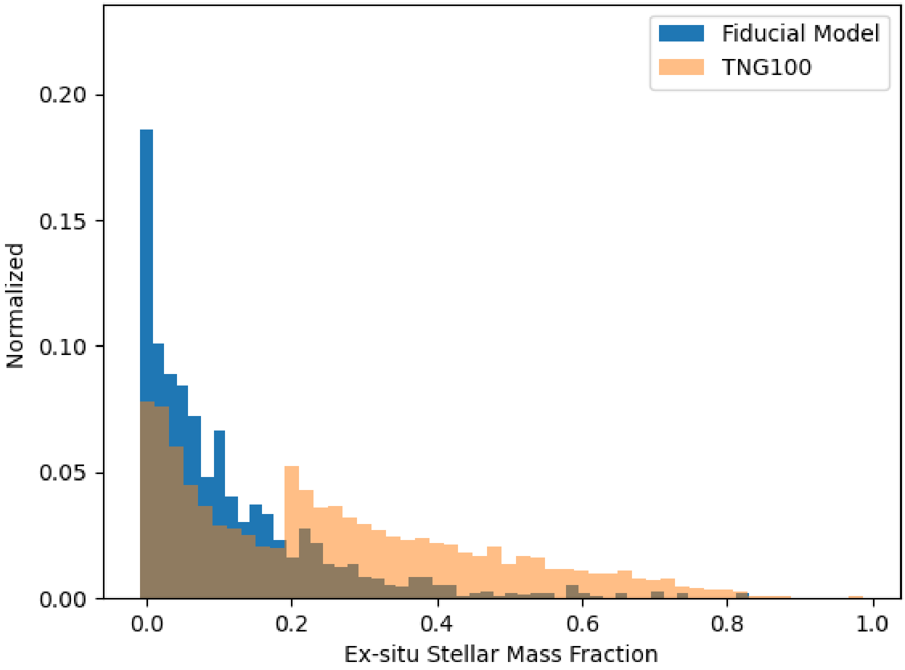

Figure 5: Histogram normalized representing the distribution of ex-situ mass fraction values for the galaxy samples used in this work from both simulations, TNG100 (blue) and the Fiducial Model from CAMELS (orange).

In Table(4) of the Appendix appears the name of the other datasets from CAMELS, which we use in this work, as well as the values of the parameters.

Observable integral properties

Besides of the ex-situ fraction, these simulations provide physical properties of galaxies as shown in Table(3), which we can use to estimate the ex-situ fraction.

Name	Unit	Description
Stellar Mass (SM)	$M_\odot$	Total mass of the stellar particles.
Gass Mass (GM)	$M_\odot$	Total mass of the gas particles.
Dark Matter Mass (DM)	$M_\odot$	Total mass of the dark matter particles.
Black Hole Mass (BHM)	$M_\odot$	Total mass of the black hole particles.
Half Mass Stellar Radius (HMSR)	kpc	Radius containing half of total stellar mass.
Spin (SP)	(kpc/h)(km/s)	Total spin per axis, computed for each the mass weighted sum of the relative coordinate times relative velocity of all member particles/cells.
Metallicity (MT)	-	The ratio Mz/Mt where Mz is the total mass of all metals elements (above He) and Mt is the total mass of elements. To convert in solar metallicity, divided by 0.0127 (the primordial solar metallicity). This value is used for all stellar particles to calculate the mass-weighted metallicity of the galaxy.
Age	Gyr	Mass-weighted mean age of each galaxy from the stellar particles contained within two times the stellar half-mass radius: $< t >_m = \frac{\sum_i^N m_i t_i}{\sum_i^N m_i}$ where N is the total number of stars, m_i and t_i is the mass and the age of each one.

Table 3: Short description of the different observational properties used to calculate the ex-situ stellar mass fraction[18].

In Figure (6), we illustrate the relationship between the properties shown in Table(3) and the ex-situ mass fraction. The blue contour represents the Fiducial Model from CAMELS, while the green contour represents the TNG100 dataset. In the first column, we observe the correlation between the ex-situ stellar mass fraction and the following properties: stellar mass, gas mass, dark matter mass, black hole mass, half-mass stellar ratio, spin, metallicity, and age. We see that the correlations are very similar for both simulations except for metallicity and HSMR. Consequently, to ensure good generalization, these parameters have not been considered in the model.

