Master thesis

Digging out the debris of the Milky Way past accretion events with Machine Learning

Amanda Aguiar Álvarez

Supervisor: Guillaume Thomas
Co-supervisor: Giuseppina Battaglia

Resumen

Nuestra posición privilegiada en el Universo hace de la Vía Láctea el laboratorio perfecto para entender los mecanismos físicos que llevan a la formación de sus diferentes estructuras. En las últimas décadas, estos estudios se han visto impulsados debido a la mejora en la calidad de los datos, gracias a proyectos como el Sloan Digital Sky Survey, que permite estudiar el desplazamiento al rojo espectroscópico para un gran número de estrellas y tomar imágenes multiespectrales, o misiones como Gaia, que proporciona un catálogo de datos astronómicos con precisiones sin precedentes. Asimismo, es de gran relevancia la mejora en la capacidad computacional, que ha impulsado el desarollo de simulaciones cosmológicas.

Por otro lado, el paradigma estándar de la cosmología, Lambda cold dark matter (ΛCDM), indica que las galaxias de menor tamaño son las primeras en formarse y que las galaxias mayores, como la Vía Láctea, son el resultado de procesos de acreción y fusión de galaxias de menor tamaño, junto a la acreción del gas. Estos procesos de acreción y fusión de galaxias dejan marcas observables en la actualidad y que esperan encontrarse, esencialmente, en el espacio de las integrales de movimiento de las estrellas del halo en forma de cúmulos. No obstante, existen varios procesos, como la fricción dinámica o el aumento de la masa de la Vía Láctea con el tiempo, que hacen que estas cantidades no se conserven en su totalidad.

El objetivo del presente trabajo es desentrañar la historia del halo estelar de la Vía Láctea mediante la identificación de estos cúmulos en el espacio de fases, haciendo uso para ello de técnicas de Machine Learning no supervisado. Específicamente, se ha recurrido a un modelo de mezcla Gaussiana (Gaussian Mixture) tras comprobarse que, de entre los métodos considerados, es el que conduce a la mejor identificación de las diferentes sobre-densidades como grupos independientes. Este modelo se basa en la probabilidad de que un cierto punto pertenezca a una distribución en forma de Gaussiana multi-dimensional y permite obtener sus parámetros característicos (pesos, valores medios y matrices de covarianza), los cuales son iniciados haciendo uso del método de Machine Learning conocido como K-Means. Concretamente, se emplea el método de la Bayesian Gaussian Mixture, que emplea la regla de Bayes para encontrar el número adecuado de cúmulos dado un un límite superior en el número de componentes que puede determinar. A su vez, en este modelo se emplea una asignación a priori de las probabilidades asociadas a cada uno de los componentes mediante el llamado proceso de Dirichlet. Posteriormente, el modelo óptimo es encontrado a través del algoritmo de esperanza-maximización. Se utilizan valores como el Bayesian Information Criterion (BIC) o la log-likelihood para poder comparar los diferentes modelos.

El primer paso necesario para desarrollar este método ha sido la familiarización con esta técnica haciendo uso de conjuntos de datos controlados; concretamente, de datos generados mediante Gaussianas cuyos parámetros son conocidos. De este modo, se aprecia el efecto que tiene la variación de los diferentes parámetros de entrada que requiere el método, así como sus limitaciones. Esto ha permitido concluir que, efectivamente, es posible recuperar los puntos generados por las diferentes Gaussianas como cúmulos independientes mediante el modelo de Bayesian Gaussian Mixture.

El siguiente paso ha sido implementar estos métodos para trabajar con halos simulados en el paradigma ΛCDM de la colaboración Auriga, correspondientes a simulaciones magneto hidrodinámicas de alta resolución de galaxias análogas a la Vía Láctea. En este caso, las partículas de estrellas cuentan con una etiqueta que indica su origen (como podría ser una galaxia pequeña acretada, a la que nos referiremos como su progenitor), de modo que es posible comparar lo obtenido con los modelos de Bayesian Gaussian Mixture con los resultados que serían esperables.

A continuación, con el fin de familiarizarnos con los datos de las simulaciones, se ha empezando haciendo una inspección visual del espacio constituido por la energía total y el momento angular en torno al eje perpen-

dicular al plano del disco de las partículas pertenecientes al halo estelar, ampliamente usados para estudiar los procesos de acreción/fusión en la Vía Láctea, para diferentes rangos de radio en torno al centro Galáctico y diferentes rangos de metalicidad total. Seguidamente, se han visualizado diferentes espacios de las cantidades en las que se espera encontrar sobre-densidades asociadas a cada progenitor, es decir, a cada galaxia satélite a la que pertenecían las estrellas antes de que los procesos de acreción tuviesen lugar. Esto ha demostrado la dificultad intrínseca de la tarea que se pretende realizar, debido a la superposición existente entre galaxias satélite en los espacios considerados y al hecho de que un progenitor no se asocia a una única sobre-densidad. Esto último lleva, además, a que no es posible recuperar cada uno de los progenitores como una única Gaussiana. No obstante, este efecto es más importante en el caso de los progenitores más masivos.

Posteriormente, se ha buscado el conjunto de estrellas compuesto por los 4 progenitores más masivos en el rango de radios y metalicidades en el que estos se distinguen con mayor facilidad en el espacio constituido por la energía total y el momento angular a lo largo del eje perpendicular al plano del disco. Luego, se ha aplicado en este subconjunto de datos el método de Bayesian Gaussian Mixture en los diferentes espacios en los que se espera que las estrellas que perteneciesen a un mismo progenitor aparezcan como cúmulos, obteniéndose resultados muy semejantes. En consecuencia, se ha decidido centrar la atención en el espacio constituido por la energía total y el momento angular vertical, junto al momento angular perpendicular, por ser más fáciles de interpretar.

Una vez hecho esto, se ha aplicado la Bayesian Gaussian Mixture en el rango completo de radios y metalicidades a los datos correspondientes a los 4 progenitores más masivos, así como a otros 4 cuyas masas se encuentran en un rango intermedio. De este modo, se han identificado las diferentes sobre-densidades como múltiples Gaussianas independientes, si bien no se ha establecido aún ningún enlace entre ellas y las galaxias satélites originales. Por este motivo, a continuación, se ha procedido a intentar relacionar las differentes Gaussianas haciendo uso de las distancias de Mahalanobis y del método de enlace pesado de forma jerárquica. Se ha llegado con esto a que, si bien no ha sido posible determinar el origen de las sobre-densidades en el espacio de integrales de movimiento al relaciolarlas entre ellas para progenitores de mayor masa, esta sí es una opción viable en rangos de masa menores.

Con esto, se llega a que sería necesario desarrollar un método alternativo que permita estudiar los progenitores más pesados, para luego poder estudiar únicamente los de menor masa y aplicar métodos de Machine Learning de agrupamiento, junto a métodos de enlace, para así identificar los cúmulos restantes. Asimismo, sería de interés hacer otro tipo de pruebas con simulaciones con datos más realistas, es decir, con una mayor semejanza con los datos observacionales, así como realizar un estudio más profundo de la información que puede extraerse de las metalicidades, con el fin de cumplir el objetivo presentado.

Table of contents

1. Introduction

Since the beginning of time, the human race has admired the night sky, with a white, diffuse glowing band that stretches across the sky being one of its most noticeable features. This magnificent spectacle led to our ancient forebears to often perceive it as some sort of path: early Hindus considered it the way that guided the god Aryaman to his throne, early Nordics saw it as the road to Valhalla and Iroquois found in it the path that would lead them to eternal life. Nevertheless, we usually refer to this river of light as the Milky Way (from the Latin word *Via Lactea*), due to the Greek myth that says that this is the milk of the goddess Hera, spilled by the demigod Hercules (Tom Burns 2021). This road is our own Galaxy, the only one from which we can extract information such as close-ups of its structures and contents in great detail, which motivates our efforts to try to understand its features.

Fig. 1: Panoramic view of the Milky Way above the Gran Telescopio de Canarias (GTC) (Figure credit: Daniel López 2019).

Several new astronomical surveys have emerged during the last decade and have led to large improvements in our ability to study our home Galaxy. A few examples are the European *Gaia* space telescope, intended to obtain the largest existing astrometry catalogue with an unprecedented quality, the incoming *WEAVE* survey, a multi-object spectrograph to be used on the William Herschel Telescope of the Roque de los Muchachos observatory, one of the state-of-the-art facilities, or the fifth Sloan Digital Sky Survey (*SDSS-V*) a multi-epoch spectroscopic survey. This large amount of new data is expected to allow us to use the Milky Way, a rather common type of galaxy, as a cosmological laboratory that provides us with deeper information, for example, regarding the stellar populations and the internal mechanisms of galaxies, in comparison with what has been available in the past.

One of the essential objects of interest is the formation history of the Milky Way, since it could grant us information regarding its own evolution but also about the evolution of other galaxies and, ultimately, of the Universe itself. According to the Lambda-Cold Dark Matter (ΛCDM) paradigm, galaxies grow in a hierarchical way; that is, low-mass systems would have been the first to be formed and would have eventually merged into larger systems through the accretion/merging of smaller building blocks and gas accretion (White & Rees 1978). Even if the debris of these events is not spatially coherent anymore, their imprints are preserved over several Gyr in dynamical and chemical spaces and they are expected to be detected, in principle, by the current and upcoming astrometric missions (Helmi et al. 1999).

The Milky Way evidences bottom-up structure formation, as it has been demonstrated with the discovery of highly structured features, e.g., the tidal debris emerging from the Sagittarius dwarf galaxy (Ibata et al. 1994), detected by means of the movement of stars towards the Galactic center; or, more recently, the stellar debris from Gaia-Enceladus-Sausage (Helmi et al. 2018) or Sequoia (Myeong et al. 2019), both using dynamical parameters and chemical abundances of individual stars. Nevertheless, the phase-space distribution of the local halo also incorporates clumpy substructures at smaller scales (Lövdal et al. 2022; Ruiz-Lara et al. 2022).

However, the accretion relics are not expected to be found all over the Milky Way. Disc galaxies like our own include several stellar components (i.e., a bulge, a thin disc, a thick disc or a stellar halo) contained within a dark matter halo. Moreover, they contain gas and dust, though, because of our purposes, we do not

consider them further in the present work. We are mostly interested in the stellar halo, a large and diffuse spheroid of about $\sim 10^9 M_\odot$ that extends to ~ 200 kpc in radius, while the disc corresponds to $\sim 10^{11} M_\odot$ and has an extension of ~ 20 kpc in radius. According to ΛCMD cosmology, the stellar halo is the component that is almost exclusively formed by the accretion/merging processes of destroyed galaxies among the stellar components, whose origin is found in the smaller systems. That is, the stellar halo acts as a time capsule of the merging history of the Galaxy. This is due to the dynamical timescale in that region being of several Gyr (Binney & Tremaine 2008) and to the fact that the disruption of satellite galaxies by tidal forces left star trails that ended up leading to the formation of a spheroidal component (Helmi & de Zeeuw 2000). This disruption has been studied by using cosmological simulations created under the ΛCDM model of structure formation (e.g., Bullock & Johnston 2005; Grand et al. 2017), according to which the bottom-up growth of the structures (or galaxies) does lead to the stellar halo being formed mainly from accreted systems. This means that, in first approximation, stars that have escaped their progenitor conserve the integrals of motion values this last one had at the moment of the ejection. Even after several Gyr, when all the stars of a given progenitor are phase-mixed and spatially sparse, they keep their dynamical information mostly conserved.

Fig. 2: An artistic representation of the encounter between the Milky Way and the Sausage dwarf galaxy (Figure credit: V. Belokurov, Juan Carlos Muñoz (ESO) 2018).

The signatures of accretion events depend on various factors, such as when they were accreted, since they might still appear as spatially coherent features like streams of stars, which is the case for the Sagittarius stream, or with their stars having lost any spatial coherence, like for Gaia-Enceladus or Sequoia. In the latter case, it has been proposed that the stars that originated from a given satellite conserve their location in phase space. Some examples are the vertical angular momenta and energy space, (e.g., Helmi & de Zeeuw 2000; Naidu et al. 2020), the actions space (e.g., McMillan & Binney 2008; Myeong et al. 2018; Malhan et al. 2022), or the orbital velocities space (Koppelman et al. 2019). These quantities are promising tools to disclose the merger history due to them being conserved for stars that originated from the same progenitor (that is, to the initial galaxy before its accretion by the host galaxy) under the assumption that the potential of the Milky Way is preserved over several Gyr. Thus, one should be able to search for these destroyed galaxies as clumps in the phase space.

On the other hand, similar chemical abundance patterns are expected to be found on stars that once belonged to the same satellite galaxy (e.g., Freeman & Bland-Hawthorn 2002; Helmi 2020), though this is only feasible if the stars of the same progenitor present a homogeneous composition and each one of the disrupted satellite galaxies is chemically different. Nevertheless, the validity of these conditions when trying to use chemistry alone remains unclear (Casamiquela et al. 2021), so this assumption must be carefully handled.

Therefore, in order to unravel the processes that led to the Milky Way's formation history, we rely on the expectation of being able to reproduce the merging history of the Galaxy. In order to so do, we intend to use both the motions and the chemical composition of the individual stars that constitute the stellar halo in order to find substructures. Learning about the initial physical properties of the merging galaxies is essential to reveal the formation history of the Milky Way. However, it is also important to take into account that the accreted galaxies of the Milky Way galaxy are not free from external perturbations, such as interaction, kinematical heating, tidal effects, and dynamical friction (Koppelman et al. 2019). Moreover, the gravitational potential of the Milky Way itself has been varying over time. These effects may lead to multiple clumps in the integrals

of motion space that have their origin in the same structure (Fattahi et al. 2019), as well as to the overlap of different destroyed satellites in phase space. Furthermore, there are other complications: some quantities are conserved only for axi-symmetric potentials, which might not be the case for the Milky Way, and its structure might depend, for example, on the mass of the Large Magellanic Cloud.

In short, in the perfect case of a static galaxy with an axi-symmetric gravitational potential and no external perturbations, signatures of accretion are expected to be found in the integrals of motion space as clumps. In reality, what we found is that these values are well conserved; hence, we want to explore techniques in order to find these clumps and tackle the limitations of our premises by using galaxies obtained from cosmological simulations.

The non-trivial practice of trying to reconstruct the merging history of the Milky Way has been done numerous times before in several different ways: e.g. Helmi & de Zeeuw 2000, using a Friends-of-Friends algorithm-based algorithm in the integrals of motion space; Naidu et al. 2020, by manual selection, or, Lövdal et al. 2022, by means of a data-driven and statistically based clustering algorithm in the integrals of motion space. In addition, during the last few years, the use of Machine Learning (ML) algorithms to perform classification and clustering tasks in order to predict complicated patterns when dealing with large volumes of data has rapidly increased. This is closely related to the fact that the more recent surveys allow having good quality input information, which is imperative to developing a suitable ML algorithm, as well as to the increasing computing power over the last decades. It is, therefore, not surprising that there are several other works that have been published in recent times in which attempts have been made to use different Machine Learning clustering methods to do the same, like Yuan et al. 2018, applying an unsupervised ML algorithm based on a self-organizing map to the stars' kinematics, or Borsato et al. 2020, by using data mining and numerical and statistical techniques. Nonetheless, these methods usually lead to results that are very difficult to interpret.

Thus, the quantification of the level of structures in the Milky Way's stellar halo in the integrals of motion space, which allows us to make a detailed reconstruction of its formation processes, has been proven to be of great relevance by recent works, like the ones that have been mentioned, in order to find the accretion signatures. At the same time, the rapid development of Machine Learning techniques in the last few years, along with the higher-quality and more reliable data provided by the newest surveys, makes the clustering algorithms a promising tool to improve those analyses. This is the motivation that led to the fundamental goal of this project being to use Machine Learning methods (specifically, clustering methods in the Integrals of Motion space) to determine the substructures of the Galaxy that merged for the Milky Way to be as we know it today. Unsupervised Machine Learning techniques have been used; that is, meaningful patterns are found in unlabelled databases, with those patterns being associated with probability densities.

In this case, we are dealing with an unknown number of clusters and multi-dimensional and complex data that cannot be linearly separable. Different methods have been used to classify the data and find clumps in the phase space, with the Gaussian Mixture Model (particularly, the Bayesian Gaussian Mixture Model) being ultimately selected in order to parametrize them. This method associates each point with a given Gaussian, whose parameters (weights, mean values, and covariance matrices) are determined with the Expectation-Maximization method. Furthermore, different quantities that allow us to make a numerical evaluation of the performance of the Bayesian Gaussian Mixture methods under different circumstances will be considered (BIC, log-likelihood). This method has the advantage that it is easy to interpret, since each star has a probability of belonging to each of the identified clusters and, presumably, to a unique accretion event.

The main tool that has been used in this work is the `Scikit-learn` open source Python package, which includes very diverse tools that allow both supervised and unsupervised Machine Learning techniques (Pedregosa et al. 2011).

2. Objectives

Our main interest lies in unraveling the formation history of the Milky Way in order to reproduce the merging history of the Galaxy and to put constraints on the characteristics of the star formation and chemical evolution of its individual structures. To this end, we want to look for signatures left by the different accreted structures in the stellar halo. The assumptions we rely on are the following:

- Galaxies experience hierarchical structure formation as dictated by the ΛCDM cosmology.
- The stellar halo contains signatures left by the different accreted structures.
- Each disrupted satellite galaxy is constituted by stars with similar integrals of motion values, physical quantities are mostly conserved over the Hubble time (Helmi & de Zeeuw 2000). Thus, we expect to find the signatures of accretion as clusters (specifically, as overdensities) in the spaces of these values.

