Stellar Streams in FIREbox

Benjamin Hanf

Dr. Francisco Mercado and Professor Jorge Moreno, Advisors

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts
in Physics and Astronomy

POMONA COLLEGE
Claremont, CA
May 6, 2025

Abstract

A stellar stream is a group of stars orbiting around a galaxy, formed when a disrupted satellite galaxy is accreted onto a larger host galaxy. Stellar streams encode the history of galactic mergers and of their progenitors. Now is a particularly exciting time to study galactic substructure, as new observations from the Vera Rubin Observatory, Euclid, ARRAKIHS, and the Nancy Grace Roman Space Telescope will reveal stellar streams in other galaxies. Finding stellar streams generally requires kinematic and metallicity information. One particularly useful kinematic transformation yields action-angle coordinates, which confine similar orbits to similar regions of action space. In this phase space, clustering algorithms can extract overdensities to identify stellar stream candidates. I have developed code which combines simulation data with machine learning techniques to identify and analyze stellar streams around Milky Way analogs in FIREbox. FIREbox is a novel, state-of-the-art cosmological volume simulation which contains a statistical sample of interacting galaxies. I provide tools for future work connecting theory to observations, in preparation for collaborations investigating galaxy formation and theories of dark matter.

Acknowledgments

To complete this thesis I relied on the help of many generous people. I am immensely grateful to Dr. Francisco Mercado who took me under his wing and set me toward this fascinating endeavor. I would also like to thank Profe Moreno, who shared with me invaluable information about graduate school and the rest of the Universe. I am thankful for Prof. Tamayo at Harvey Mudd for immersing me in computational astrophysics and the world of academia. I am grateful for my friends and community who have created such a wonderful place here, and for my parents whose love and support made me who I am today. I am grateful most of all to my brother whose kindness and unending curiosity inspire me wherever I go. Thank you.

Contents

Chapter 1

Introduction

1.1 Stellar Streams and Galactic Substructure

Now is a particularly exciting time to study galactic substructure. Substructures in galaxies include prominent features, such as the Milky Way's stellar disk, as well as more subtle ones, like faint streams of stars in a galaxy's halo [Helmi, 2020]. In the next decade, the Vera Rubin Observatory[1] and space telescopes like Euclid[2], ARRAKIHS[3] and the Nancy Grace Roman Space Telescope[4] will reveal thousands of these stellar streams in galaxies hundreds of Megaparsecs away [Pearson et al., 2022b]. This rapid development in extragalactic astronomy makes it enticing for theorists to try to model the dynamics and properties of these substructures, which will in turn allow us to answer questions about the history of the Milky Way and its neighbors, as well as the properties of dark matter.

Our Milky Way is a typical disk galaxy embedded in an invisible dark matter halo which comprises a majority of the galaxy's mass [Bland-Hawthorn and Gerhard, 2016]. This halo's shape, density profile, and degree of homogeneity (lumpiness) are still the subject of continuous investigation. We expect that each galaxy in the Universe is similarly embedded within a dark matter halo [Helmi, 2020] as in Figure 1.1 (but see Moreno et al. [2022] for exceptions). The properties of dark matter halos for other galaxies are even more poorly constrained than for our own. The cosmological constant plus cold dark matter (ΛCDM) model is the most popular framework for explaining galaxy formation [White and Rees, 1978, Frenk and White, 2012]. It prescribes that galaxies form hierarchically from mergers of dark matter halos, which are themselves just local overdensities of cold collisionless dark matter. As dark matter is predicted to dominate the mass of baryonic matter by a factor of six [Helmi, 2020], much of the structure and dynamics of the galaxies depend on their dark matter halos.

In the ΛCDM model, galaxies are built up from a series of mergers of dark matter halos.