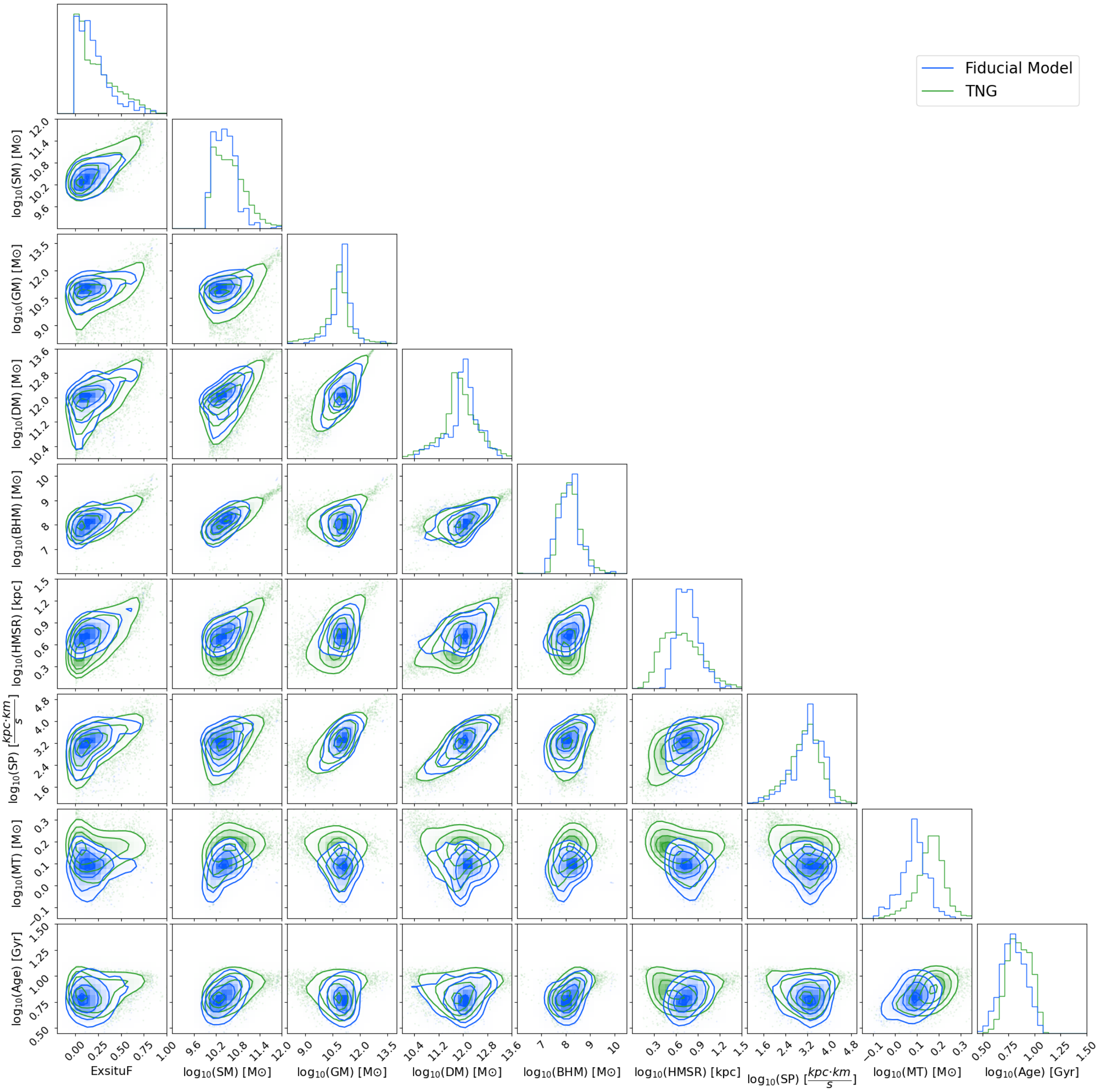

Figure 6: Correlation between different observational properties for the TNG (green contours) and Fiducial Model (blue contours). The diagonal shows the distribution of each property for both simulations.

4.4 High-Performance Symbolic Regression

One of the main issues with Neural Networks is the lack of interpretability. To overcome this we explore ways to obtain analytic expressions relating the ex-situ fraction to observable properties, as an alternative. To accomplish this, we employ High-Performance Symbolic Regression (PySR), which unlike traditional regression methods, allows us to derive complex non-linear relationships[19]. PySR is an open-source symbolic regression library, which falls under the category of machine learning[20], as it seeks to discover mathematical expressions that describe the relationship between input and output[21].

The following provides a summary of how this algorithm operates, but for more detailed information, refer to the publication 'Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl' by Miles Cranmer[20]:

1. **Ingredients of the Symbolic Regression**: this library provides us with a set of operators and functions such as addition, subtraction, exponentiation, logarithmic, etc. Using these elements, the algorithm generates a set of different random equations[22]. Besides, it has a fitness function or a function that evaluates how well each equation fits the data, defined as the difference between the true and the predicted value and a set of mutation operators or collections of different transformations that can be applied to the equations in the population (change a constant, function, etc.).

2. **Selection of equations**: the algorithm selects a small random group from this set, which competes with each other to determine the best candidate. Typically, this tournament involves a competition between two individuals.

3. **Evaluation of the fitness function**: each member of the small subset is evaluated using the fitness function, as mentioned in step 1, to provide us with information about how well it fits the data.

4. **Winner of the tournament**: this algorithm has an assigned probability of about 90 per cent to select the fittest individual. This is because we would otherwise always get the fittest candidate, and this adds a random factor so that the least qualified individuals can also be selected. Therefore, when the winner of the tournament is selected, a random number between 0 and 1 is generated, and if it is less or equal than the previously probability (in this case 0.9), they are selected as the winner. On the other hand, if this

number is bigger, the candidate is removed from the subset and it repeats step 3 with the remaining individuals. If there is only one candidate, it is automatically the winner.

5. **Mutation of the equation**: once the winner is chosen, a copy of it is created and a randomly-selected mutation is applied to it. These variations allow for finding better equations to fit.

6. **Replacement of the weakest**: the fitness function is applicated in the group of equations obtained in step 1 and the weakest, in other words, the one with the high score, is removed and replaced for the mutation equation obtained in step 5.

After this step, the process repeats until a certain number of iterations is reached to improve the equations of the solution. Then, the best solution is selected within the same complexity degree and printed in a hall of fame, i.e. the best equations are displayed. Finally, the best expression from this set is selected by balancing the score and the complexity. In Figure(7), we can see a summary picture of how this algorithm works. For more information in detail, see [20].

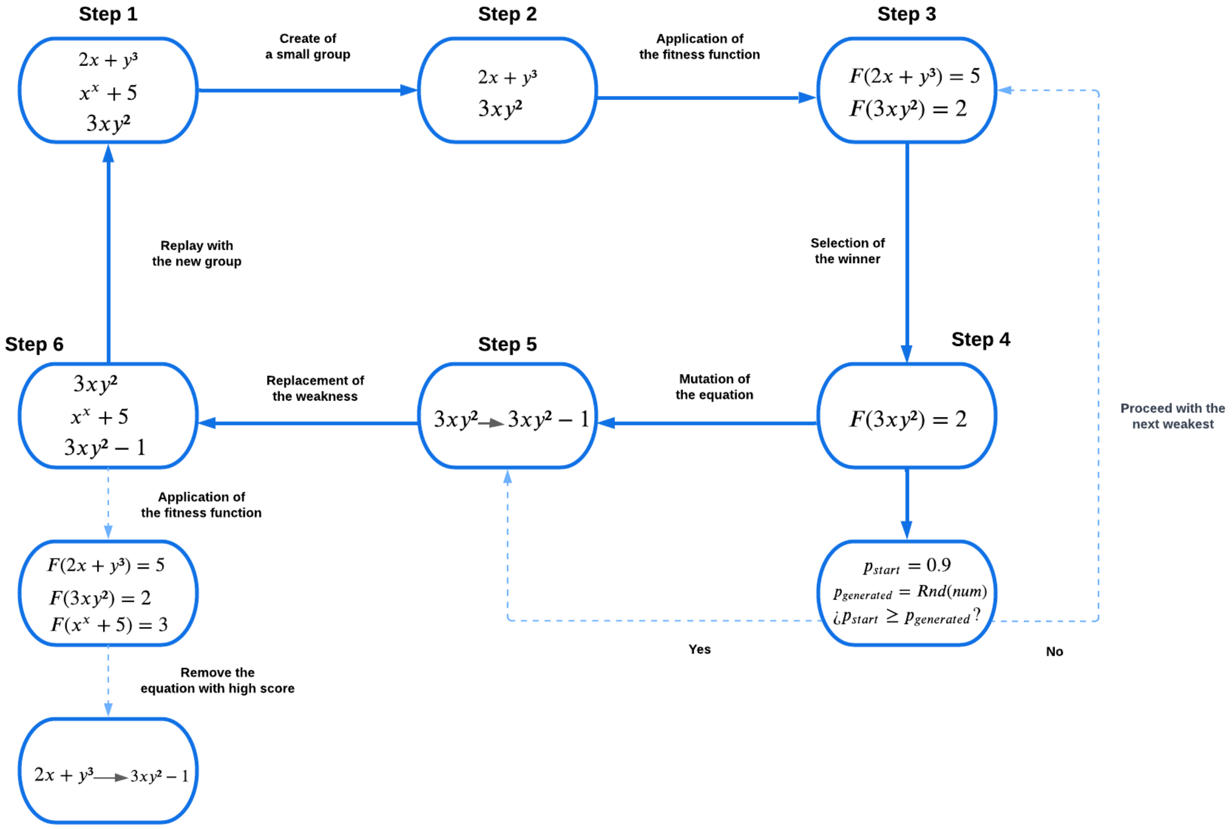

Figure 7: Scheme of how High-Performance Symbolic Regression works. It is a cyclical process that follows the direction of the thick blue arrows. The dashed arrows indicate alternative paths of the algorithm or an extension of how that step is performed.