Specifically, we want to find and quantify the number of accretion events to make a detailed study of their internal peculiarities, like their masses and metallicity. This is achievable because we expect to find as many clumps in those spaces as the number of accretion or merging events (or, at least, to set a lower limit on said number), even after the complete spatial-mixing has taken place. The selected approach has been to design a procedure based on Machine Learning clustering techniques, for which the following steps have been taken:

1. Creating a variety of toy datasets randomly generated from Gaussian distributions with known parameters and assignation of labels to each dataset point.
2. Using these toy datasets in order to try different clustering methods and to become familiar with how they work. By doing this, the Bayesian Gaussian Mixture Model has proven to lead to the assignation of the points to the different clusters that most closely recover the original labelling.
3. Learning about the effects of the different initial parameters for Bayesian Gaussian Mixture models with the same toy datasets.
4. Using this knowledge to create models for Milky Way-like galaxy halo simulations obtained from the Auriga project, with the idea of further developing the method and, ultimately, applying it to the real Milky Way's observational data.

This work is structured as follows: Section 3 describes the clustering methods that have been considered (Subsection 3.1), along with tests of those techniques applied to toy datasets (Subsection 3.2). Section 4 introduces the Auriga simulations (Subsection 4.1), with a description of the different quantities whose spaces can be used to search for substructures (Subsections 4.1.1 and 4.1.2). Then, the Auriga data is explored in order to determine how to achieve our goal. Subsection 4.2 includes the main features of the most massive progenitors, along with the distribution of the star points that belonged to each one of the 4 most massive progenitors in the total energy versus vertical angular momentum space in different radius and metallicity bins. Subsection 4.3 presents the distribution of the 4 most massive progenitors and 4 intermediate-mass progenitors in the different spaces where clustering is expected, along with a metallicity histogram. Next, Section 5 incorporates the Bayesian Gaussian Mixture Model results for the two subdatasets that have been considered (see Subsection 5.1) and the process that has been followed in order to link different Gaussians together (see Subsection 5.2). Lastly, Section 6 summarizes the results and presents future improvements to be made. In addition, four appendixes are also included: a detailed explanation of the Expectation-Maximization algorithm, used in the Bayesian Gaussian Mixture model (Appendix A), two comparisons of the classical and Bayesian Gaussian Mixture method results for controlled datasets (Appendix B), the distribution in the total energy against vertical angular momentum space for 4 intermediate-mass progenitors (Appendix C) and the Bayesian Gaussian Mixture Model results for the 4 most massive progenitors in a certain radius range for 5 different combinations of quantities (Appendix D).

3. Methodology

As it has been already mentioned, to achieve our goal of using dynamical and chemical information of halo stars to find the residual signatures of the pristine galaxies that have been accreted or have merged with the Milky Way along its history, we look for signatures of accretion in the form of clusters. Unsupervised Machine Learnings methods, such as the Gaussian Mixture model, are perfect to achieve that objective, as they have already proved their efficacy in identifying clusters in a non-labelled dataset (see Section 3.2).

We will introduce the mathematical notation used to describe a set of points as a set of multi-dimensional Gaussians and discuss the methodology used to find the number and parameters of the Gaussians that best fit the sample under consideration.

3.1. Clustering algorithms

As it was already stated, we are interested in using clustering algorithms to group the data in a given feature space, that is to say, to make a classification into groups according to their similarity. Some of the clustering methods that have been considered in this master project are the following:

- **K-Means (Lloyd 1982):** The samples are separated into groups of equal variance.
- **DBSCAN (Ester et al. 1996):** Density-based spatial clustering of applications with noise that considers the different clumps as regions of high density separated by areas of lower density.
- **Gaussian Mixture Model (Duda & Hart 1973):** The data classification is based on the likelihood that a point belongs to a given multi-dimension Gaussian probability distribution.

After a research process in which all three methods have been tested with toy datasets constituted by clearly discernible Gaussian overdensities (see Subsection 3.2), the Gaussian Mixture Methods have proven to lead to the best outputs in the scenario we are interested in, since we expect the clumps in the Milky Way stars' phase space to have Gaussian-like distributions. Therefore, specifically, all throughout this project, we have been mainly using two clustering methods: the classical Gaussian Mixture Model, which we will refer to simply as the Gaussian Mixture Model, and a variant of this method called the Bayesian Gaussian Mixture Model.

3.1.1. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a probabilistic clustering method in which each component is a multivariate Gaussian density defined by its mean and covariance matrix, as well as by its weight. A given normalised multi-dimensional Gaussian probability distribution function (Normal distribution) is represented in an arbitrary number of dimensions n as follows:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det\,(\boldsymbol{\Sigma})}} exp\left[-\frac{1}{2}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right], \tag{1}$$

with $\mathbf{X} = \mathbf{x} - \boldsymbol{\mu}$, where $\mathbf{x}$ is a vector with n dimensions, $\boldsymbol{\mu}$ is the Gaussian's mean value and $\boldsymbol{\Sigma}$ its the covariance matrix, of $n \times n$ dimensions.

The covariance matrix is a square matrix whose diagonal elements correspond to the variance (Var) in each of the n dimensions and the other elements to the covariances (Cov) of $n \times n$ dimensional data as it follows:

$$\mathbf{\Sigma} = \begin{bmatrix} Var\left(X_1\right) & \dots & Cov\left(X_1, X_n\right) \\ \vdots & \ddots & \vdots \\ Cov\left(X_n, X_1\right) & \dots & Var\left(X_n\right) \end{bmatrix} \tag{2}$$

with X_k being each of the n components of $\mathbf{X}$.

When using this method, we will be representing our N datapoints as a linear combination of G multivariate normal distributions. Given that multiple representations are possible, the following question arises: how do we know which model is the best model among those where convergence is reached?

Below, we describe two of the essential quantities that are used to evaluate the output of the GMM:

- **Likelihood and Gaussian density distribution:**

  The likelihood of a given point $\mathbf{x}_i$ (marginalized likelihood) of our N points dataset is defined as follows in the case of a GMM:

$$p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) = \sum_{j=1}^{G} \alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right), \tag{3}$$

  where $\mathcal{N}$ being Gaussian probability distribution function defined as seen in Eq. 1, whose parameters $\boldsymbol{\theta}$ include $\boldsymbol{\mu}_j$, the vector of mean values, $\mathbf{\Sigma}_j$, the covariance matrix and the normalization factor for each of the individual G Gaussians α_j so that $\sum_{j=1}^{G} \alpha_j = 1$.

  On the other hand, the likelihood of the entire dataset would correspond to the product of the different likelihoods:

$$L = \prod_{i=1}^{N} p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) \tag{4}$$

  However, the natural logarithm of the likelihood is used, mostly because of numerical limitations. This can be done because it is a monotonically increasing function. The log-likelihood for the entire sample that is being considered is given by:

$$lnL = \sum_{i=1}^{N} ln\left[\sum_{j=1}^{M} \alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)\right], \tag{5}$$

  and it has to be maximized to determine the best parameters of the Gaussian Mixture that reproduce the observed dataset.

  Therefore, in this particular case, we make the assumption that every point $\mathbf{x}_i$ has been generated from an individual Gaussian j, which implies that all data can be sorted in G subsamples, depending on the Gaussian

that originated them. We are capable of determining the probability that each point belongs to a certain cluster or class probability, which can be defined as:

$$p\left(j|\mathbf{x}_i\right) = \frac{\alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}{\sum_{j=1}^{G} \alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \rightarrow \sum_{j=1}^{G} p\left(j|\mathbf{x}_i\right) = 1, \tag{6}$$

that is, the sum of the probabilities of each point belonging to a given Gaussian is the unit.

Afterwards, the Expectation-Maximization (EM) algorithm is the method that is used to attempt to find the parameters that maximize the likelihood function for a given dataset (see Appendix A).

- **Bayesian Information Criterion (BIC) or Schwarz criterion:**

  One way to qualify if adding more freedom to a model is meaningful is with the BIC. This quantity is often used to select the optimal number of clusters that allow us to simulate the data set to which the GMM is applied. It is based on the maximization of the data likelihood.

  The BIC is given by the following equation:

$$BIC = klnN - 2lnL, \tag{7}$$

  with N the total number of data points and lnL the maximum value of the log-likelihood. k denotes the number of free parameters, which are α_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ for each one of the Gaussians. The model with the smallest BIC would generally correspond to the preferable model for two reasons: it is related with lower penalties of free parameters (the larger k, the larger and, accordingly, worse, is the BIC) and, the larger the lnL, the smaller (and, thus, better) is the BIC value (Ivezic et al. 2020).[1]

However, though, according to their own description, the model with the largest log-likelihood and/or smallest BIC is expected to provide the best description of the data, this may not be the case if the number of free parameters is different between models. A larger number of free parameters, that is proportional to the number of Gaussians, will mostly lead to a larger likelihood, in the same way that the position of 3 points could be fitted with a large degree polynomial, but this would not make it the optimal model. Nevertheless, this is not always the case, mostly for numerical reasons. Therefore, finding the proper number of Gaussians remains a challenging task for the Gaussian Mixture Method.

3.1.2. Bayesian Gaussian Mixture Model

The Bayesian Gaussian Mixture Model (BGMM) is a variant of the Gaussian Mixture Model in which the number of effective components can be inferred from the dataset, as long as an upper limit for this value is selected. That is, we select the maximum number of Gaussians that can be retrieved, G_{max}, which can be different from the number of Gaussians the model predicts, G. Nonetheless, it is worth mentioning that the upper value is only needed because of the limited possibilities of computers, since, theoretically, it could be infinite.

The main difference between classical and Bayesian techniques is the fact that the latter adds extra information (also called hyper-parameters) to the analysis, which is usually referred to as *prior*, extending the

[1] There are other quantities that are used to compare models, such as the Akaike information criterion (AIC), closely related to the BIC. The reason to choose the BIC over the AIC is that its value is larger (less adequate) when more model parameters are considered, since $BIC \propto k$, leading to the selection of simpler models.

likelihood concept. It describes the expected distribution of the model free parameters. Therefore, the foundation of this algorithm is the same as in the expectation-maximization algorithm, though, in this case, since we are dealing with a variational method, regularization is added by means of the integration of information obtained from prior distributions.

In this case, the Bayes' rule is applied to the likelihood function as follows:

$$p\left(m, \boldsymbol{\theta}|D, I\right) = \frac{p\left(D|m, \boldsymbol{\theta}, I\right) p\left(M, \boldsymbol{\theta}|I\right)}{p\left(D|I\right)}, \tag{8}$$

where m represents the model, whose parameters are not directly estimated by the algorithm, but whose distribution of possible values is estimated. It would incorporate k distributions, one for each model parameter, represented the $\boldsymbol{\theta}$ vector. On the other hand, I stands for the prior information and D for the data. The main idea is that Bayes' rule allows us to obtain an improved model by means of the combination of an initial belief and the incorporation of the data.

The different members of the rule are interpreted as follows:

- $p\left(m, \boldsymbol{\theta}|D, I\right)$: *posterior* probability distribution function corresponding to the model m and $\boldsymbol{\theta}$ parameters, given the data D and the I prior information.
- $p\left(D|m, \boldsymbol{\theta}, I\right)$: the likelihood given certain m and θ along with I.
- $p\left(m, \boldsymbol{\theta}|I\right) = p\left(\boldsymbol{\theta}|m, I\right) p\left(M|I\right)$: *prior*, which represents the shared probability associated with m and the parameters $\boldsymbol{\theta}$.
- $p\left(D|I\right)$: *probability of the data* or prior predictive probability of the data, which is used to normalize the posterior probability distribution function.

The best model parameters would correspond to the maximum posterior probability density function $p\left(m|D, I\right)$, obtaining what is called the *maximum a posteriori*. The main idea is that the knowledge is continuously refined, starting with no data (*prior*) and being updated by using the data in order to get to the *posterior* (Ivezic et al. 2020).

In this case, an infinite mixture model with the Dirichlet Process as a prior for the weights distribution has been used (Pedregosa et al. 2011). On the other hand, the weights, means and precisions of the different components must be initialized under a given criteria so, to that end, the K-means method has been used.

- **Dirichlet process:**

  The Dirichlet distribution takes place over a k-dimensional vector of real numbers between 0 and 1, so that the total sum is one. The Dirichlet process prior is used to assign mixing probabilities to an infinite number of components so that each probability must be between 0 and 1, and the total sum should give the unit: this is the so-called stick-breaking process, which contains a Beta distribution prior Γ, defined as follows:

$$p\left(\mathbf{x}, \alpha, \beta\right) = \frac{\Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \mathbf{x}^{\alpha-1} \left(1 - \mathbf{x}\right)^{\beta-1}, \tag{9}$$

  with $0 < \mathbf{x} < 1$ and this function being described by two parameters $\alpha > 0$ and $\beta > 0$.

- **K-means:**

  The **K-means** method aims to find the values that minimize the sum-of-squares objective function, often called inertia or within-cluster sum-of-squares-function, that is defined as:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2, \tag{10}$$

  with C_k being each one of the K subsets of equal variance in which the data is divided, N_k the number of points of each partition, $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i \in C_k} \mathbf{x}_i$ the mean of the points of each set and $C(\mathbf{x}_i) = C_k$ referring to the fact that the $\mathbf{x}_i$ class is C_k (Ivezic et al. 2020).

  In order to minimize the inertia, each one of the centroids $\boldsymbol{\mu}_k$ is initialized and then, the cluster at which each point is closest to is selected so that $C(\mathbf{x}_i) = arg\,min_k ||\mathbf{x}_i - \boldsymbol{\mu}_k||$. After that, new centroid values are found and each point is assigned to a cluster until the difference between the $\boldsymbol{\mu}_{k-1}$ and $\boldsymbol{\mu}_k$ values is smaller than a certain threshold. Even though there is a possibility of not finding a global optimal minimum for Equation (10), but a local minimum (the minimum value for each one of the EM iterations), the difference between the global and the local minimum does not increase with the K value (Pedregosa et al. 2011).

Therefore, in summary for the BGMM, the method begins with the *prior* such as all mixing probabilities are equally likely (Dirichlet process prior) and the Gaussian parameters are initialized by using the K-means method. Afterwards, the likelihood is obtained and the model is updated by means of Bayes' rule in order to obtain the *posterior*. The Expectation-Maximization algorithm is then used (Chlon 2020) in the same way as for the GMM, though in this case we aim to find hyper-parameters that describe the parameter distributions that maximize the posterior.

3.1.3. Input parameters of the Gaussian and Bayesian Gaussian Mixture Models

Taking all of this into account, the Gaussian Mixture models must be built by selecting the following values:

- Number of mixture components, that is, the maximum number of Gaussians explored, denoted as G_{max}. It must be noted that G, the number of Gaussians that is found to best reproduce the data, is such that $G = G_{max}$ for the GMM and $G \leq G_{max}$ for the BGMM.
- Tolerance (tol), that is, the convergence threshold so that, when the lower limit of the average gain of the likelihood of the training data with regard to the model is smaller than this value, the EM algorithm stops iterating.
- Regularization (reg), the value that is added to the diagonal of the covariance matrix to avoid getting negative values for the individual elements of the diagonal. For the same reason, this value also imposes a lower threshold on the size of the Gaussians that can be found by the model.
- Maximum number of EM iterations (max_iter).
- Number of K-means initializations (n_init). The result with the largest minimum likelihood value is kept, so we avoid being stuck in a local minimum.

On the other hand, in the case of the Bayesian Gaussian Mixture, the Dirichlet concentration of each component in the weight distribution must be selected. This value is such that, the higher it is, the more mass is put in the centre, which leads to favouring a solution with more components. On the contrary, the lower it is,

the more mass is put at the edge of the mixture components (Pedregosa et al. 2011).

The number of clumps G that we expect to find in the datasets is unknown since multi-dimensional spaces are used. Therefore, manually selecting G, as the GMM requires, forces us to do many models in order to explore how many Gaussians we need and implies adding a constraint on what we are able to reproduce; which is not (or less) the case with the BGMM. With the BGMM, it is enough to select the upper limit on the number of clusters G_{max}, as long as this value is not large enough for it to not be feasible, and this leads to the prediction of a smaller G value. In that case, the model automatically recovers the best model possible with G Gaussians so that $G_{max} \geq G$. Additionally, the GMM has the disadvantage that it can produce extremely different solutions since it has a larger sensitivity to the number of the model parameters and simply forces the results to G Gaussians however it finds feasible.

3.2. Validation of the Gaussian and Bayesian Gaussian Mixture Model with toy datasets

Here is where one of the most crucial questions of this project arises: how do we know if we have retrieved the actual number of past accretion events that built our Galaxy? Or, according to how the GMM works, if we didn't know the actual number of Gaussians that generated the dataset, how would we select the number of clusters for the model?

In order to develop the method that would allow us to sort the points into different clusters and show their performance, toy datasets have been employed. These datasets are composed of points that belong to a pre-determined number of multi-dimensional Gaussians that are randomly generated and whose weights, mean values, and covariance matrices are known. As a consequence, we are able to compare the original Gaussian parameters with the Gaussians predicted by the models. Python scripts have been used both to generate the toy datasets and to apply the GMM and BGMM with different input parameters.

The first step that was taken was applying the GMM to these datasets with different input parameters, which led to less satisfactory and very different results. This instability indicates that the model does not seem suitable for this case, since it seems to find only a local minimum for the BIC and a local maximum for the log-likelihood, while we are interested in their global values. Moreover, as it has already been mentioned, if the actual number of mixture components does not match the number of Gaussian distributions G, the model is forced to be adapted to a determined number of clusters as it finds suitable, leading to either over-fitting or under-fitting. That is, we have to know the number of clusters beforehand.

Since we want to create a model that is able to automatically determine the number of Gaussians that gave rise to the toy dataset, the BGMM is the most suitable choice. As it was already stated, in this case, an upper limit on the number of Gaussians G_{max} is imposed, and the number of Gaussians that fit the data best G is determined. Two comparisons between the GMM and BGMM results that explicitly prove why the BGMM is the adequate option in this case can be found in Appendix B. Therefore, from now on, we will focus entirely on the BGMM method.

The BIC and total log-likelihood of the model are used to evaluate the quality of the outcome and, most importantly, to compare models obtained with different input parameters. Therefore, the first thing to do is try to fit the data with different BGMM to learn the effect the different initial conditions (G, tol, reg, max_iter and n_init), as well as find a criteria to choose the most fitting.

First of all, the max_iter, the upper limit on the number of EM iterations, which stops once convergence is reached, must be large enough for this condition to be always satisfied. This value can be selected by means

of trial and error. On the other hand, the weight concentration prior has been fixed as $1/G_{max}$ so that we do not favour the mass either on the centre or in the edge of the components.

Regarding the n_init parameter, though a single K-means initialization may allow us to find convergence (if the max_iter value is large enough), this can lead to the proper number of Gaussians never being retrieved. Nonetheless, a high number of n_init is needed to efficiently explore the parameters space, since the solutions are likely non-monotonic and a local maximum of the likelihood and not the global maximum. However, this is computationally expensive and it is not possible to determine a suitable value by using the BIC or the log-likelihood to compare the BGMM outputs when considering different n_init. Therefore, a value that is as large as possible and yet workable for the computer is selected. In this case, it has been fixed to $n\_init = 100$.

Lastly, several combinations of *tol* and *reg* values have been considered in order to determine the impact of varying these parameters when applying the model to the data. Four different combinations, corresponding to four orders of magnitude for each one of them, are used as input parameters to enhance the similarities and differences of the considered BGMM. The *tol* has been selected to be an order of magnitude larger than the *reg* value to avoid obtaining a *fake* convergence. In this case, the BIC, log-likelihood and number of predicted Gaussian values are used to compare them so that, according to their own definitions, it can be assumed that the best possible model would correspond to the BGMM with the largest log-likelihood value or, reciprocally, the smallest BIC value. However, in this case, since our main interest lies in determining the most acceptable number of predicted Gaussians, k, the number of free parameters (see Equation (7)), has been selected so that the $k = \left(1 + n + n^2\right) \times G$, with n being the number of dimensions.

The comparison of the models that have been obtained for these four combinations of *tol* and *reg* values with different G_{max} (8 to 15) for a toy dataset generated by 10 different Gaussians is presented in Fig. 3 (see the upper left panel of Fig. 4). However, this same method has been also applied to other toy datasets generated by smaller and larger numbers of Gaussians in order to determine the consistency of the results (see BGMM results in Appendix B).

Fig. 3 shows that the total log-likelihood and the BIC parameters have opposite behaviors: the greater the log-likelihood value, the smaller the BIC value. This suggests that the log-likelihood and this BIC have a similar interpretation, despite the first one being computed with the k value associated with G_{max} and the second one with G only. It also reflects that, as G_{max} increases, the BIC and total log-likelihood curves' behavior approaches a plateau, which implies that the models converge to a certain value of these quantities so that the difference between them is negligible.

Both of these conditions are satisfied for all combinations of the *tol* and *reg* values for 10 components, for which we predict that the number of Gaussians that originated the dataset is also 10. Furthermore, the BIC and log-likelihood values are really similar in almost all cases. The exception is the $tol = 1$, $reg = 1 \times 10^{-1}$ one, where the BIC values are always considerably larger, the total log-likelihood values are always noticeably smaller than the other parameter combinations, and the predicted number of Gaussians differs when trying to fit the model with 12 and 14 components. This implies that these input parameters do not allow us to find the maximum log-likelihood, likely because of the *tol* value being excessively large. It imposes a threshold in the lower bound average gain value of the log-likelihood that makes the EM iterations stop too soon. On the other hand, the *reg* value is also way too large, so a very large quantity is added to the diagonals of the covariance matrices. However, in all other three cases, the BIC and total log-likelihood values are very close and lead to the same G, our main concern, so it would not be prudent to state if one model is better than the others. Nevertheless, it must be considered that, the smaller the *tol* and *reg* values, the larger the computational time, so the less computationally expensive model should be preferred.

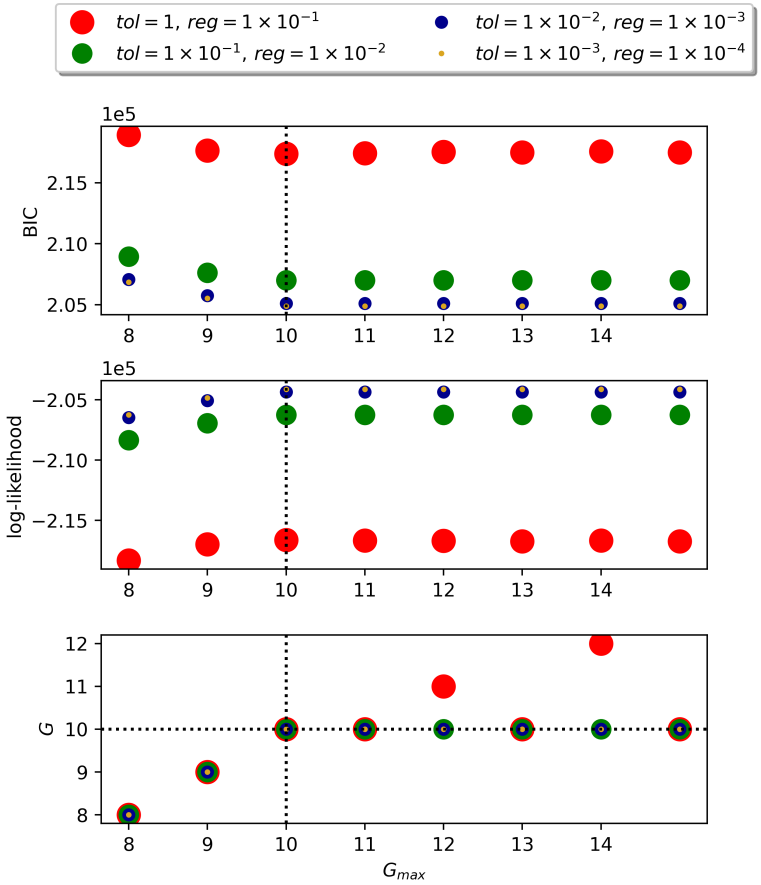

Fig. 3: BIC (*top*), total log-likelihood (*middle*) and predicted number of Gaussians G (*bottom*) predicted for each of the number of components that have been considered for each G_{max} of the different BGMM. The different colors of the points refer to the different *tol* and *reg* values, as indicated in the legend, and the vertical lines correspond to the number of components for which the maximum total log-likelihood and the minimum BIC are reached. The actual number of Gaussians used to generate this dataset is 10 and is indicated with a horizontal line in the bottom figure.

Consequently, it can be concluded that the parameters that lead us to the best BGMM for this toy dataset could be of the three *tol* and *reg* combinations, as long as the *n_init* and *max_iter* values are large enough to obtain a reasonable output. Nonetheless, it must be taken into account that the *reg* value must be carefully selected since the addition of a value that is too large in the diagonal of the covariance matrix may lead to either an overestimation or an underestimation of the number of predicted Gaussians.

The results that were obtained when using this BGMM can be seen in Fig. 4. The left column panels reflect the diversity of the Gaussian shapes in the toy model, so we can determine if our method has limitations that have to do with this aspect. It is clear from the colored-plot that the distribution of points of each cluster is expected to be reproduced by a Gaussian, though the density plot may lead us through visual inspection to the conclusion that this dataset contains only 8 well separated overdensities with a Gaussian-like shape. Furthermore, it manifests the great difficulty that would imply having to determine the existence of the Gaussian number 10 because of its small density, as well as of Gaussian number 7, which is hard to distinguish from number 4. This demonstrates the need to use a ML clustering method where the number of effective

components is automatically detected by the algorithm. There is also a great similarity between the original dataset density plot in the upper right panel and the generated sample density figure in the upper right panel, so one can conclude that the BGMM model is suitable for the generated toy dataset. Lastly, the bottom right panel of Fig. 4 shows that the shape and location of all 10 Gaussians have been successfully reproduced by our BGMM independently of their shape. This is particularly relevant for the bottom left Gaussian (number 6), which may seem to have a larger resemblance to a straight line than to an actual Gaussian. In addition, the model was able to find Gaussians 10 and 7 as independent clusters. Nevertheless, it must be noted that, though the extremely thin Gaussian 6 is identified as an independent cluster, its corresponding clump in the generated sample is considerably thicker.

Therefore, it has been demonstrated that, if the clusters in the data distribution have a Gaussian shape,

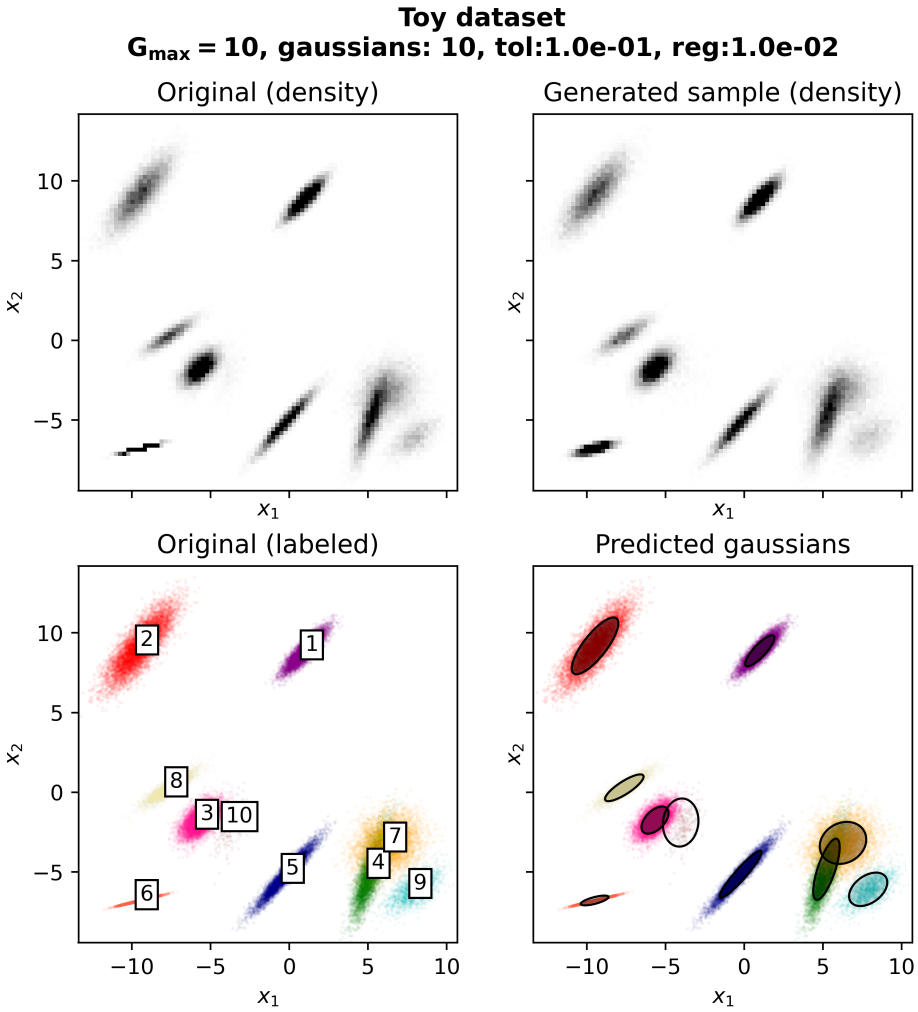

Fig. 4: **Upper left:** Density plot of the original toy dataset. **Upper right:** Random sample generated from the fitted Gaussian distribution. The more the sample resembles the original distribution, the better the model. **Bottom left:** Original dataset with the different Gaussians color-coded. **Bottom right:** Location and shape of the Gaussians predicted by the BGMM as ellipsoids in such a way that the opacity of the interior region of the ellipsoids is directly proportional to their weight.

it is possible to apply the Bayesian Gaussian Mixture Model to them to find the number of clumps they are composed of.

In short, regarding the tol, reg, n_init and max_iter values, a large number of tests have led us to the conclusion that the results are very similar as long as the reg value, which describes the minimum size allowed for a Gaussian, is small enough to detect the smallest substructure. On the other hand, we must ensure that the tol, n_init and max_iter are such that the convergence is reached. In addition, enough n_init are needed for the results to be consistent with the overdensities we are able to detect by looking at the star distribution in the considered spaces. Regarding G_{max}, it is selected so that we can be confident that it is larger than some visual estimation of the G we expect to find, but not so excessively large that there is an obvious overestimation of the number of clusters. Therefore, in the case of Fig. 4 we should select $G_{max} > 8$, since this is the number of clusters we undoubtedly identify, but, for example, $G_{max} < 20$, since it is evident that we are not going to recover that many independent Gaussians.

4. Data: Stellar haloes in simulated Milky Way-like galaxies

In order to evaluate the ability of the BGMM on more realistic datasets, we use high-resolution simulations of Milky Way-like haloes, specifically, simulated galaxies from the Auriga suite (Grand et al. 2017). These are magneto-hydrodynamic cosmological simulations in the ΛCDM paradigm that include several physical mechanisms of great relevance in galaxy formation processes, such as gravity, feedback effects, star formation, and gas cooling.

4.1. Description of Auriga

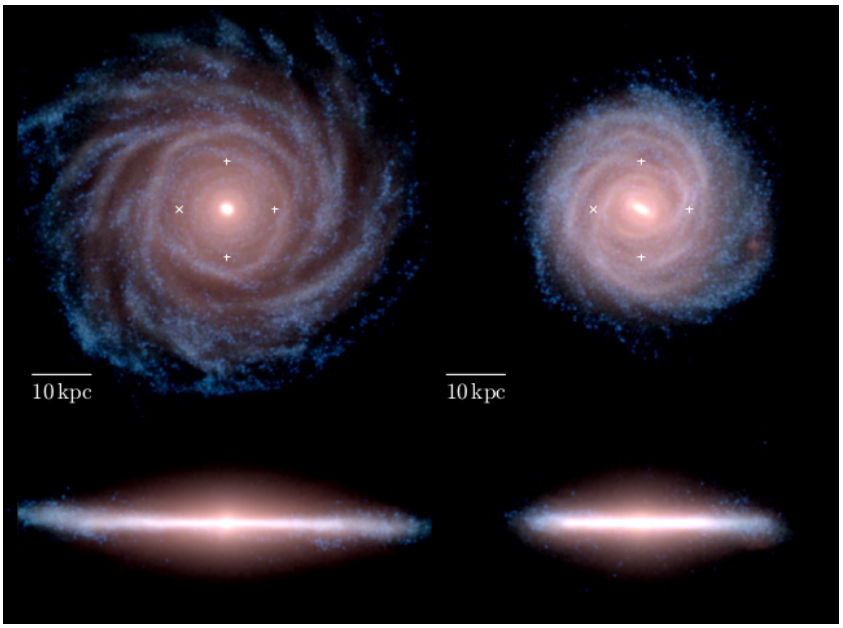

Fig. 5: Face-on and edge-on projection of stellar densities from two simulated galaxies in Auriga (Au 24 and Au 27) at $z = 0$. Younger stars are shown in bluer colors and older stars appear in redder tones (Figure credit: Grand et al. 2018).

The Auriga simulations provide us with information about the progenitor from which stellar particles at $z = 0$ came from, so they can be used to test the ability of the developed method, finding its weaknesses and determining the quality of the base premises. The use of the BGMM clustering algorithm has the drawback of assuming that all data points belong to a certain cluster, and, while this is true in the case of the accreted particles of the Auriga simulations, we expect a large amount of noise in the observational data, that is, the presence of unstructured points. Thus, it must be taken into account that a method that can handle noise might be needed when dealing with real data, despite this not being an issue with the Auriga galaxies. However, it has been recently suggested that $\gtrsim 95\,\%$ of the Galaxy's halo came from progenitor galaxies (Naidu et al. 2020), so assigning every data point to a given Gaussian seems like a sensible assumption. On the other hand, it is worth mentioning that Naidu et al. 2020 over interpreted their results, since their prior was that all stars belong to a given progenitor. As a consequence, the use of the method that will be developed for observational data must be carefully handled.

These catalogues incorporate gas cells, dark matter particles, wind particles and black holes, as well as star particles that correspond to in-situ stars (stars bound to the main halo when they were formed), stars still bound to other subhaloes and accreted stars part of the main halo at $z = 0$. Moreover, they include not only the kinematical and chemical information of the star particles that constitute them, but also the main progenitor branch each of them was a part of before merging when the object reached its peak mass. This implies that we know the origin of each star particle, and it is possible to compare the results to which the model leads and the actual distribution of the star particles that belonged to the different progenitors of a simulated galaxy whose properties are similar to those of the Milky Way (Grand et al. 2017, 2018).