5 Results and Discussion

5.1 Predictions Vs True Values

Once the adjustments described in Section(4.3) have been made, we train the model on the TNG100 training set and obtain the predictions both on the TNG100 and the CAMELS Fiducial Model test sets. These are plotted alongside the actual values to verify their proximity. This is demonstrated in Figure(8) for the TNG100 simulations and the Fiducial Model. The Fiducial simulation from the CAMELS project has been chosen to ensure that the results align with expectations before proceeding to the remaining datasets.

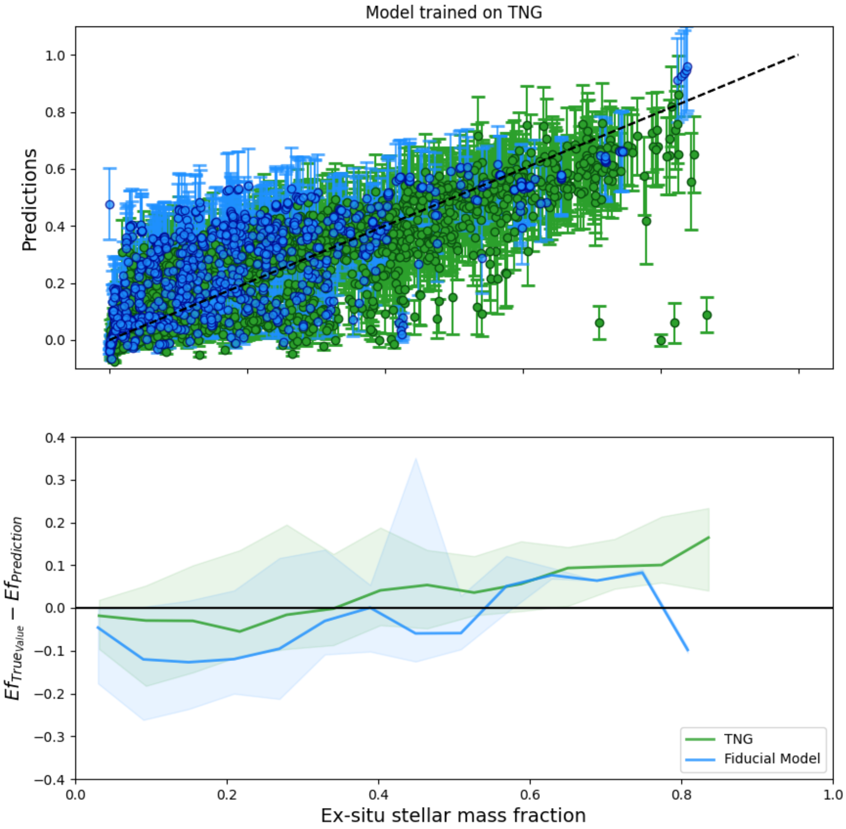

Figure 8: Evaluation on the Fiducial Model from CAMELS (blue) and the TNG100 test set (green). At the top, we show the predictions of the models versus the ground truth of the ex-situ stellar mass fraction. The black dashed line on the prediction panels marks the 1:1 relation and serves as a guide to the eye. The objective is to have the prediction points clustered as closely as possible to this line. At the bottom, the average error appears, where the shaded region contains the 68% of the data[1].

As shown in Figure (8), both plots follow the same trend, as expected, since the Fiducial Model shares the same cosmological and astrophysical parameters as TNG100 but with a smaller volume. It proves however that the resolution does not significantly impacts the results. The bottom panel of the figure shows the average difference between the true values and those predicted by the model. We notice some differences but both estimates are statistically compatible.

5.2 Comparison with the Fiducial Model

Once the baseline accuracy has been verified by predicting the ex-situ fraction using the Fiducial Model, we now explore how the predictions depend on cosmological and astrophyisical parameters. Figure (9) illustrates the average errors calculated as the differences between the predictions obtained for the Fiducial Model and for simulations with different cosmological and astrophysical parameters. Two plots have been generated for each parameter to enhance line visibility. The legend displays the value of each parameter, with the black line representing the value of the Fiducial Model.

In Figure (9), it can be observed that, for the majority of cases, the errors fall within the range of (-0.3, 0.3). However, there are a few instances where the behavior deviates slightly, such as in the case of $\Omega = 0.1000$ or $\mathrm{A}_{SN2} = 0.5000$, among others. This indicates that these particular values may not perform well in the machine learning model being used. In addition to analyzing the difference in prediction values, a plot has been generated to illustrate the discrepancy in average standard deviation values. This statistical tool provides us with a better understanding of the data behavior, as two lines may have similar average values but different standard deviations. We observe that the standard deviations are essentially consistent across datasets. Both figures therefore suggest a good generalization of the Neural Network model, which seems to be able to marginalize over astrophysics and cosmology except for some extreme cases.