In this study, accreted star particles of the main halo at $z = 0$ are used and treated as individual stars. We focus on the stellar halo for the reasons stated in Section 1; in short, it is assumed that the stellar halo consists

of mostly accreted stars that once belonged to satellite galaxies that were disrupted.

This work focused on two haloes in the highest resolution simulations of the Auriga suite, that have been selected because of the presence of a Gaia-Enceladus analogue in them, which we would expect to be more useful to develop a method that will be applied to actual Milky Way data. These are Halo 24 (Au 24, $R_d = 6{,}1$ kpc), that presents structures in $E_{tot}\, vs\, L_z$ for the stellar halo star particles with a qualitative resemblance to those seen in the Milky Way stellar halo; and Halo 27 (Au 27, $R_d = 3{,}2$ kpc), due to the fact that its scale radius is closest to the Milky Way's and because of its interesting satellite interactions. The results that are shown in the present work correspond to Au 27, since both simulations led us to different results but similar conclusions (Grand et al. 2018).

These simulations were kindly provided by Robert Grand along with a script to read the accretion history of each star particle within R_{200} of the main halo at $z = 0$. Afterwards, a cross-match between the information included in these files and the raw snapshot data has been performed by means of a Python script. This code allows us to get all the parameters information we need, along with an identification number that lets us know which star particles belonged to the same disrupted galaxy. Then, the accreted star particles of the Au 27 halo were selected and the parameters needed to calculate the quantities that will be used by BGMM were obtained. All this information, the dynamical and chemical quantities of each star particle and the labels that dictate their origin, allows us to compare the labels predicted by the model, that is, the assignation of each point to a given mixture component (a given Gaussian) with the actual progenitor they belong to.

4.1.1. Dynamical parameters

First of all, in order to develop our method, we have to select the dataset that will be used to search for the clusters, i.e., we need to select the kinematically-related spaces where we expect to find clusters of star particles that once belonged to the same progenitor.

The parameters we use to define our primary workspace are the total energy, E_{tot}, the amplitude of the angular momentum vector along the z axis (symmetry axis), L_z, and the perpendicular angular momentum, $L_\perp$. However, while E_{tot} and L_z are fully conserved when an axi-symmetric potential is under consideration, as we can assume in first approximation, $L_\perp$ is only roughly conserved, as it typically varies slowly. Nevertheless, Helmi & de Zeeuw 2000 showed that $L_\perp$ is quasi-constant and, therefore, can be used to search for substructures. It has to be noted that, despite the Auriga simulations not having an axi-symmetric potential, it is close enough that the assumption of E_{tot} and L_z being integrals of motion remains valid. This would constitute a limitation of our model, though it is used for simplicity and it is a standard practice in the astronomical community. Moreover, the reason why we use $L_\perp$ despite this value not being well conserved is to reduce the chances of overlap. This selection has been motivated by the large number of previous works that have demonstrated the legitimacy of the $E_{tot} - L_z - L_\perp$ space to retain information about the accretion events due to the high degree of clustering it contains (e.g, Helmi et al. 1999; Helmi & de Zeeuw 2000; Naidu et al. 2020; Lövdal et al. 2022).

These three quantities of our primary workspace are defined as follows:

$$E_{tot} = \frac{1}{2}v^2 + \Phi \tag{11}$$

$$L_z = x \cdot v_y - y \cdot v_x \tag{12}$$

$$L_\perp = \sqrt{L_x^2 + L_y^2}, \tag{13}$$

where $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$ denotes the total velocity of the particle, Φ is the gravitational potential and $L_x = y \cdot v_z - z \cdot v_y$ and $L_y = z \cdot v_x - x \cdot v_z$, with both the coordinates and the velocities being centred at the Galaxy centre. The gravitational potential provided by the Auriga simulations is directly used in the computation of the total energy instead of a fitted axi-symmetric potential values, since we are interested in testing the cluster selection method. However, a continuation of this project may include applying the model with a fitted axi-symmetric potential.

On the other hand, it is also possible to search for substructures in the Milky Way's halo in the action space (see Myeong et al. 2018; Malhan et al. 2022). These quantities are also integrals of motion, so their invariance is conserved under the effect of slow changes (e.g, Helmi et al. 1999; McMillan & Binney 2008; Myeong et al. 2018). Therefore, the action spaces $\frac{J_\phi}{J_{tot}}$ and $\frac{(J_z - J_r)}{J_{tot}}$, with J_z describing the vertical oscillation of a given orbit, J_r the radial oscillation and J_ϕ the azimuthal oscillation are considered. Unfortunately, action can only be analytically computed for a few specific potentials like the Stäckel potential, in which case the motion will be fully integrable. For practical reasons, they have been computed by using the AGAMA Python package (Vasiliev 2018a), which requires an axi-symmetric potential, so we fitted a realistic axi-symmetric model composed of 2 discs, a bulge and a NFW halo to the potential of the Auriga simulation, which only differs slightly from this type of model. J_{tot} is defined as $J_{tot} = J_r + J_z + |J_\phi|$ and the action space is defined by $(J_z - J_r)/J_{tot}$, that is, the normalized difference between the vertical and radial actions, vs J_ϕ/J_{tot}, the normalized azimuthal action, so the projected action space is studied (following Binney & Tremaine 2008; Myeong et al. 2019; Vasiliev 2018b).

Since the properties of dynamical substructures depend on the merger history of Au 27, the velocities can also be used in order to study it (Koppelman et al. 2019; Lövdal et al. 2022). These velocities are defined in a right-handed spherical coordinate system with the origin at the Galactic centre as follows:

$$v_r = \frac{xv_x + yv_y + zv_z}{\sqrt{x^2 + y^2 + z^2}} \tag{14}$$

$$v_\phi = \frac{xv_y - yv_x}{\sqrt{x^2 + y^2}} \tag{15}$$

$$v_\theta = \frac{\left[v_z\left(x^2 + y^2\right) - z\left(xv_x + yv_y\right)\right]\sqrt{x^2 + y^2}}{x^2 + y^2 + z^2}, \tag{16}$$

where v_r denotes the radial velocity, v_ϕ the azimuthal velocity and v_θ the zenithal velocity. We would expect some degree of clustering in the $v_\phi \, vs \, v_r$ space.

4.1.2. Chemical parameters

In addition, individual stars also retain information about their origin in their chemical abundances and do not change when scattering in the phase space takes place (Freeman & Bland-Hawthorn 2002; Lee et al. 2015). Therefore, the metallicity of the elements that are heavier than He, $[M/H]$ is defined as follows:

$$[M/H] = log_{10}\left(\frac{Z_{Auriga}}{Z_\odot}\right) \tag{17}$$

Z_{Auriga} refers to the mass fraction of all elements that are not hydrogen or helium, that is obtained from the Auriga simulations, and $Z_\odot$ is the corresponding standard solar value.

4.2. Main characteristics of the most massive progenitors

Before applying the BGMM to these more realistic datasets, a preliminary exploratory process of this data must be conducted in order to evaluate how suitable we expect the method to be, as well as to determine the model input parameters that might fit our interests best.

In general, the more massive progenitors have experienced more nuclear reactions, that is, the mean metallicity of the stellar component is larger with respect to smaller progenitors, and the metallicity distribution function of the stars extends over a larger range of $[M/H]$ values. On the contrary, less massive progenitors tend to be more metal poor and their whole metallicity distribution covers a smaller range of $[M/H]$ values. At the same time, the more massive they are, the more likely it is for the integrals of motion to not be conserved due to the dynamical friction effect, which produces a loss of the angular momentum and kinetic energy of stellar systems because of their gravitational interactions with their surrounding matter (Binney & Tremaine 2008). The reason for this is that enough time has passed for the dynamical friction to act, which leads to a single satellite galaxy giving place to several clumps (Jean-Baptiste et al. 2017; Grand et al. 2019; Koppelman et al. 2019).

The main features of the 40 most massive progenitors are summarised in Table 1 and Fig. 6. As it can be seen in Table 1, for the Au 27 galaxy, the contribution of the 4 most massive progenitors constitutes $\sim 70\,\%$ of the entire stellar halo. Moreover, the percentages associated with the different progenitors rapidly become smaller as we go towards smaller satellites. At the same time, even though there is not a strict tendency, the mean values of total metallicity, $[M/H]$, are larger for the most massive progenitors.

However, all of this is true as long as the whole distance and metallicity ranges are being considered. Therefore, before jumping into conclusions, the next thing to do is to study the peculiarities in clustering effects that are found in different distance and metallicity bins.

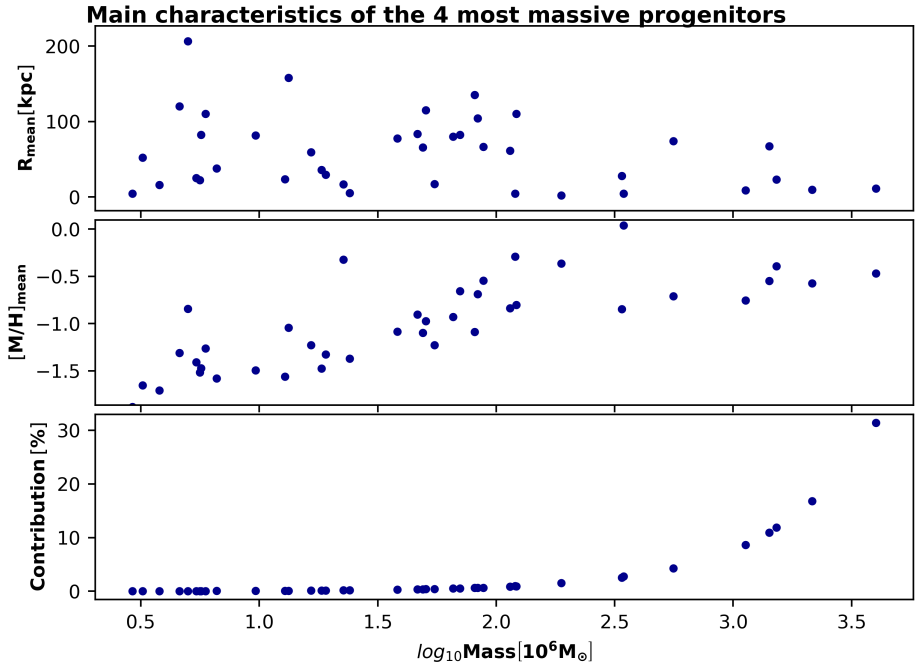

Fig. 6: Mean galactocentric radius values, mean $[M/H]$ values and contribution of the disrupted satellite galaxies to the accreted star particles of the main halo at $z = 0$ against their stellar mass in a logarithmic scale.

Order	**Mass** $[10^6 M_\odot]$	$\mathbf{R_{mean}}$ **[kpc]**	$\mathbf{[M/H]_{mean}}$	%
1	4011.57	11.2	-0.47	31.38
2	2154.77	9.7	-0.57	16.79
3	1522.79	23.2	-0.39	11.87
4	1420.95	67.3	-0.55	10.95
5	1129.91	8.8	-0.75	8.66
6	559.47	74.2	-0.71	4.28
7	345.23	4.5	0.04	2.78
8	338.72	28.0	-0.85	2.56
9	188.33	2.0	-0.37	1.53
10	120.31	4.7	-0.29	0.94
11	121.66	110.3	-0.80	0.92
12	114.24	61.3	-0.84	0.87
13	88.21	66.7	-0.55	0.65
14	83.43	104.3	-0.69	0.63
15	81.04	135.3	-1.09	0.61
16	70.35	82.6	-0.66	0.53
17	65.76	79.9	-0.93	0.50
18	54.95	17.2	-1.23	0.42
19	50.49	115.1	-0.97	0.38
20	49.06	65.6	-1.10	0.37
21	46.46	83.8	-0.90	0.35
22	38.24	77.6	-1.09	0.29
23	24.10	5.3	-1.37	0.18
24	22.63	16.7	-0.32	0.18
25	19.06	29.5	-1.33	0.14
26	18.32	36.1	-1.48	0.14
27	16.53	59.5	-1.23	0.13
28	12.83	23.8	-1.56	0.10
29	13.30	158.1	-1.04	0.10
30	9.64	81.8	-1.49	0.07
31	6.61	38.1	-1.58	0.05
32	5.93	110.3	-1.26	0.04
33	5.42	25.4	-1.41	0.04
34	5.69	82.6	-1.47	0.04
35	5.61	22.5	-1.52	0.04
36	5.00	206.3	-0.84	0.04
37	4.60	120.3	-1.31	0.03
38	3.78	16.2	-1.71	0.03
39	3.22	52.3	-1.65	0.02
40	2.91	4.5	-1.88	0.02

Table 1: Main features of the 40 most massive progenitors (out of 159 total progenitors) in Au 27. The first column indicates the weight order, with 1 being associated with the most massive progenitor. Their stellar mass is found in the second column. The following columns correspond to the mean radius centred at the Galaxy center and the mean $[M/H]$ values. The last one shows the contribution of each progenitor, that is, the fraction of star particles that belong to each one of them over the total number of accreted star particles in the entire radius and metallicity range. We will study a dataset constituted by the 4 most massive progenitors (highlighted in blue) and another dataset that only includes 4 intermediate mass progenitors (highlighted in green) in Sections 4.3 and 5.1.

We are interested in studying several radii and metallicity ranges since, as it can be seen in Fig. 7, the contribution of the different progenitors is highly dependent on the selected ranges of these quantities. In this figure, the fraction of star particles belonging to each of the 4 most massive progenitors (see their masses in Table 1) is presented, as well as the fraction of star particles that once belonged to other disrupted satellite galaxies. As expected, the dominance of the most massive progenitors is greater toward higher metallicity values, while the dominance of the other progenitors increases as we go towards a lower metallicity range. Furthermore, as it was already shown in Table 1 and Fig. 6, even when considering the entire radius and metallicity range, the dataset is largely dominated by these 4 most massive progenitors, so we might need to determine their formation history before analysing the less massive ones.

Since we expect $E_{tot}\,vs\,L_z$ to constitute the space where the initial clumping is better preserved even after spatial-mixing, we use it to visualize the distribution of the star particles that belonged to different progenitors.

The distribution of the points that belong to each one of the 4 most massive progenitors in the $E_{tot}\,vs\,L_z$ space in different radius and metallicity ranges is displayed in Fig. 8. It can be seen that progenitor number 3 (blue) dominates the lower L_z side and progenitors 1 (red) and 2 (green) are super-imposed in the middle part of the $E_{tot}\,vs\,L_z$ space, in a similar region to where the Gaia-Enceladus-Sausage structure in the Milky Way is found. Progenitor 4 (orange), on the other hand, is constituted by several overdensities towards the larger L_z values extending across the entire E_{tot} range, which may be caused by the total energy and angular momentum of the progenitors not being conserved during an interaction. Regarding the entire distribution, the most drastic discrepancies are seen in the different radius bins, with the scattering of the data along the energy range decreasing as the radius increases. The reason why this happens is the diminution of the depth of the gravitational well with the radius. On the other hand, as it has been already explained, the larger the $[M/H]$ value, the more massive the progenitors tend to be. Star particles that once belonged to the same satellite galaxy appear in different clumps that occupy a large area of the $E_{tot}\,vs\,L_z$ and there is a noticeable overlap of the different accreted structures. Though the use of a kinematics-related space of more than 2 dimensions alleviates this fact, and despite some clearly visible substructures, it becomes evident that the clustering task is far from trivial. A certain over-density can be a consequence of the overlap of star points originated from different satellites (as for progenitor 1 and progenitor 2), and reversely, many clumps can be the signature of a single accretion event (as for progenitor 4). Moreover, it is clear that the location and shape of the clusters depend on the radius and metallicity ranges that are being considered. It must also be noted that not all of the accreted star particles are distributed in clumps, since some of them present a more diffuse distribution, with no possible over-density with which they may be associated.

Finally, the fact that the contribution of smaller progenitors is really small in comparison to the most massive ones would imply that they are harder to recover when applying the BGMM to the entire dataset. We are interested in seeing if the integrals of motion space make it easier to recover the progenitors of lower-mass accretion events. The main difference between this case with regards to the most massive satellite galaxies is that, since dynamical friction depends on their mass, it is expected that the location of the stars in the energy and angular momentum space to be very well conserved over time. The $E_{tot}\,vs\,L_z$ space distribution when considering intermediate-mass progenitors (12th, 19th, 26th and 33rd most massive progenitors, with from $1 \times 10^6 M_\odot$ to $1 \times 10^8 M_\odot$, as it can be seen in Table 1) is shown in Fig. C.1.

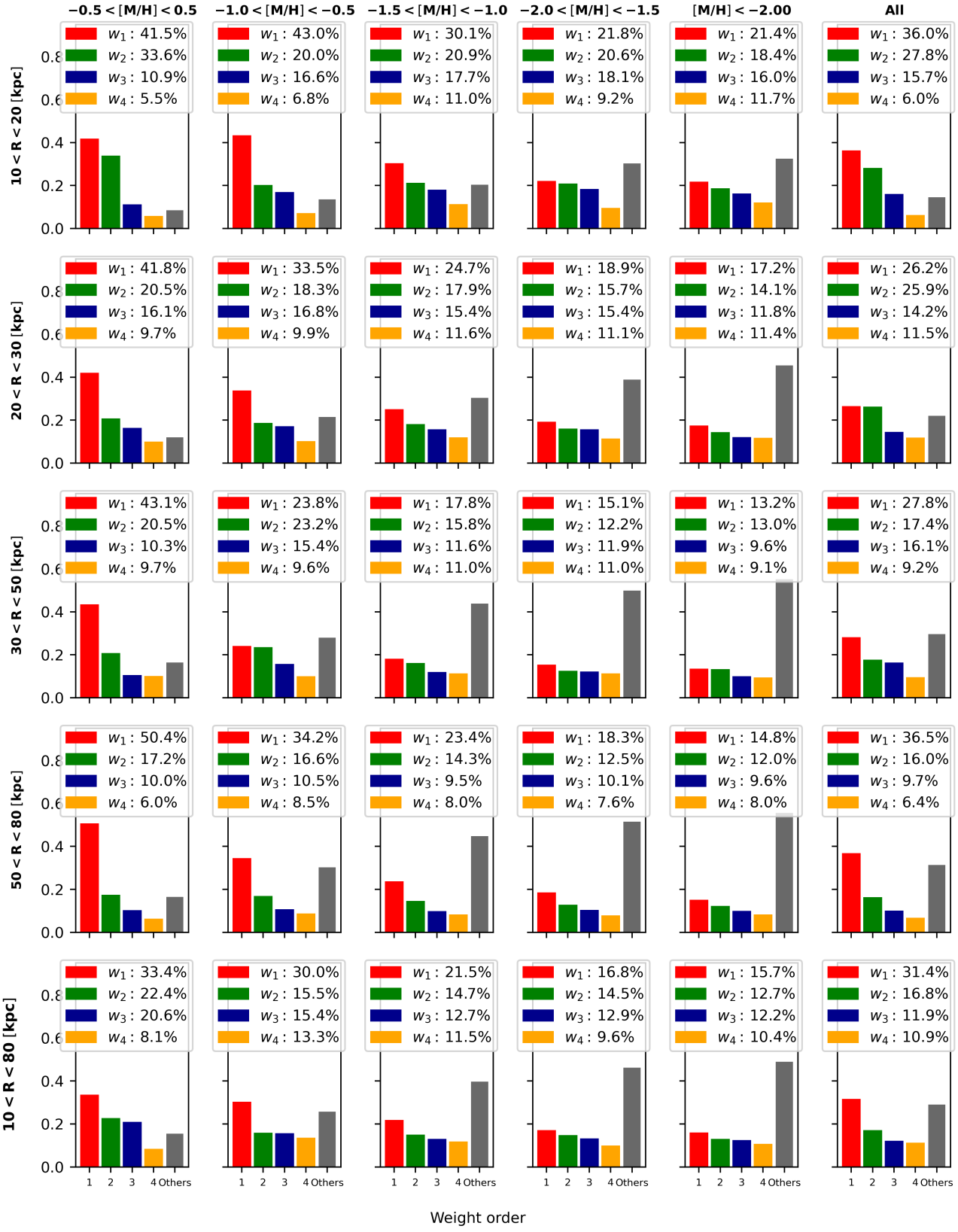

Fig. 7: Histogram of the contribution of each one of the 4 most massive progenitors in Au 27 in different distance (rows) and metallicity (columns) ranges, as indicated in the titles of the figure panels. These progenitors are color-coded as follows, in decreasing weight order: red, green, blue and orange, with the percentual contribution of the smaller-mass progenitors being shown in grey. The exact value of the contribution of each progenitor is included in the legend of each subfigure.

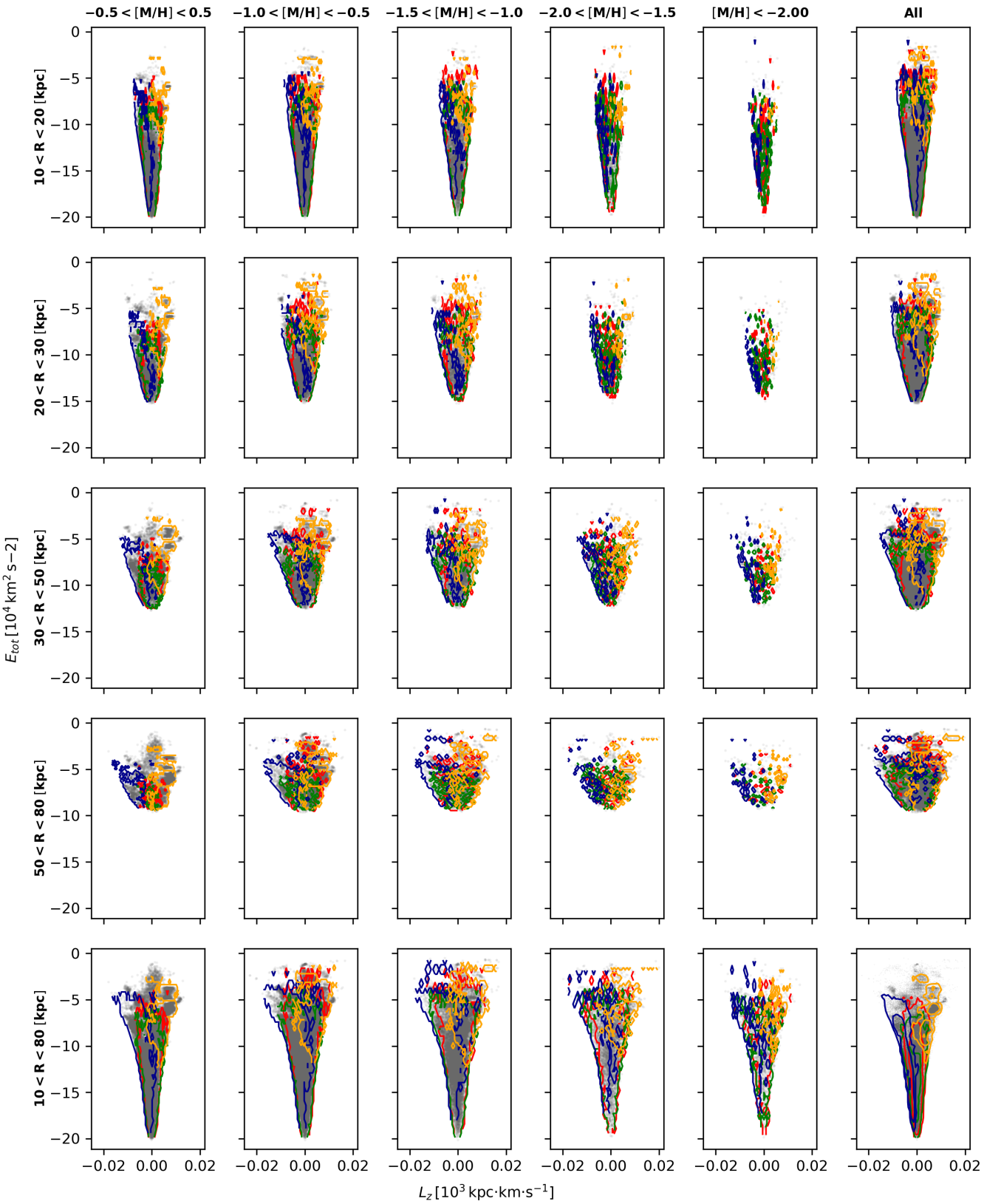

Fig. 8: $E_{tot} - L_z$ contour plots of the star particles that belonged to the 4 most massive progenitors in different distance and metallicity bins on top of the density distribution of the entire dataset. They appear color-coded as in Fig. 7.

4.3. Analysis of the different spaces

The next step to be taken would be to study the star point distribution in other spaces we might be interested in using. However, since it is really difficult to find physical arguments that motivate the selection of certain quantities over others in order to apply the BGMM, a visual inspection of these parameters must be carried out.

It is important to understand the behaviour of the different progenitors in the set of quantities we can consider applying the BGMM to. Thus, in order to analyze the star particles clustering in the different spaces, some of the most relevant quantities (total energy, angular momentum, actions, velocities, spatial distribution and metallicity) are represented for the entire radius and metallicity range.

Fig. 9 presents these spaces for the 4 most massive progenitors. It is evident that there is, indeed, clustering in each one of the spaces that are being shown, with the metallicity distribution presenting quite some similarities for all 4 progenitors. However, the clustering is more noticeable in the $E_{tot}\,vs\,L_z$, $L_\perp\,vs\,L_z$, $v_\phi\,vs\,v_r$ and $\frac{(J_z-J_r)}{J_{tot}}\,vs\,\frac{J_\phi}{J_{tot}}$ spaces, where the star particles that belong to different progenitors have less overlapping. Therefore, we would expect them to be useful to try to find clusters and, while doing so, be able to unravel the process of the accretion history of the Milky Way. At the same time, the severity of the overlap between the different progenitors becomes more than obvious by looking at the different distributions and because of the fact that the $[M/H]$ histograms also present strong similarities. As a consequence, Fig. 9 provides more clear evidence about the intrinsic difficulties of identifying the different progenitors as clumps, regardless of the method that is selected to find them.

On the other hand, Fig. 10 shows the equivalent plot but for the intermediate-mass progenitors already mentioned in the previous section (see Fig. C.1), which leads to different conclusions. The clumping of the different star particles is more noticeable in the $E_{tot}\,vs\,L_z$ and $L_\perp\,vs\,L_z$ spaces than in the case of the most massive progenitors and, at the same time, these intermediate-mass progenitors are more separated in their properties. It can be seen that, when considering only these satellite galaxies, only progenitors 26 and 33 (blue and orange in both Fig. 10 and Fig. C.1) overlap. Thus, in this case, retrieving the different overdensities by using a clustering method seems like a more plausible idea. Nonetheless, it also further shows that resolving a single Gaussian shape for each progenitor is not a trivial, if even possible, task.

Our first idea was to apply the BGMM to conserved quantities of accreted halo populations, which we expect to be distinguishable from each other by means of their different orbital and chemical properties. However, we have seen that the analysis of the different quantities shows that, especially for the most massive progenitors, each one of them leads to several clumps. In addition, there is a non-negligible degree of overlap in all studied cases that varies when considering different radius and total metallicity values, as seen in Section 4.2. As a consequence, in order to obtain a reasonable model that allows us to recover the different over-densities, the predicted number of Gaussians G will not match the number of progenitor galaxies in the input data. Furthermore, the quantities and radius and metallicity bins that are used to apply the BGMM must be selected carefully.

Therefore, in order to select the quantities in whose subspaces we pretend to look for substructures, the $50 < R < 80$ kpc bin in a subdataset with only the 4 most massive progenitors (see Fig. 9) has been selected, since the $E_{tot}\,vs\,L_z$ space in this bin includes the smaller degree of overlap between the different progenitors and includes several overdensities that are clearly distinguishable (see Fig. 8). The BGMM is then applied to 4 different combinations of quantities in order to compare the results of the different models, as can be seen in Appendix D. This has led us to the conclusion that, since the results are not significantly improved when considering more than the E_{tot}, L_z and $L_\perp$ values and for computational efficiency reasons, it is best to stick to these quantities to search for substructures. In addition, even when considering this subdataset where the

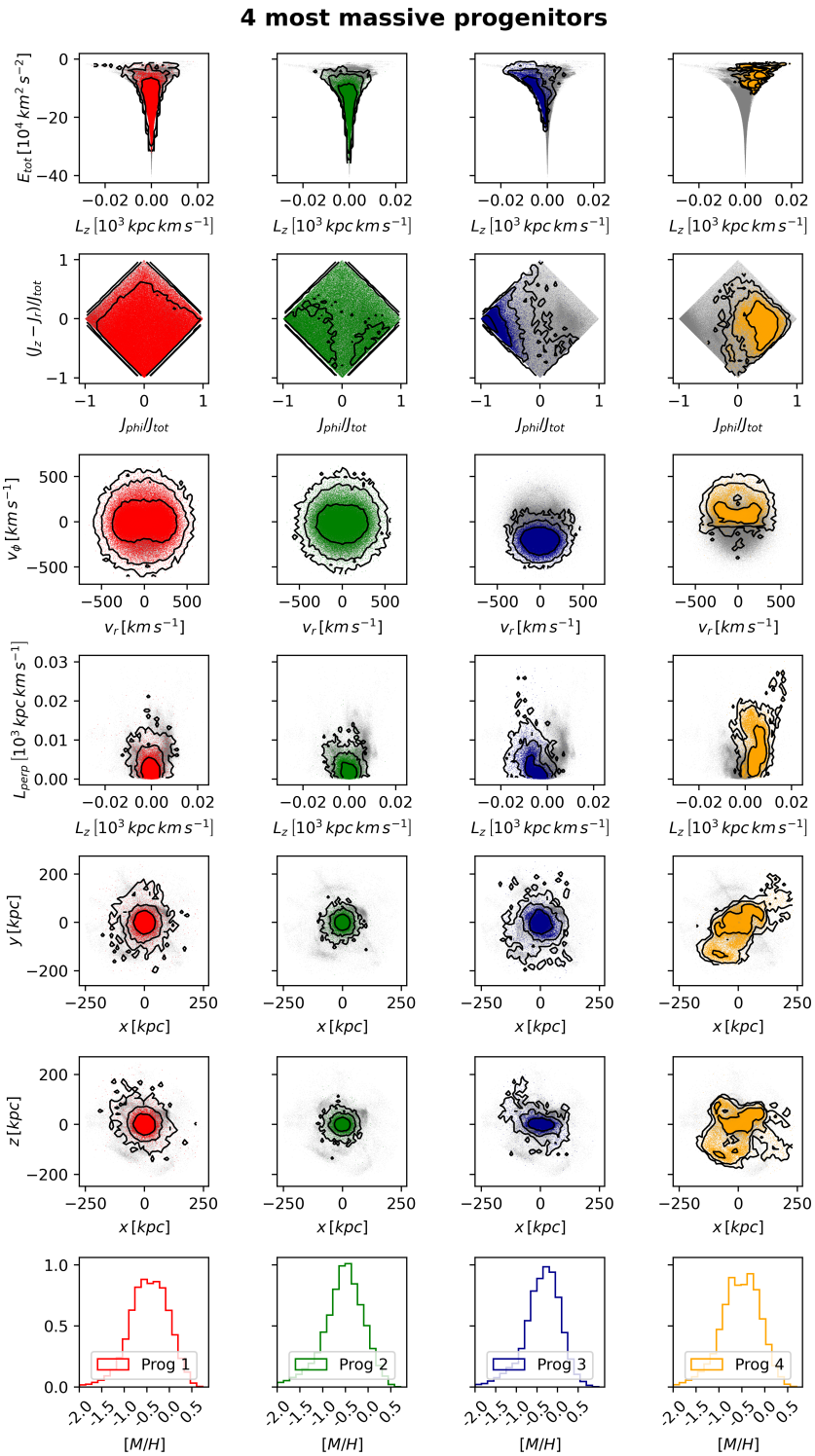

Fig. 9: Distribution of star particles for the 4 most massive progenitors (from largest to smallest) for Au 27 of the Auriga simulations in the integral of motion, spatial coordinate and velocity spaces, along with their metallicity distribution. The entire distribution of the entire accreted particles from Au 27 is plotted in grey, while the points that are associated with the same progenitor appear in a different color. The contour lines are used to show the star particle density in the different regions, with each column corresponding to each one of the 4 most massive progenitors in descending order of mass.

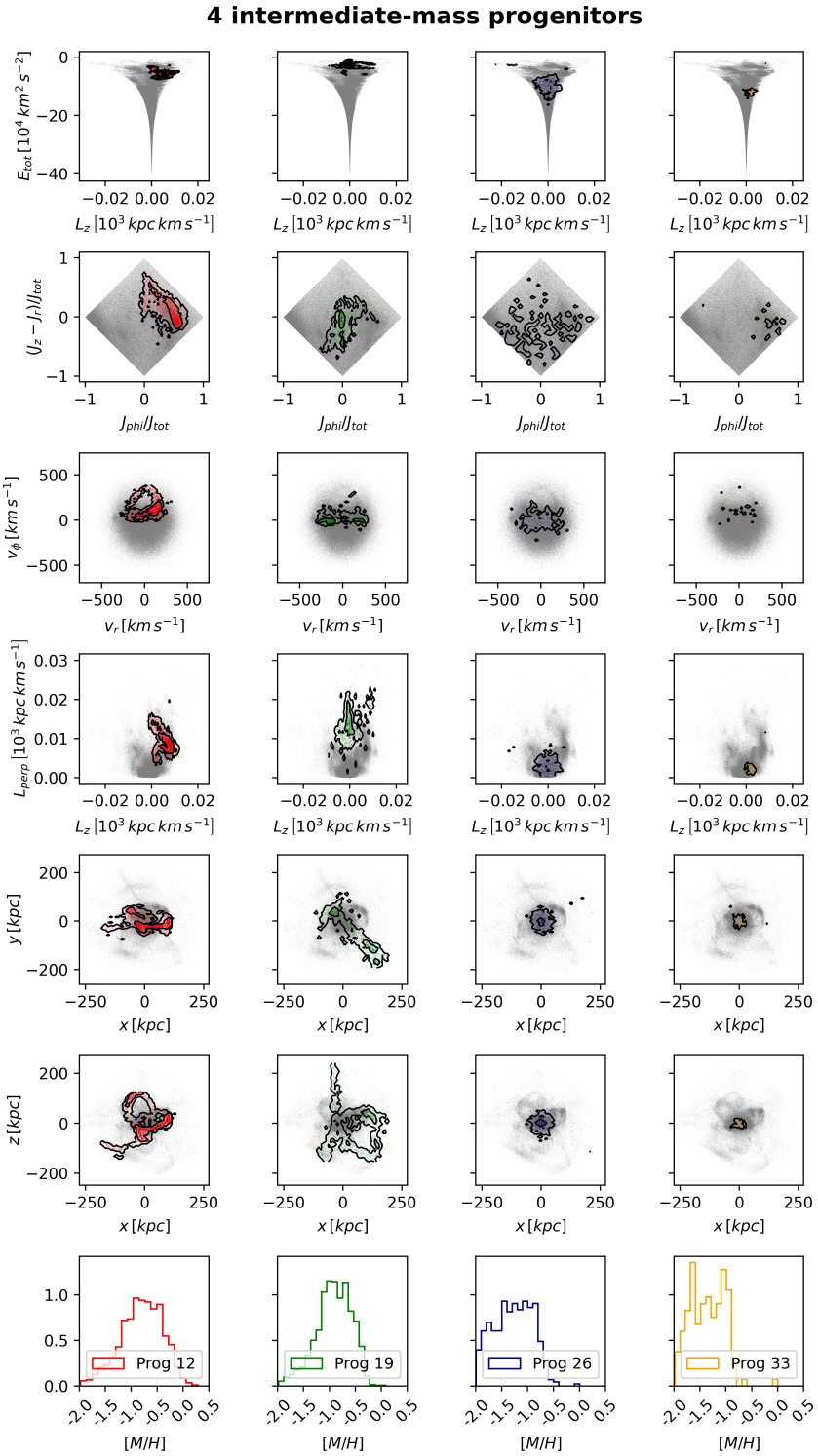

Fig. 10: Distribution of star particles for the 12th, 19th, 26th and 33rd most massive progenitors (from largest to smallest) for Au 27 of the Auriga simulations. Different spaces are shown, as well as a total metallicity histogram for each one of them.

progenitors are easier to discern, a number of Gaussians significantly larger than the number of progenitors being analyzed is retrieved. As a consequence, from now on, we will focus our study on progenitors with different mass ranges but in the entire radius and metallicity ranges, since we have definitely ruled out the possibility of assigning a single Gaussian to each progenitor, with the goal of recovering all of the individual overdensities.