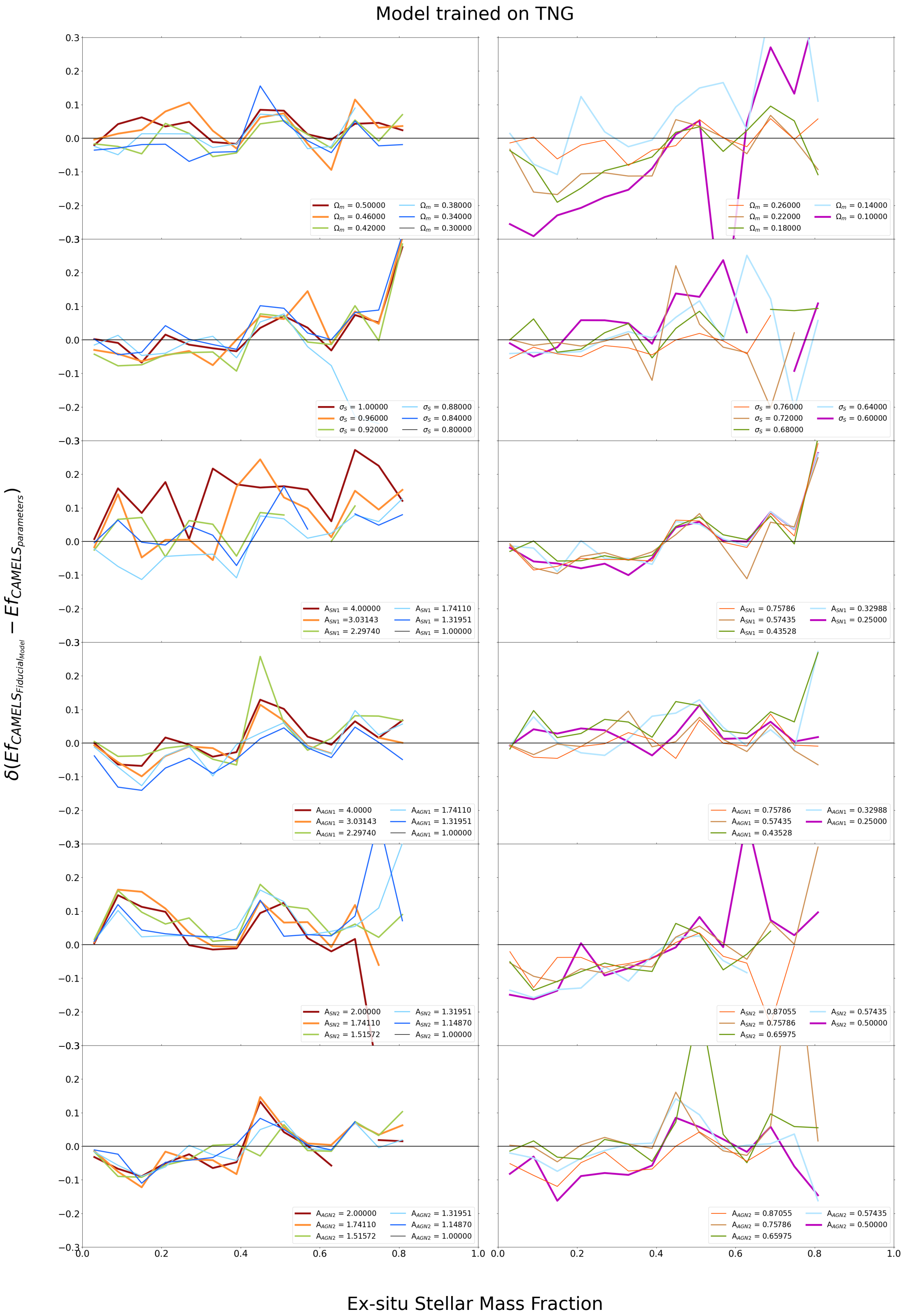

Figure 9: Error plot obtained as the difference between the prediction of the Fiducial Model and the other simulations provided by CAMELS. In the legend, the value of the astrophysical or cosmological parameter is displayed, with the black line corresponding to the Fiducial Model.

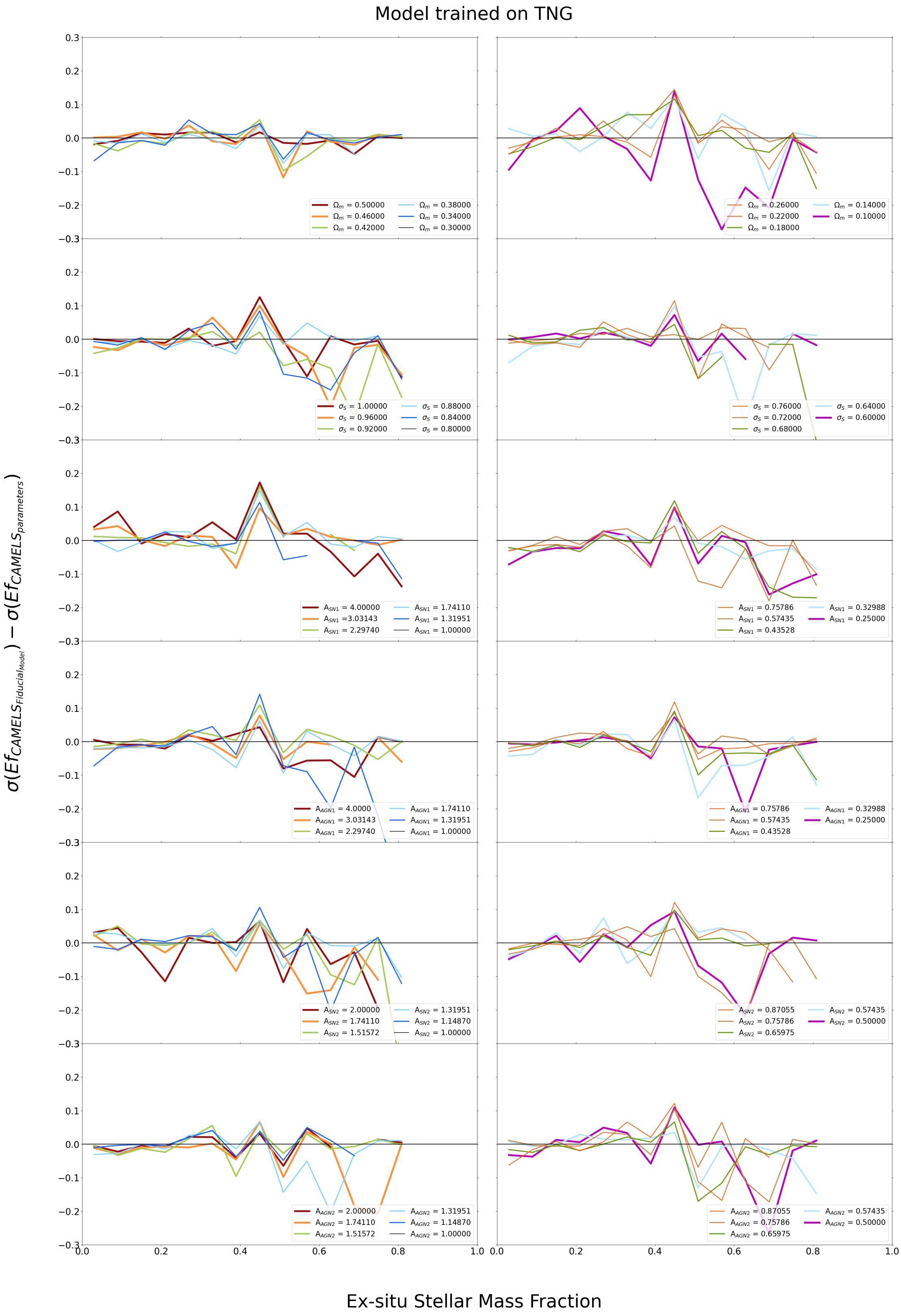

Figure 10: Differences in the standard deviation media of the prediction between the Fiducial Model and the other datasets provided by CAMELS are depicted. Similar to the previous plot, the legend provides the values of the different astrophysical or cosmological variables, with the black line indicating the value of the corresponding constant for the Fiducial Model.

This might appear surprising in the first place. In order to better understand the behavior, we explore the correlations between the different galaxy properties used in the model and the ex-situ stellar mass fraction in thedifferent CAMELS models. To verify this, corner plots are performed, which allow us to see how properties are related to each other. Figure(11) shows an example for the case of $\Omega = 0.1000$.