5. Results

The BGMM have been built for several different datasets in order to see how they perform in different radius and metallicity ranges, as well as with progenitors with different masses. However, here only two representative datasets are considered in order to simplify the study: for the 4 most massive satellites and for 4 intermediate-mass progenitors, both in the entire metallicity and radius range.

Moreover, it is worth mentioning that, since the overlap of the different star points in the spaces that are being considered is hard to visualize, a Python script has been developed in order to create what we have called a dominance figure (see 4th and 5th columns of Fig. 11). It allows us to divide each distribution into several 2D bins and, then, show every bin with the color that is associated with the progenitor or Gaussian that contributes to a larger number of star particles in said bin. These types of figures can be used to see the relevance of each progenitor/Gaussian in every region of the different spaces. Along with the contour plots, it makes it easier to see both the overlap and the Gaussian that should be assigned to each one of the progenitors.

5.1. Bayesian Gaussian Mixture results

Given their dominance, we first apply the BGMM to the 4 most massive progenitors. As it has been discussed in Section 4.3 and shown in Fig. 8, one single progenitor is associated with several overdensities, so we must find the different overdensities and, then, develop a method that will allow us to relate them to each other. Therefore, in order to do so, we select a large number of components ($G_{max} = 50$) for the best possible BGMM outcome despite the computational cost (that is, with a small convergence threshold, $tol = 1 \times 10^{-5}$) and with essentially no lower limit on the cluster size ($reg = 1 \times 10^{-6}$).

Fig. 11 presents the results obtained with the BGMM for the 40 % of the dataset associated with the 4 most massive progenitors in the $E_{tot}\, vs\, L_z\, vs\, L_\perp$ phase space, with this selection being motivated by the limitation on the computational power. The prediction of such a large number of Gaussian ($G = 18$) seems to be adequate, as it can be seen both in the second column panels of Fig. 11 and on Fig. 8, and due to the great resemblance between the original density panel and the generated sample density panel. Moreover, the entanglement of the different satellite galaxies contributes to the need to have a large G in order to fit the data. In any case, 18 Gaussians are found for a dataset with 4 progenitors yet; at the same time, eye inspection of all three spaces ($E_{tot}\, vs\, L_z$, $E_{tot}\, vs\, L_\perp$ and $L_\perp\, vs\, L_z$) may lead to the recovery of a different number of substructures. This shows that the model identifies more than the most obvious clumps. The degree of overlap in the second column panel, corresponding to the distribution of the different progenitors, is far larger than for the predicted clusters, shown in the third column. Nevertheless, in both cases, the dominance figures are representative of the contribution of each progenitor or Gaussian in each one of the 2D histogram bins, since the cluster that gives the color of each one of them contributes with > 75 % of the star particles in nearly all cases. Therefore, they are useful to see which Gaussians should be associated with the satellite galaxies. Moreover, it can be seen that the star particles found on the outskirts of the distributions are associated with Gaussians with small weights. A more in-depth comparison between the original density plot and the generated sample density panels (see 6th and 7th columns of Fig. 8) indicates that the models are capable of reproducing the different overdensities in all three planes except for the upper right region of the $E_{tot}\, vs\, L_z$ and $E_{tot}\, vs\, L_\perp$ spaces, where some overdensities that are elongated along the horizontal axis are not reproduced. Thus, this demonstrates that the model could be improved by decreasing the tolerance and regularization values and/or increasing the number of K-Means initializations or the maximum number of EM iterations, though a large computational efficiency would be needed to this end.

On the other hand, Fig. 12 shows the results that correspond to the entirety of the dataset constituted by 4 intermediate-mass satellites presented in Fig. 10, in a figure that is equivalent to Fig. 11. Though, in this case, the overlap of the different progenitors is still present, along with a single progenitor giving rise to more than a single over-density, these effects are less severe than for the 4 most massive progenitors. The resemblance

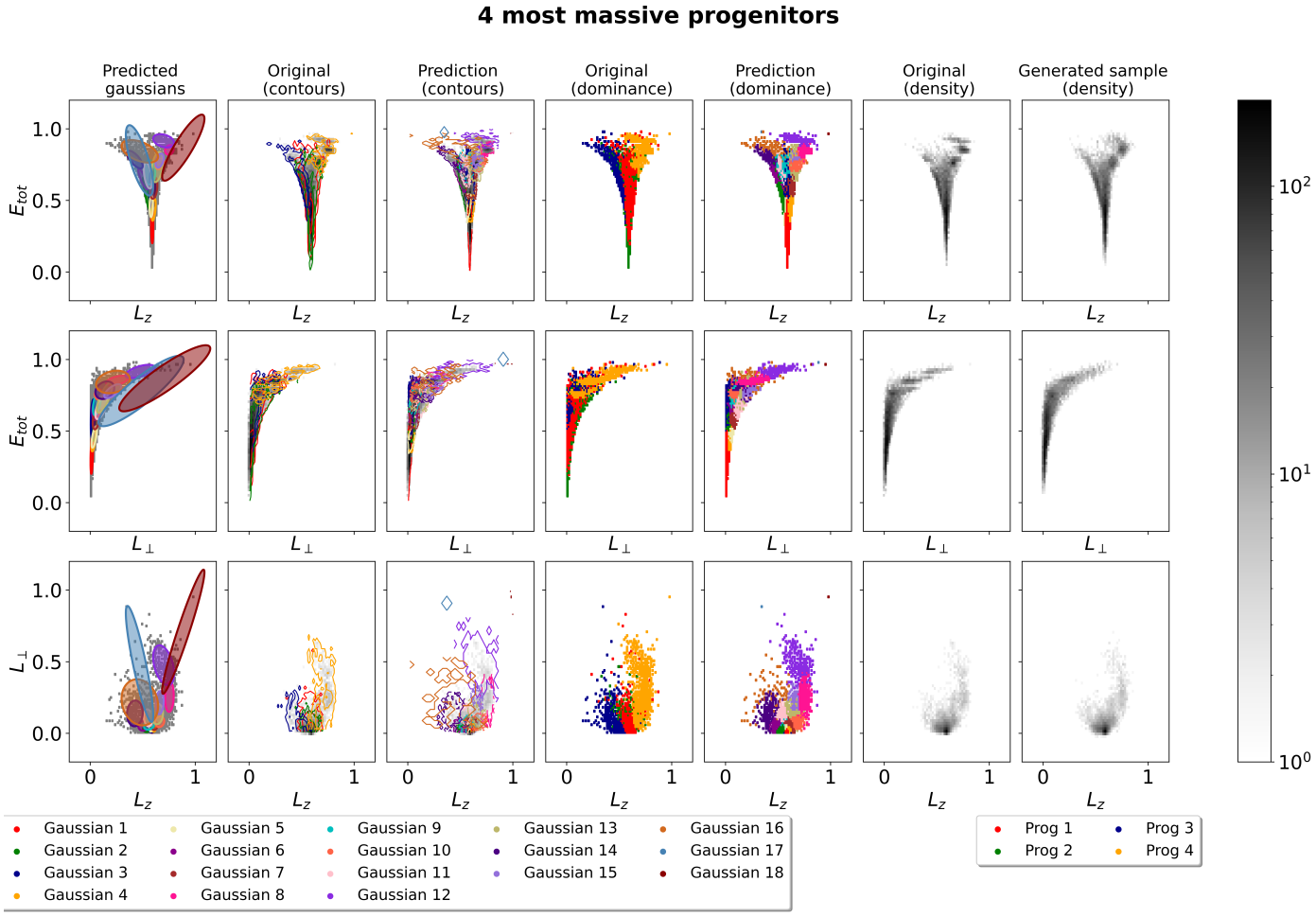

Fig. 11: BGMM results for the 4 most massive progenitors in the entire metallicity and radius range in all 3 planes: E_{tot} *vs* L_z, E_{tot} *vs* $L_\perp$ and $L_\perp$ *vs* L_z. Their values have been normalized to simplify the analysis of the Gaussian parameters **First column:** Star particles' distribution in 2D histograms overlaid by the Gaussian distribution found by the BGMM, with each Gaussian in a different color, as indicated in the bottom left legend. **Second column:** Contour distribution of the star particles color coded by the progenitors (Prog) they are associated with, as shown in the bottom right legend. **Third column:** Same plots as in the second column but with the color-coding associated with the Gaussian the model predicts they belong to. **Fourth column:** Dominance panel, made by using the progenitors each star particle actually belongs to, but obtained by dividing the entire subfigure into different 2D bins so that each bin appears with the color of the progenitor that has the larger number of particles in said bin. **Fifth** Equivalent panels to the fourth column for the predicted clusters. **Sixth column:** 2D histogram showing the density distribution of the star particles for the original dataset. **Seventh column:** Same as in the sixth column but for the sample generated by the BGMM. Both the progenitors and the Gaussians appear in decreasing weight order in the legends.

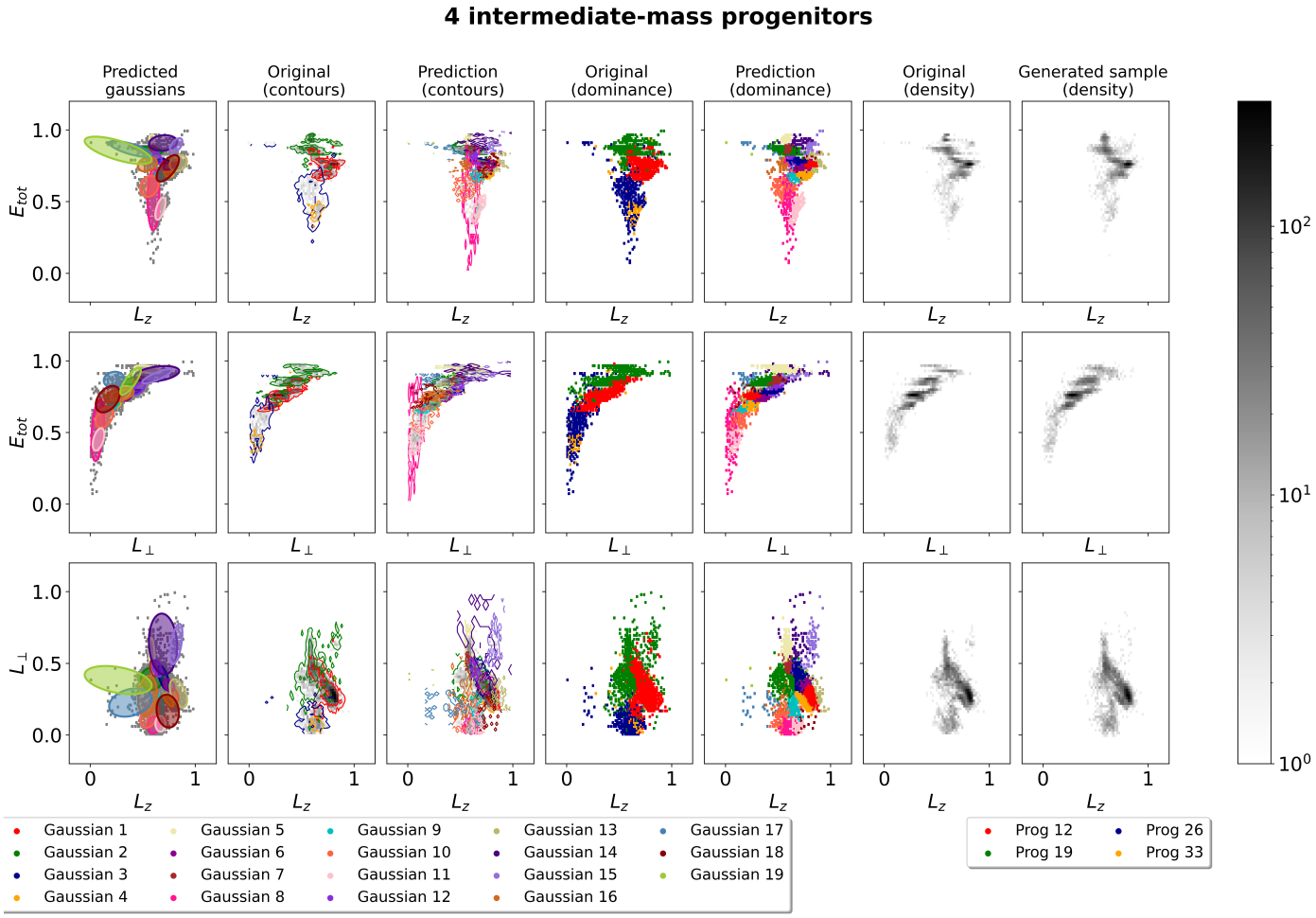

Fig. 12: BGMM results for 4 intermediate-mass progenitors in the entire metallicity and radius range in all 3 planes: E_{tot} *vs* L_z, E_{tot} *vs* $L_\perp$ and $L_\perp$ *vs* L_z. This figure is structured as Fig. 11.

between the original density distribution and the generated sample density plot is much larger, with the last two columns of Fig. 11 being hardly discernible except for the upper right region of E_{tot} *vs* L_z and E_{tot} *vs* $L_\perp$ spaces, where two close overdensities in the original distribution appear as a single clump in the generated one. Nevertheless, we again recover a large number of Gaussians ($G = 19$), which is unavoidable in order to properly fit our data.

This study has demonstrated that, even when considering the whole dataset for which we have available data and satellites with very different masses, the BGMM is able to recover the different overdensities as independent Gaussians, though it is not possible to assign a single Gaussian to each of the progenitors. Therefore, it indicates that the kinematical information is not enough to reproduce the merger events in a Milky Way's analogue outer halo and an extra step must be taken in order to relate the different predicted clusters together. However, due to a smaller degree of entanglement and the fact that each progenitor is associated with a smaller number of clumps in the intermediate-mass case than in the most massive progenitors case, the first one seems to be more promising for reproducing the galaxy structure by means of the method that has been developed in the present work.

5.2. Linking different Gaussians to a single event

The previous analysis still leaves our main question unanswered: how many independent clusters do we actually have?

In order to get an answer to this issue, the potential relations between the clusters that have been found must be considered, for which the separation between the Gaussians in all three dimensions and the internal hierarchy in the clustering spaces between clusters are used. That is, since it has been seen that it is impossible to try to reproduce the merging history by using one Gaussian per progenitor, the classification is refined by performing hiearchical/agglomerative clustering on the input data.

The distance metric that has been selected in order to try to link the different clusters is the Mahalanobis distances between the centres of each Gaussian. This distance metric has the advantage over other distance types, like the Euclidean distance, that it takes into account the correlation that exists between the different variables, since it includes information provided by the covariance matrix. It is defined as follows:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \tag{18}$$

where $\mathbf{x}$ is the vector of the quantities that are being considered, $\boldsymbol{\mu}$ the mean value of the Gaussians, and $\boldsymbol{\Sigma}$ is the covariance matrix.

The linkage method works as follows: we start by having a forest of clusters and, according to the Mahalanobis distance between two clusters, s and t, and the linkage method, if the corresponding condition is satisfied, they are combined into a new cluster, u, after which s and t are replaced in the forest of clusters by u, with the algorithm being stopped when only one cluster remains in the forest. The two closest clusters are associated on each iteration for all linkage methods, though the distance between the resulting cluster u and the remaining clusters are different for each one of the different linkage methods.

Several linkage methods have been tested in order to compute the distance between the different Gaussians. They have been applied by using the `SciPy` open-source software (Virtanen et al. 2020). The so-called *weighted* method has been favoured over others because it has been determined to be the best one for this particular case, since the nearest two clusters are combined and, then, the arithmetic mean of the distances between each one of these two clusters and the clusters that remain in the forest is computed. Thus, each of the clusters that had been previously linked has the same relevance when calculating the new distance.