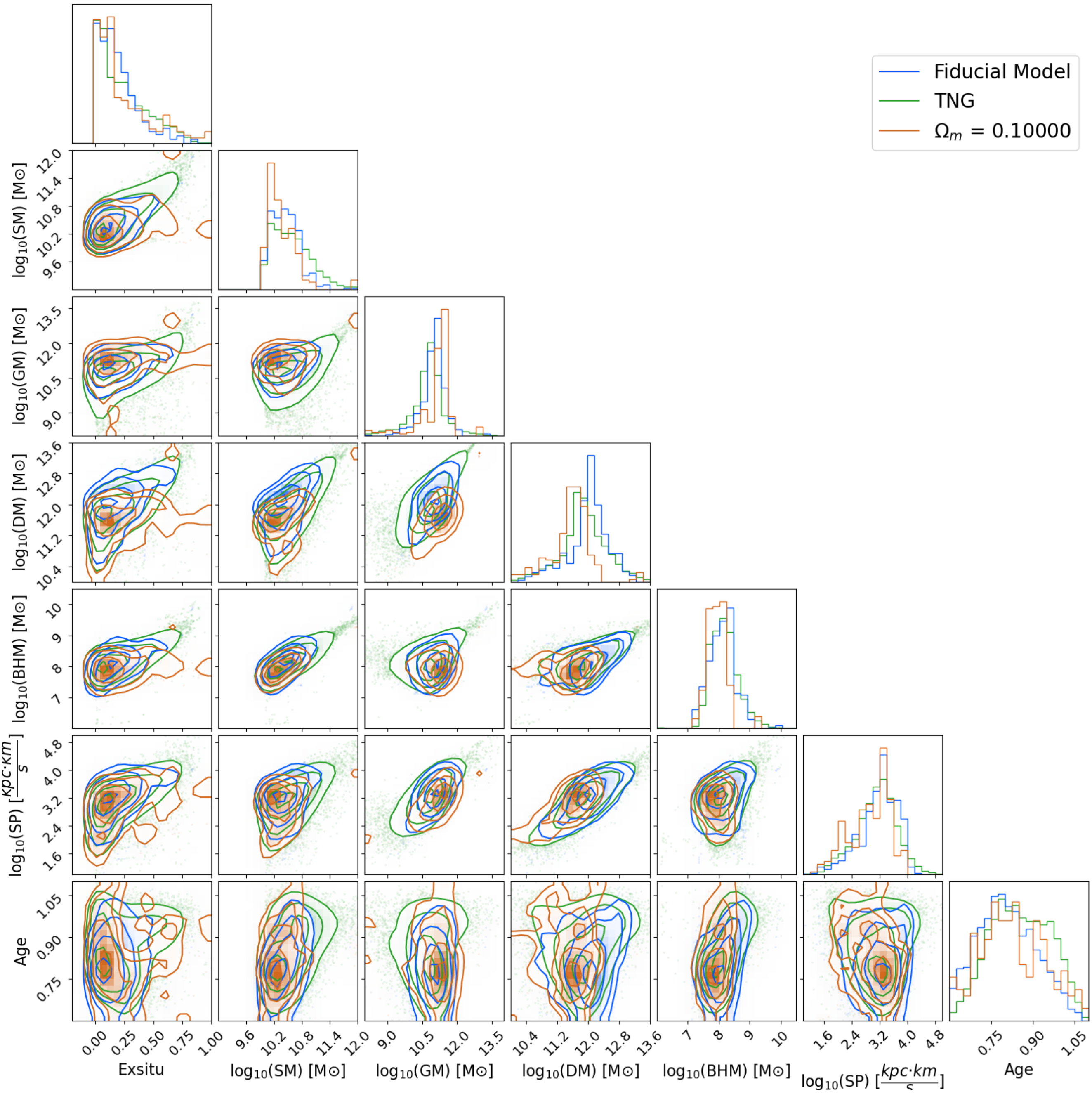

Figure 11: Correlation of the different galaxy properties for the Fiducial Model (blue contour), TNG (green contour), and the CAMELS simulation where the omega parameter is varied (orange contour).

As it can be observed, the relations between the ex-situ stellar mass fraction and the observable integrated parameters acquired from the CAMELS simulation with $\Omega = 0.1000$ show differences from the ones acquired from the TNG100 and the CAMELS Fiducial Model. This behaviour can affect the accuracy of our model's predictions. Therefore, we consider this to be one of the reasons why there is such a difference between the estimated prediction for the Fiducial Model and the predictions obtained for some of the variations of the cosmological and astrophysical parameters. However, this only happens in some cases and the errors are not very significant, thus confirming the generality of the model.

5.3 High-Performance Symbolic Regression

As mentioned in Section(4.4), the objective of this library is to derive an equation that enables the calculation of the ex-situ fraction using observational parameters. In our case, we use the data from TNG100 to derive an analytical expression, and we obtain the residuals when applying the resulting equations to the Fiducial CAMELS model. Figure (12) illustrates the discrepancy between the true value of the Fiducial Model and the value obtained from the equations generated by this algorithm.

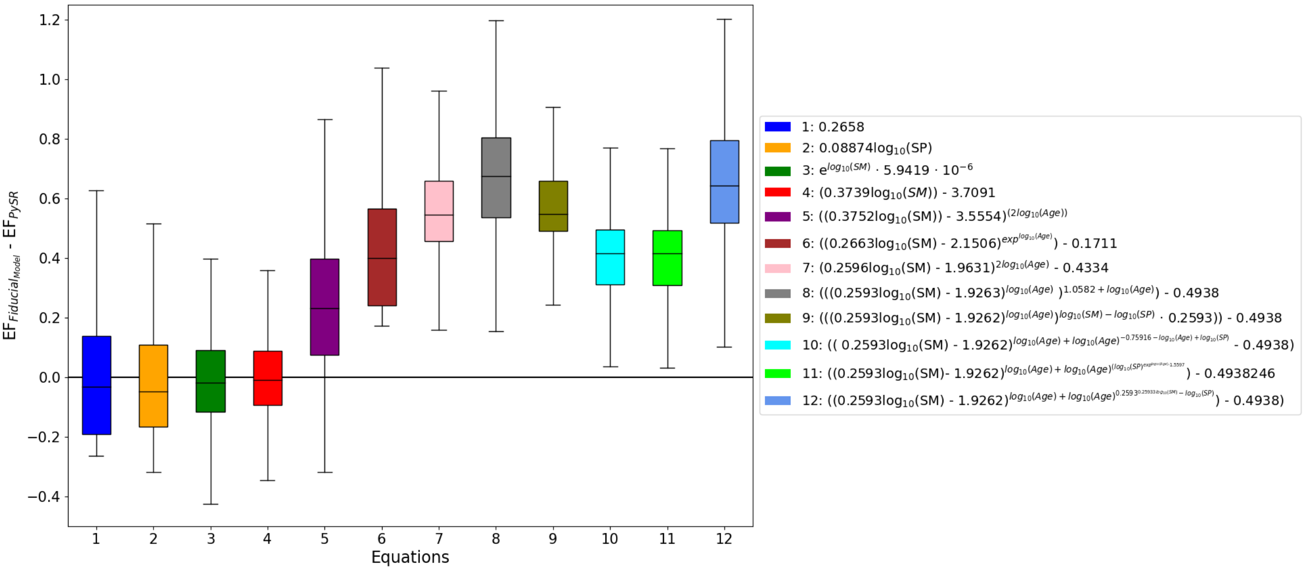

Figure 12: Difference between the true value of the Fiducial Model and the estimated value obtained from the equations of the TNG100. The black line dividing the box represents the median, which indicates the location of the 50% of the data. The lower box or first quartile represents the 25% of the data, while the upper box or third quartile represents the 75% of the data. The extreme values are represented by long lines that extend beyond the box and are commonly referred to as whiskers.