In other words, the distance between the new cluster u and each forest cluster v for the *Weighted or Weighted Pair Group Method with Arithmetic Mean* is the following (Müllner 2011):

$$d(u,v) = \frac{D_M(s_{mean}, v_{mean}) + D_M(t_{mean}, v_{mean})}{2}, \tag{19}$$

Since we would expect the star particles that belong to the same substructure to show similar metallicity distribution functions (see Naidu et al. 2020), we include the metallicity information in this step by adding an extra dimension to the Gaussian parameters. That is, the mean $[M/H]$ values of each predicted Gaussian are incorporated in the mean values vector and their variance are added in the diagonal of the covariance matrix, with the $Cov([M/H], X_k)$ values (with k being each of the n dimensions to which the BGMM was applied to) being null. Afterwards, these new mean vectors and covariance matrices are used to find the Mahalanobis distances.

In order to visualize the results, a dendogram has been used, that is, a diagram with a tree shape that allows us to illustrate the hierarchical clustering results. With this type of representation, we are able to see how close

the Gaussians are according to the Mahalanobis distance that separates them. The most similar clusters are the first that appear joined together at the bottom part of the representation, and then, at a larger height, the next two more similar clusters, among the remaining original Gaussians and the result of joining the two most similar ones, are found, and so on. In addition, a distance threshold is selected so that the linkage results are as good as they can be to reproduce the original labels of the dataset that is considered.

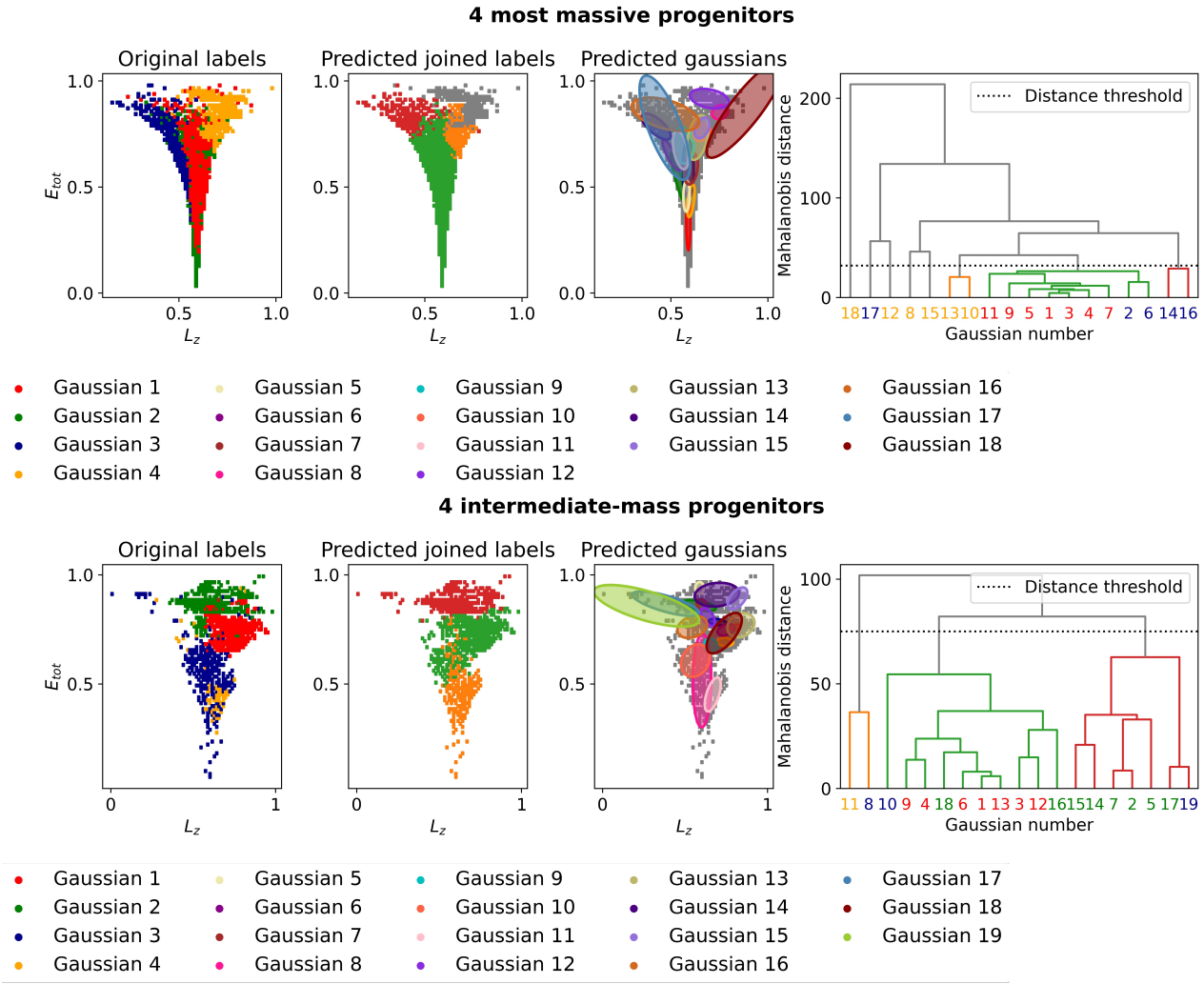

Fig. 13: Results of the linkage method for the 4 most massive progenitors (top row) and 4 intermediate-mass progenitors (bottom row). *From left to right:* **(a):** 2D histogram with the bins color-coded according to the dominant progenitor of each bin in the E_{tot} *vs* L_z space. **(b):** Same panel as in panel **(a)**, but color-coded according to the dominant Gaussian of each bin. **(c):** Distribution of the predicted Gaussians. **(d):** Obtained dendogram. The Gaussian numbers appear in the dendogram color-coded by the progenitor with a larger contribution of star particles to each one of them. The distance threshold is shown as a horizontal line.

This study has been carried for both the set of the 4 most massive progenitors and the 4 intermediate-mass progenitors, both when considering only the BGMM results and when adding to the mean and covariance matrices values the information regarding $[M/H]$. However, the results are essentially the same in these two cases, so only the outputs of the first scenario are presented.

The results are found in Fig. 13. The dendogram obtained for the 4 most massive progenitors shows that, even if we select a low distance threshold, it is not possible to link the independent Gaussians together and still obtain predicted joined labels with a larger degree of resemblance to the original labels than the one that is presented. Though in the lower part of the $E_{tot}\, vs\, L_z$ space we might identify the red area in the left panel with the green one in the following, the distribution of the other linked clumps does not match the distribution of the original progenitors. On the other hand, for the 4 intermediate-mass progenitors, one can see that the green progenitor in the original labels panel resembles the red cluster in the predicted joined labels one, while the red and blue in the first panel somewhat resemble the green and orange, respectively, in the second one. Despite the weighted linkage method leading us to join, for example, Gaussians 17 and 19, and 8 and 11, which belong to different progenitors, it is clear that the results are more plausible than for the most massive progenitors.

In conclusion, the procedure that has been followed is valid to identify the different clumps in the spaces of the quantities that we expect to contain signatures of accretion by means of the Bayesian Gaussian Mixture model. However, it is hard to link them together, especially for the most massive progenitors, whose clustering is less evident than in the case of the intermediate-mass ones. In this last case, however, the linking methods might be a promising tool to relate the different Gaussians to each other and, ultimately, identify the disrupted satellite galaxies as a group of several clumps with similar characteristics.

A way to refine this method would be to make the dendograms in two steps: joining the Gaussians with a larger weight in the first step and later adding the smallest clusters. The development of this next step would be the continuation of the current project.

6. Conclusions

In this work, we have tried to develop a clustering method based on the Bayesian Gaussian Mixture Model (BGMM) in order to unravel the merging history of the Milky Way's stellar halo so that we can study the physical processes that led to its current state. After developing a BGMM with some toy datasets generated from Gaussians with known parameters, Auriga simulations have been used to further develop this technique for a Milky Way's analogue galaxy, using the integrals of motion space to try to find clumps that could correspond to the different satellite galaxies whose accretion events built the simulated halo, with the hope that this same method could be applied to observational Milky Way data.

This work suggests that the use of the BGMM by using only kinematics in order to reproduce the merger history is not sufficient, even though the integral of motion spaces are rich in substructures. The overlap of the different progenitors, along with the presence of several overdensities per disrupted galaxy in the integrals of motion space, makes it hard to distinguish between them. At the same time, determining the number of accretion events that led to the state of the Galaxy nowadays has been proven to be a very challenging task, even when considering different quantities we expect to be conserved over time. Therefore, the different substructures in the kinematic-related spaces are not efficient enough to reveal the origin of the stars that formed them. That is, the origin of the different substructures cannot be reproduced only by finding Gaussian shapes in the $E_{tot} - L_z - L_\perp$ space, which has been selected as the primary workspace because of the conservation, to a certain extent, of these quantities in the case in which we have an axi-symmetric potential.

Therefore, the new aim of this project has been to try to determine the different overdensities in that same space as clumps to later relate them to each other by using the Mahalanobis distances between the centers of each of the predicted Gaussians.

In the case of the most massive progenitors, the outcome of the developed method has proven to be far from capable of discerning the different clumps so that the original progenitor satellites are recovered. Nevertheless, this result is significantly improved when using smaller progenitors and a linkage method that allows us to relate different Gaussians together and to determine that they once belonged to the same structure. This would be consistent with the fact that the integrals of motion are not fully conserved for the most massive satellites, with this effect being less relevant in smaller progenitors. In the future, we will be comparing the E_{tot}, L_z and $L_\perp$ values of stars in a given satellite at the time of merging with the host value with those observed at $z = 0$ in order to test this hypothesis. Moreover, a new method may be needed in order to recover the most massive progenitors and then apply clustering methods to the smaller ones.

The main outcome of this study is that we might want to rethink the often used strategy of using the integrals of motions' phase space in order to study the history of the Milky Way. Identifying the accretion signatures as clumps in those spaces carries a great difficulty. Firstly, because the clustering might be questionable, especially for the most massive progenitors, due to the large degree of overlap they show. Secondly, it has also been demonstrated that each progenitor usually gives rise to more than one overdensity, so it would be essential to take an extra step and relate the different clumps to each other in order to determine which ones have a common origin. In addition, the results are different depending on the mass range being considered, with the most massive progenitors presenting a greater challenge. This fact, along with the large amount of overlap between the different substructures, poses a great obstacle, since the most massive progenitors dominate the data sample and prevent us from finding the smaller, easier to find, merger events.

The results that have been obtained differ from most of the previous works on the topic, though some other previous works have led to some controversy on this particular topic (e.g, Jean-Baptiste et al. 2017). They also found that the use of total energy, angular momentum in the vertical axis and perpendicular angular

momentum along with other spaces in which clustering of the stars that once belonged to the same progenitor with magneto-hydrodynamic galaxy simulations is not adequate to reproduce the accretion history of the Milky Way by analysing the distribution in said spaces. Therefore, the current project may serve as a word of caution regarding the foundations of the assumptions they are based on.

We might also question the hierarchical formation of the structures in the Milky Way, since it is unclear if accretion is the dominant accretion in this procedure, as it is possible that not all star points may have once belonged to a certain progenitor. However, the most recent observational data seems to be in quantitative agreement with the ΛCDM models of galaxy formation. In addition, it must be noted that these conclusions have been obtained by using simulations that do not include observational effects that mimic the actual data, such as the AuriGaia simulations, so the next step in this study would be to consider these mock catalogues.

Moreover, a more in-depth study of not only the kinematics but also more detailed chemical information, as well as the stars' ages, would be essential in order to decipher the accretion history of the Milky Way's analogue simulations and to develop a method that will ultimately allow us to study the merging processes that ended up building the current state of our home galaxy.

Acknowledments

The student and her supervisors thank Robert Grand, PI of the Auriga project, for sharing the Auriga simulations, and for his support with handling and interpreting them, as well as for the discussions on the results obtained.

References

Basurto, J. 2022, Skiing the Galaxy

Binney, J. & Tremaine, S. 2008, Galactic Dynamics (Princeton University Press)

Borsato, N. W., Martell, S. L., & Simpson, J. D. 2020, Monthly Notices of the Royal Astronomical Society, 492, 1370

Bullock, J. S. & Johnston, K. V. 2005, The Astrophysical Journal, 635, 931

Casamiquela, L., Castro-Ginard, A., Anders, F., & Soubiran, C. 2021, Astronomy and Astrophysics, 654, A151

Chlon, L. 2020, Dirichlet Process Gaussian Mixture Models made easy, [Online; accessed March 28, 2022]

Daniel López. 2019, Panoramic view of the Gran Telescopio de Canarias (GTC), [Online; accessed March 4, 2022]

Duda, R. O. & Hart, P. E. 1973, Pattern Classification and Scene Analysis (New Yotk: John Willey & Sons)

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, 226–231

Fattahi, A., Belokurov, V., Deason, A. J., et al. 2019, Monthly Notices of the Royal Astronomical Society, 484, 4471

Freeman, K. & Bland-Hawthorn, J. 2002

Grand, R. J. J., Deason, A. J., White, S. D. M., et al. 2019, Monthly Notices of the Royal Astronomical Society: Letters, 487, L72

Grand, R. J. J., Gómez, F. A., Marinacci, F., et al. 2017, Monthly Notices of the Royal Astronomical Society, stx071

Grand, R. J. J., Helly, J., Fattahi, A., et al. 2018

Helmi & de Zeeuw. 2000, Monthly Notices of the Royal Astronomical Society, 319, 657

Helmi, A. 2020, Annual Review of Astronomy and Astrophysics, 58, 205

Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85

Helmi, A., White, S. D. M., de Zeeuw, P. T., & Zhao, H. 1999, Nature, 402, 53

Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, Nature, 370, 194

Ivezic, Z., Connolly, A., Vanderplas, J. T., & Gray, A. 2020, Statistics, data mining, and machine learning in astronomy: A practical python guide for the analysis of survey data (Princeton University Press)

Jean-Baptiste, I., Matteo, P. D., Haywood, M., et al. 2017, Astronomy and Astrophysics, 604, A106

Koppelman, H. H., Helmi, A., Massari, D., Price-Whelan, A. M., & Starkenburg, T. K. 2019

Lee, D. M., Johnston, K. V., Sen, B., & Jessop, W. 2015, The Astrophysical Journal, 802, 48

Lloyd, S. 1982, IEEE Transactions on Information Theory, 28, 129

Lövdal, S. S., Ruiz-Lara, T., Koppelman, H. H., et al. 2022, arXiv preprint arXiv:2201.02404

Malhan, K., Ibata, R. A., Sharma, S., et al. 2022

McMillan, P. J. & Binney, J. J. 2008, Monthly Notices of the Royal Astronomical Society, 390, 429

Müllner, D. 2011, Modern hierarchical, agglomerative clustering algorithms

Myeong, G. C., Evans, N. W., Belokurov, V., Sanders, J. L., & Koposov, S. E. 2018, Monthly Notices of the Royal Astronomical Society, 478, 5449

Myeong, G. C., Vasiliev, E., Iorio, G., Evans, N. W., & Belokurov, V. 2019

Naidu, R. P., Conroy, C., Bonaca, A., et al. 2020, The Astrophysical Journal, 901, 48

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Ruiz-Lara, T., Matsuno, T., Lövdal, S. S., et al. 2022, Substructure in the stellar halo near the Sun. II. Characterisation of independent structures

Tom Burns. 2021, Ancient cultures viewed Milky Way differently, [Online; accessed March 4, 2022]

V. Belokurov, Juan Carlos Muñoz (ESO). 2018, An impression of the encounter between the Milky Way and the Sausage dwarf galaxy., [Online; accessed March 20, 2022]

Vasiliev, E. 2018a

Vasiliev, E. 2018b

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261

White, S. D. M. & Rees, M. J. 1978, Monthly Notices of the Royal Astronomical Society, 183, 341

Yuan, Z., Chang, J., Banerjee, P., et al. 2018, The Astrophysical Journal, 863, 26

A. Expectation-Maximization Algorithm

If the log-likelihood is defined as it has been shown in Equation (5), its partial derivative with respect to $\boldsymbol{\theta}_j$, which can correspond to either $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ or α_j (respectively, the mean value, covariance matrix and weight), is defined as:

$$\frac{\partial lnL}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{N} \frac{\alpha_j}{\sum_{k=1}^{G} \alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \left[\frac{\partial \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}{\partial \boldsymbol{\theta}_j}\right], \tag{20}$$

where G refers to the number of Gaussians found by the model, N corresponds to the number of points in the sample and $\mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$ is each one of the Gaussian distributions.

Taking into account Equation (6), Equation (20) can be rewritten as it follows:

$$\frac{\partial lnL}{\partial \theta_j} = \sum_{i=1}^{N} \left[\frac{\alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}{\sum_{j=1}^{G} \alpha_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}\right] \left[\frac{1}{\mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \frac{\partial \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}{\partial \boldsymbol{\theta}_j}\right], \tag{21}$$

which is redefined as:

$$\frac{\partial lnL}{\partial \boldsymbol{\theta}_j} = -\sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right) \frac{\partial}{\partial \boldsymbol{\theta}_j} \left[ln\Sigma_j + \frac{\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^2}{2\boldsymbol{\Sigma}_j^2}\right], \tag{22}$$

If the derivatives of lnL with respect to $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are set to zero and the normalization condition is taken into account, the next set of parameters (also called estimators) are such as:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right) \mathbf{x}_i}{\sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right)} \tag{23}$$

$$\boldsymbol{\Sigma}_j^2 = \frac{\sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^2}{\sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right)} \tag{24}$$

$$\alpha_j = \frac{1}{N} \sum_{i=1}^{N} p\left(j|\mathbf{x}_i\right), \tag{25}$$

In the *maximization M-step*, we start with a certain guess for $p\left(j|\mathbf{x}_i\right)$ that allows us to determine the $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ and α_j by means of equations (23), (24) and (25), respectively. This is done in such a way that the algorithm does not depend on this initial guess and so that the parameters become closer to those that lead to the maximum possible value of lnL. This initial guess is made by using the assumption of different random components so that they are centered on random data points and, afterwards, we use the probability for each point of having been generated by every single one of the components.

Then, in the *expectation E-step*, Equation (6) is used to compute the $p\left(j|\mathbf{x}_i\right)$ to update the parameters at step $N+1$ and the process is repeated until the local maximum value of lnL is found (Ivezic et al. 2020).

B. Classical and Bayesian Gaussian Mixture comparison

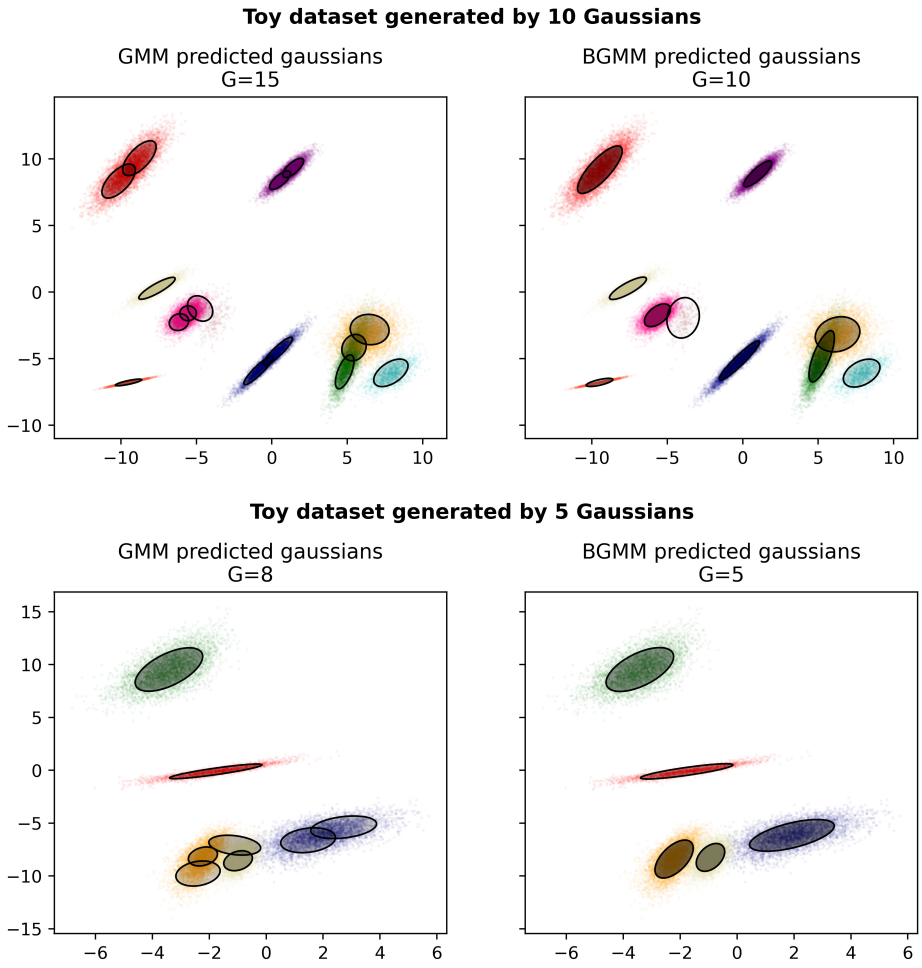

Fig. B.1: Comparison of the GMM and BGMM results for two different toy datasets, one per row. The Gaussian distributions appear in different colors and the shape of the predicted Gaussians is shown on top of said distributions. The greater the weight of the predicted Gaussians, the more opaque their interior appears. The number of predicted Gaussians is denoted as G. For the top row toy dataset, G_{max} has been set to 15 while, for the bottom row dataset, $G_{max} = 8$.

Fig. B.1 shows the comparison of the results of the classical (GMM) and the Bayesian Gaussian Mixture Models (BGMM) for two different toy datasets obtained with the same tol, reg, n_init, max_iter and G_{max} values. It is worth remembering that, for the GMM, $G_{max} = G$, that is, the number of Gaussians to be retrieved is imposed, whereas for the BGMM, $G_{max} \leq G$. The dataset on the top row is constituted by a total of 10 Gaussians and $G_{max} = 15$ has been selected. As expected by its definition, 15 Gaussians are used to fit the GMM while, for the BGMM, the correct number of Gaussians is obtained. On the other hand, for the bottom row, a toy dataset constituted by 5 Gaussians is presented and $G_{max} = 8$ has been used, with $G = 8$ being the prediction for the GMM and $G = 5$ the one for the BGMM. In both cases, the actual number of Gaussians that generated the toy dataset is obtained for the BGMM. A visual inspection of the GMM results reveals that this method performs rather poorly, since some of the independent Gaussians (e.g, upper right and upper right Gaussians in the top row dataset and the bottom three Gaussians in the bottom row datasets) contain several independent clusters according to the model outputs.

C. Distribution of star particles for 4 intermediate-mass progenitors in the $E_{tot}\,vs\,L_z$ space in different radius and metallicity bins

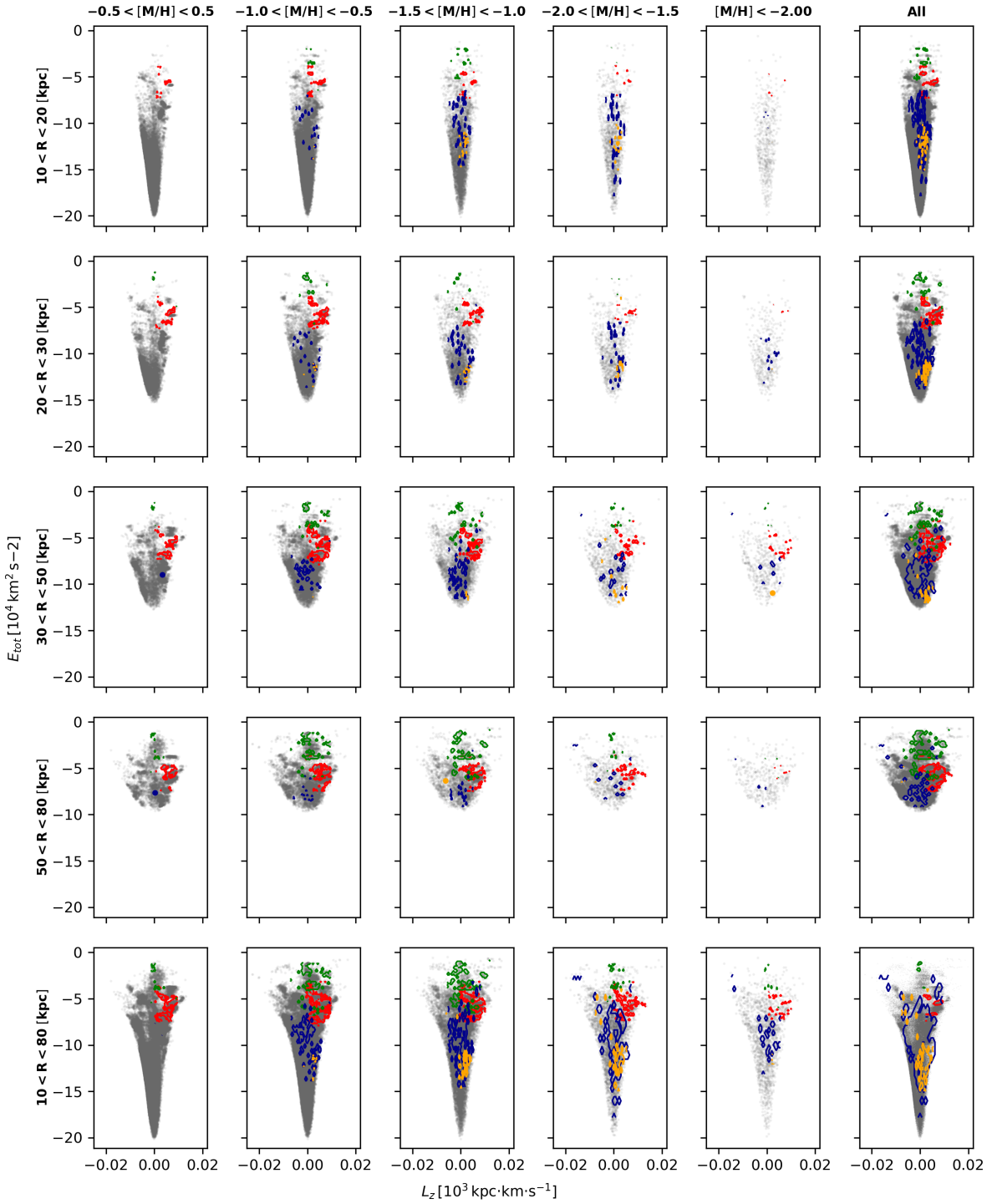

Fig. C.1: Equivalent plot to Fig. 8 for 4 intermediate-mass progenitors. The most massive among those 4 progenitors appears in red, the second one in green, the third one in blue and the fourth one in orange.

Fig. C.1 shows that, for these 4 intermediate-mass progenitors, their contribution to the dataset is much smaller than for the 4 most massive ones. The clustering of the star points that belonged to the same disrupted galaxy is noticeable and, at the same time, though they also give place to several overdensities, the overlap is less severe than for the most massive satellite galaxies. Moreover, there are very few star points in the $-0{,}5 < [M/H] < 0{,}5$ range, which is consistent with the mean values found in Table 1, since mean $[M/H]$ values decrease with the mass of the progenitors.

D. Selection of the quantities to apply the BGMM to

The distribution of the star particles of the 4 most massive progenitors in the different quantities that have been considered to find clumps and in the $50 < R < 80$ kpc galactocentric radius range is shown in Fig. D.1. Even though there is an obvious overlap between the different progenitors, more importantly for the two most massive ones (red and green), we would expect the other two to be reproducible by the BGMM output, especially when considering the $E_{tot} - L_z - L_\perp$ space.

It must be taken into account that, when applying the BGMM model, all features are considered equally relevant, so the selection of the quantities that are used in order to find the clusters is essential. This has been done both in order to select the quantities in which clumps are expected that will be used for the model and to make a first evaluation of the performance of this model. The reason for selecting this specific radius range is that, as it can be seen on Fig. 8, this is the bin where these progenitors are more easily distinguished.

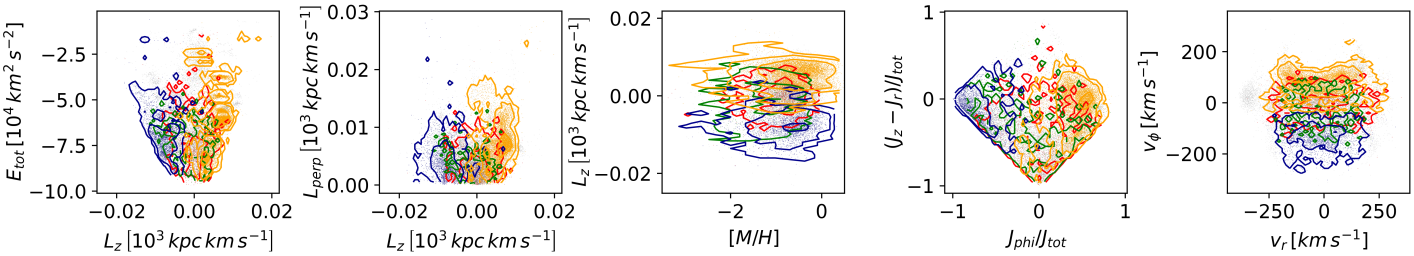

Fig. D.1: Contour plots of the different spaces that are considered to apply the BGMM to for the 4 most massive progenitors. The $50 < R < 80$ kpc galactrocentric radius range has been considered.

The BGMM is applied to 4 different combinations of quantities in order to compare the results of the different models:

- L_z, $L_\perp$, E_{tot}.
- L_z, $L_\perp$, E_{tot}, $[M/H]$.
- L_z, $L_\perp$, E_{tot}, v_r, v_ϕ.
- L_z, $L_\perp$, E_{tot}, J_ϕ/J_{tot}, $(J_z - J_r)/J_{tot}$.

When applying the BGMM, the data is normalized because of the fact that the different quantities have different scales and scaling simplifies handling the Gaussian parameters. Furthermore, for computational efficiency reasons, a subdataset of 40 % of the total star particles has been used to find the clusters by means of the BGMM.

Fig. D.2 shows a $E_{tot}\,vs\,L_z$ density 2D histogram both for the original dataset and for the sample generated by the BGMM in the different spaces and with the same model input parameters. Moreover, the number of Gaussians predicted by the model, as well as their total log-likelihood (lnL) value, are presented. The reason why the $E_{tot}\,vs\,L_z$ is used to compare them is that we expect to find more discernible clumps and because those are the two parameters that are the true integrals of motion for an axi-symmetric potential.

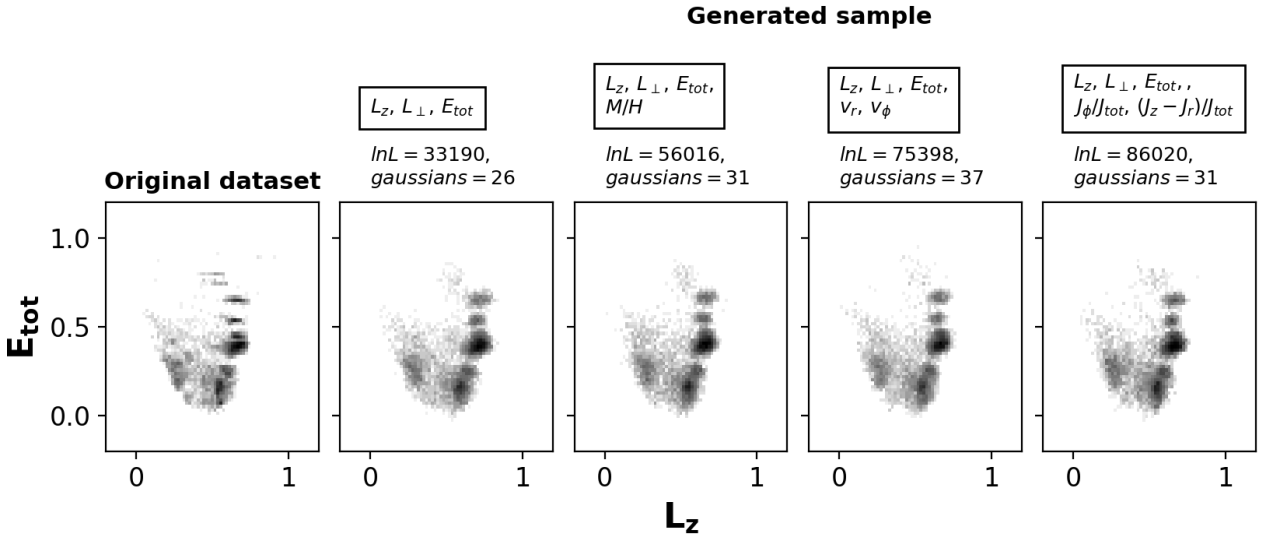

Fig. D.2: 2D density histogram of the original dataset in the $E_{tot}\, vs\, L_z$ space in the $50 < R < 80$ kpc range and of the samples generated by applying the BGMM to different combinations of quantities with the same upper limit on the number of Gaussians, tolerance, regularization value, maximum number of iterations and number of K-means initializations. The number of Gaussians and the total log-likelihood predicted by each model are also included.

Even though the largest log-likelihood corresponds to the BGMM obtained by using L_z, $L_\perp$, E_{tot}, J_ϕ/J_{tot}, $(J_z - J_r)/J_{tot}$, we have previously realised that the clustering is more evident in the $v_\phi\, vs\, v_r$ space than in the action space. In addition, some quantities are redundant, not independent from each other (i.e; J_ϕ and L_z), so the BGMM is not suitable for using these quantities. On the other hand, the generated sample in all cases is very similar and in neither of the cases is able to reproduce the small overdensities in the upper right part of the figures, which indicates that adding information does not improve the models. Accordingly, the addition of the $[M/H]$ value, along with the velocities and actions spaces, does not contribute with useful additional information, since it does not make a meaningful difference despite reducing overlap.

Furthermore, these results show that the number of Gaussians predicted with each one of the different combinations of parameters, as expected, is way larger than the desirable output, which would be one Gaussian per progenitor. And, while a smaller upper limit on the number of clusters obviously leads to a smaller number of predicted Gaussians and convergence might be reached, a visual inspection of the results shows that the BGMM is not able to retrieve the star particles that belonged to the different progenitors as independent clusters. At the same time, this output is not surprising due to the indisputable presence of non-Gaussian shapes in the space distribution.

This may indicate that the BGMM is not enough to reproduce the accretion events, since it only recovers the different overdensities, and a way to associate the different Gaussians that are predicted with the original satellites must be developed. In addition, errors are made because of the presence of non-Gaussian shapes.

Therefore, since the results are not significantly improved when considering more than the E_{tot}, L_z and $L_\perp$, we will stick in our search for substructures with those values both for simplicity, since these spaces are easier to interpret, and in order to improve the computational efficiency.