As we can see, this library provides us with different equations, ranging from more to less levels of complexity. Due to its simplicity and lower error compared to the others, we selected the Equation 4 as the candidate for calculating the ex-situ fraction using observational parameters. It is shown as follow

$$EF = 0.3739 \cdot log_{10}(SM) - 3.7091, \tag{6}$$

where EF is the ex-situ fraction and SM is the Stellar Mass.

The fact that the equation mainly depends on stellar mass, which is also present in almost all the analytic expressions obtained, is justified by the fact that more massive galaxies have generally undergone more mergers with other galaxies, and therefore, a significant portion of their mass is attributed to these mergers.

5.4 Comparison between the Neural Network and the Symbolic Regression

Once the equation has been obtained through High-Performance Symbolic Regression, it is compared with the results obtained from the machine learning model. In the top of the Figure(13), the ex-situ mass fraction is plotted against the stellar mass to visualize the comparison.

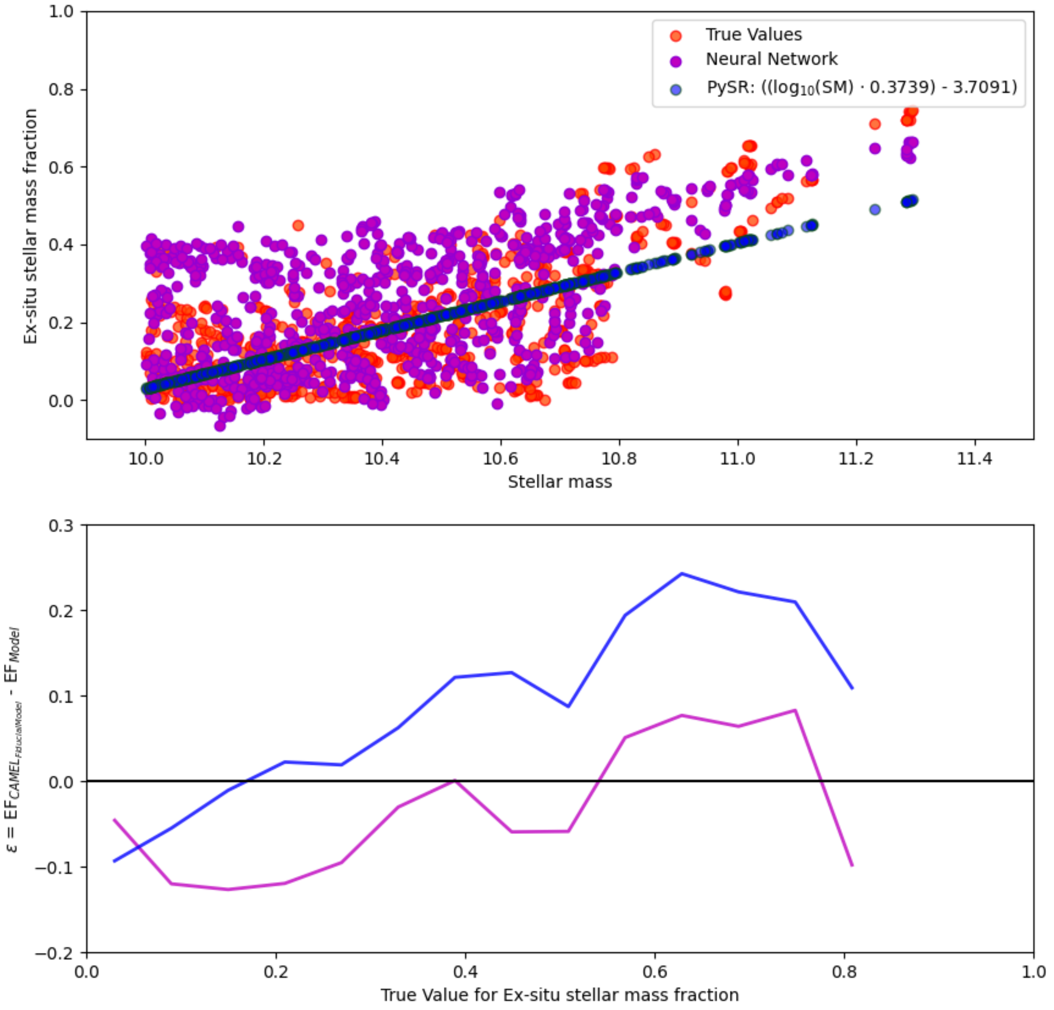

Figure 13: Comparison of the results of the ex-situ fraction obtained from the Machine Learning model (purple) and Symbolic Regression (blue), along with the True Values (reddish pink), in the upper graph. In the lower graph, the errors are displayed as the average differences between the true value of the Fiducial Model and the calculated value obtained from both methods.

As observed in the upper part of the plot, the predictions generated by the Machine Learning model effectively capture the scattering of the real values within the simulation, in contrast to the equation derived from Symbolic Regression. The latter, as it is a linear expression dependent on mass, is hence represented as a straight line with a distinct slope.

Regarding the lower part, the average error is displayed as the difference between the real value of the Fiducial Model simulation and the estimated value obtained from the two methods used to calculate the ex-situ mass fraction. The Neural Network, represented by a purple line, is closer to the real(black line) value than the equation represented by a blue line, which indicates that the first is more accurate than the second.

In certain cases, it is indeed possible for a Machine Learning model to make better predictions due to its methodology, which involves learning patterns through training. As a result, it can adapt better to non-linear behaviours. On the other hand, Symbolic Regression, although it is a useful tool capable of finding an analytical expression of the data, also has its limitations when it has to find the most suitable function, especially when the relationships are complex and non-linear.

6 Conclusion

This study utilized a machine learning model trained on the TNG100 simulation to assess its generalization capabilities using simulations provided by CAMELS, specifically the IllustrisTNG suite, which alters various astrophysical and cosmological parameters compared to the training simulation.

To improve prediction accuracy, a corner plot was generated, representing both TNG100 and the Fiducial Model, which, as mentioned earlier, is a simulation with the same initial conditions as TNG100 but at a smaller scale. This analysis revealed a poor correlation between the half mass stellar radius and the metallicity with the ex-situ stellar mass fraction. Consequently, these variables were not considered during model training to avoid compromising accuracy.

Then, the Neural Network was trained using TNG100, and predictions were calculated for both TNG100 and the CAMELS Fiducial Model. Subsequently, the predictions of the simulations were presented based on the real value of the ex-situ fraction. As observed, both simulations follow the same trend but exhibit different errors, which were calculated as the difference between the real and the predicted values. This difference in error was justified by the disparity in resolution between the two simulations..

Furthermore, other simulations from CAMELS were included, varying the cosmological parameters compared with TNG100. To compare these simulations with the Fiducial Model, two different plots were generated. The first plot depicted the average error, calculated as the difference between the predictions of the Fiducial Model and the model's predictions for different magnitudes. The second plot considered the mean standard deviation using the same calculation. It was observed how the model performed well for the majority of cases, and only in some, mostly those with more extreme values, significant errors were obtained compared to the overall trend. One possible explanation found was that the correlation between the observational properties considered during model training, as observed in the corner plot, and the ex-situ fraction may not be strong, leading to imprecise predictions..

In conclusion, this Neural Network model exhibits errors primarily due to resolution and performs well for certain changes in cosmological and astrophysical variables. However, the errors are not very significant, thus confirming the generality of the model.

To explore alternative methods for calculating this magnitude, High-Performance Symbolic Regression was employed, allowing the derivation of an analytical expression that relates observational parameters, such as the age or the stellar mass, with the ex-situ mass fraction. Using TNG100, a set of equations was derived and represented on a graph, as a function of the error calculated as the difference between the actual value of the Fiducial Model and the value obtained from the analytical expression. The equation with the lowest error was selected as the candidate for calculating the ex-situ fraction using observational parameters. This analysis revealed a dependence on stellar mass, which can be justified by the fact that a significant portion of the mass in highly massive galaxies is typically attributed to mergers with other galaxies.

A comparison was made between the values obtained using Machine Learning and Symbolic Regression in a plot. The scatter plot demonstrated better representation for the Neural Network compared to the expression, which appeared as a linear relationship with a certain slope. Additionally, the error was presented, calculated as the average difference between the real and predicted values using both methods. The results indicate that the model performed better in calculating the ex-situ fraction compared to Symbolic Regression, although the latter did not yield significantly larger errors compared to the former. Therefore, it can be concluded that the model used provides greater precision for calculating this magnitude, while the obtained expression can provide an initial approximation of its value.

As future work, it would be valuable to further investigate why the model fails for certain values of cosmological and astrophysical parameters in order to contribute to its improvement. Additionally, combining this model with High-Performance Symbolic Regression could lead to the derivation of more complex equations for calculating this magnitude.

Bibliography

[1] Eirini Angeloudi, Jesú s Falcón-Barroso, Marc Huertas-Company, Regina Sarmiento, Annalisa Pillepich, Daniel Walo-Martín, and Lukas Eisert. ERGO-ML: Towards a robust machine learning model for inferring the fraction of accreted stars in galaxies from integral-field spectroscopic maps. *Monthly Notices of the Royal Astronomical Society*, jun 2023. doi: 10.1093/mnras/stad1669. URL `https://doi.org/10.1093%2Fmnras%2Fstad1669`.

[2] Andrea Cimatti, Filippo Fraternali, and Carlo Nipoti. *Introduction to Galaxy Formation and Evolution: From Primordial Gas to Present-Day Galaxies.* Cambridge University Press, 2017.

[3] Alina Boecker, Nadine Neumayer, Annalisa Pillepich, Neige Frankel, Rahul Ramesh, Ryan Leaman, and Lars Hernquist. The origin of stars in the inner 500 parsecs in TNG50 galaxies. *Monthly Notices of the Royal Astronomical Society*, 519(4):5202–5235, dec 2022. doi: 10.1093/mnras/stac3759. URL `https://doi.org/10.1093%2Fmnras%2Fstac3759`.

[4] Ludwig Oser, Jeremiah P. Ostriker, Thorsten Naab, Peter H. Johansson, and Andreas Burkert. The two phases of galaxy formation. *The Astrophysical Journal*, 725(2):2312–2328, 2010. doi: 10.1088/0004-637X/725/2/2312.

[5] Christopher C Lovell, Viviana Acquaviva, Peter A Thomas, Kartheik G Iyer, Eric Gawiser, and Stephen M Wilkins. Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 490(4):5503–5520, 2019. doi: 10.1093/mnras/stz2849. URL `https://doi.org/10.1093/mnras/stz2849`.

[6] The illustristng project. `https://www.tng-project.org/about/`.

[7] Dylan Nelson. The illustristng simulations: Public data release. https://arxiv.org/abs/1812.05609, 2021.

[8] CAMELS team. Camels release 0.1, May 2023. URL https://camels.readthedocs.io/en/latest/description.html. Software release.

[9] Jordi Cepa. *Cosmología física*. Editorial UOC, Barcelona, España, 2^{a} ed. edition, 2015. ISBN 9788490645035.

[10] Vicente Rodriguez-Gomez, Annalisa Pillepich, Laura V. Sales, Shy Genel, Mark Vogelsberger, Qirong Zhu, Sarah Wellons, Dylan Nelson, Paul Torrey, Volker Springel, Chung-Pei Ma, and Lars Hernquist. The stellar mass assembly of galaxies in the illustris simulation: growth by mergers and the spatial distribution of accreted stars. *Monthly Notices of the Royal Astronomical Society*, 458(3):2371–2390, feb 2016. doi: 10.1093/mnras/stw456. URL https://doi.org/10.1093%2Fmnras%2Fstw456.

[11] Author's Name. The abstract for the article. *The Astrophysical Journal*, 292:371, 1985. doi: 10.1086/163168. URL https://ui.adsabs.harvard.edu/abs/1985ApJ...292..371D/abstract.

[12] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies – i. results at z = 0. *Monthly Notices of the Royal Astronomical Society*, 328:726–750, 2001. doi: 10.1046/j.1365-8711.2001.04912.x. URL https://doi.org/10.1046/j.1365-8711.2001.04912.x.

[13] K. Dolag, S. Borgani, G. Murante, and V. Springel. Substructures in hydrodynamical cluster simulations. *Monthly Notices of the Royal Astronomical Society*, 399:497–514, 2009. doi: 10.1111/j.1365-2966.2009.15034.x. URL https://doi.org/10.1111/j.1365-2966.2009.15034.x.

[14] Vicente Rodriguez-Gomez, Shy Genel, Mark Vogelsberger, Annalisa Pillepich, Laura V. Sales, Debora Sijacki, Paul Torrey, Gregory F. Snyder, Dylan Nelson, and Volker Springel. The merger rate of galaxies in the illustris simulation: a comparison with observations and semi-empirical models. *Monthly Notices of the Royal Astronomical Society*, 449(1):49–65,

2015. doi: 10.1093/mnras/stv147. URL https://academic.oup.com/mnras/article/449/1/49/1099789.

[15] Vicente Rodriguez-Gomez, Laura V. Sales, Shy Genel, Annalisa Pillepich, Jolanta Zjupa, Dylan Nelson, Brendan Griffen, Paul Torrey, Gregory F. Snyder, and Mark Vogelsberger. The role of mergers and halo spin in shaping galaxy morphology. *Monthly Notices of the Royal Astronomical Society*, 467(3):3083–3098, 2017. doi: 10.1093/mnras/stx305. URL https://doi.org/10.1093/mnras/stx305.

[16] Androbomb. Simple nn with python multi-layer perceptron. https://www.kaggle.com/code/androbomb/simple-nn-with-python-multi-layer-perceptron, Year of publication or last update.

[17] Marc Huertas-Company. PART II: FOUNDATIONS OF 'SHALLOW' NEURAL NETWORKS. PowerPoint presentation.

[18] TNG Collaboration. Tng project data specifications. https://www.tng-project.org/data/docs/specifications/, n.d.

[19] MilesCranmer and PySR contributors. Pysr github repository. https://github.com/MilesCranmer/pySR, 2022.

[20] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023.

[21] Zi Wong, Robert Zinkov, and Miles Cranmer. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems*, 2020.

[22] PySR. Pysr documentation. https://pysr.readthedocs.io/, 2022.

Appendix

Name	Ω_m	σ_s	A_{SN1}	A_{AGN1}	A_{SN2}	A_{AGN2}
1P_1_n5	0.10000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_n4	0.14000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_n3	0.18000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_n2	0.22000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_n1	0.26000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_1	0.34000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_2	0.38000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_3	0.42000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_4	0.46000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_1_5	0.50000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_2_n5	0.30000	0.60000	1.00000	1.00000	1.00000	1.00000
1P_2_n4	0.30000	0.64000	1.00000	1.00000	1.00000	1.00000
1P_2_n3	0.30000	0.68000	1.00000	1.00000	1.00000	1.00000
1P_2_n2	0.30000	0.72000	1.00000	1.00000	1.00000	1.00000
1P_2_n1	0.30000	0.76000	1.00000	1.00000	1.00000	1.00000
1P_2_1	0.30000	0.84000	1.00000	1.00000	1.00000	1.00000
1P_2_2	0.30000	0.88000	1.00000	1.00000	1.00000	1.00000
1P_2_3	0.30000	0.92000	1.00000	1.00000	1.00000	1.00000
1P_2_4	0.30000	0.96000	1.00000	1.00000	1.00000	1.00000
1P_2_5	0.30000	1.00000	1.00000	1.00000	1.00000	1.00000
1P_3_n5	0.30000	0.80000	0.25000	1.00000	1.00000	1.00000
1P_3_n4	0.30000	0.80000	0.32988	1.00000	1.00000	1.00000
1P_3_n3	0.30000	0.80000	0.43528	1.00000	1.00000	1.00000
1P_3_n2	0.30000	0.80000	0.57435	1.00000	1.00000	1.00000
1P_3_n1	0.30000	0.80000	0.75786	1.00000	1.00000	1.00000
1P_3_1	0.30000	0.80000	1.31951	1.00000	1.00000	1.00000
1P_3_2	0.30000	0.80000	1.74110	1.00000	1.00000	1.00000
1P_3_3	0.30000	0.80000	2.29740	1.00000	1.00000	1.00000
1P_3_4	0.30000	0.80000	3.03143	1.00000	1.00000	1.00000
1P_3_5	0.30000	0.80000	4.00000	1.00000	1.00000	1.00000
1P_4_n5	0.30000	0.80000	1.00000	0.25000	1.00000	1.00000
1P_4_n4	0.30000	0.80000	1.00000	0.32988	1.00000	1.00000
1P_4_n3	0.30000	0.80000	1.00000	0.43528	1.00000	1.00000
1P_4_n2	0.30000	0.80000	1.00000	0.57435	1.00000	1.00000
1P_4_n1	0.30000	0.80000	1.00000	0.75786	1.00000	1.00000
1P_4_1	0.30000	0.80000	1.00000	1.31951	1.00000	1.00000

1P_4_2	0.30000	0.80000	1.00000	1.74110	1.00000	1.00000
1P_4_3	0.30000	0.80000	1.00000	2.29740	1.00000	1.00000
1P_4_4	0.30000	0.80000	1.00000	3.03143	1.00000	1.00000
1P_4_5	0.30000	0.80000	1.00000	4.00000	1.00000	1.00000
1P_5_n5	0.30000	0.80000	1.00000	1.00000	0.50000	1.00000
1P_5_n4	0.30000	0.80000	1.00000	1.00000	0.57435	1.00000
1P_5_n3	0.30000	0.80000	1.00000	1.00000	0.65975	1.00000
1P_5_n2	0.30000	0.80000	1.00000	1.00000	0.75786	1.00000
1P_5_n1	0.30000	0.80000	1.00000	1.00000	0.87055	1.00000
1P_5_0	0.30000	0.80000	1.00000	1.00000	1.00000	1.00000
1P_5_1	0.30000	0.80000	1.00000	1.00000	1.14870	1.00000
1P_5_2	0.30000	0.80000	1.00000	1.00000	1.31951	1.00000
1P_5_3	0.30000	0.80000	1.00000	1.00000	1.51572	1.00000
1P_5_4	0.30000	0.80000	1.00000	1.00000	1.74110	1.00000
1P_5_5	0.30000	0.80000	1.00000	1.00000	2.00000	1.00000
1P_6_n5	0.30000	0.80000	1.00000	1.00000	1.00000	0.50000
1P_6_n4	0.30000	0.80000	1.00000	1.00000	1.00000	0.57435
1P_6_n3	0.30000	0.80000	1.00000	1.00000	1.00000	0.65975
1P_6_n2	0.30000	0.80000	1.00000	1.00000	1.00000	0.75786
1P_6_n1	0.30000	0.80000	1.00000	1.00000	1.00000	0.87055
1P_6_1	0.30000	0.80000	1.00000	1.00000	1.00000	1.14870
1P_6_2	0.30000	0.80000	1.00000	1.00000	1.00000	1.31951
1P_6_3	0.30000	0.80000	1.00000	1.00000	1.00000	1.51572
1P_6_4	0.30000	0.80000	1.00000	1.00000	1.00000	1.74110
1P_6_5	0.30000	0.80000	1.00000	1.00000	1.00000	2.00000

Table 4: Table with the values of the parameters for the 1P CAMELS simulations in the IllustrisTNG suite.

The names of these simulations follow this format 1P_X_Y, where X varies from 1 to 6 and represents the considered parameter: 1(Ω_m), 2(σ_s), 3(A_{SN1}), 4(A_{AGN1}), 5(A_{SN2}), and 6(A_{AGN2}). In the case of Y, it indicates the values taken from n5 to 5, with the first being the smallest and the latter being the largest values for the same X parameter [8].