[1] https://rubinobservatory.org

[2] https://www.esa.int/Science_Exploration/Space_Science/Euclid

[3] https://www.arrakihs-mission.eu

[4] https://science.nasa.gov/mission/roman-space-telescope/

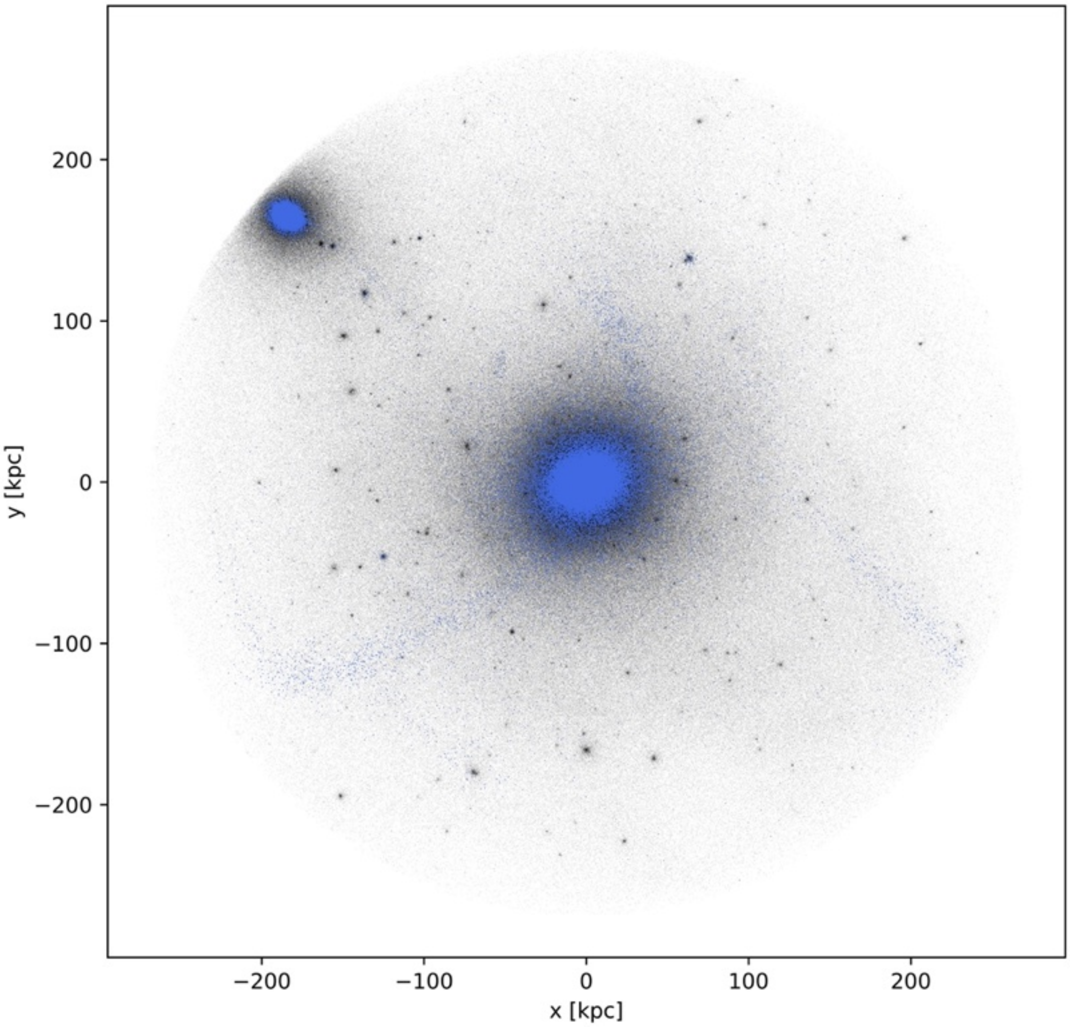

Figure 1.1: Dark matter (black) and star (blue) particles in a simulated Milky-Way-mass FIREbox galaxy [Feldmann et al., 2023]. The black clumps are invisible dark matter overdensities called subhalos. The blue cluster in the top left is a satellite galaxy orbiting around the potential of the host.

This means that smaller halos and their galaxies are absorbed into larger "host" galaxies, disrupting their existing structure. Depending on the relative masses and the orbit of the progenitor, these interactions can form structures within galaxies, including stellar streams [Bullock and Johnston, 2005]. **A stellar stream is a group of stars orbiting around a**

galaxy, often formed from a disrupted satellite galaxy which has been accreted onto a larger host galaxy. Thus, the history of galactic mergers can be studied through the stellar structures within them [Helmi, 2020]. This is the central idea behind galactic archaeology, that stars contain a chemical and kinematic memory of their origin, which can form patterns that make it possible to reconstruct the past.

The disruption of globular clusters can also form stellar streams, although these globular clusters create thinner and less populated streams due to their shallow potential [Bonaca and Price-Whelan, 2025]. However, my work makes use of the FIREbox cosmological simulation which has a baryonic mass resolution on the order of $10^5 M_\odot$, so globular clusters are not well-resolved [Feldmann et al., 2023]. The focus of this thesis therefore focuses on the (wide) stellar streams resulting from dwarf galaxy dissolution.

When a galaxy is disrupted by tidal (gravitational) forces, the stars within it continue to follow a similar path to their original system, as shown in Figure 1.2. Their orbits can be characterized by integrals of motion such as their total energy, angular momentum, or by a transformed action (see Section 2.1.3). Since these quantities will be conserved over time, we expect objects originating from a single accreted satellite to stay clustered in these kinematic spaces.

Streams also place constraints on their galaxy's dark matter distribution, testing predictions of the subhalo population [Menker and Benson, 2024]. When satellites are tidally disrupted by the host, they spread into long thin streams. Due to their low velocity dispersion (further discussed in Section 2.3), these streams are susceptible to gravitational disruption from the dark matter subhalos throughout the galaxy (demonstrated in Figure 1.3). When subhalos collide with streams they leave a visible gap whose shape and length depends on properties of the collision, such as the mass of the subhalo [Sanders et al., 2016]. Observations of the frequency and size of stream gaps reveal information about the masses and orbits of dark matter subhalos, which could lead to constraints on the properties of dark matter and the limits of galaxy formation.

1.2 Stream Identification

Amina Helmi, a professor at the Kapteyn Astronomical Institute[5], discovered some of the earliest evidence for stellar streams in 1999 using kinematic and chemical observations from Hipparcos[6] [Helmi, 2020]. Since then, rapidly improving observations and new methods of detection have revealed more than 100 stellar streams in the Milky Way, sparking new insights into our galaxy's merger history [Bonaca and Price-Whelan, 2025] (see Figure 1.4). Among the most well-known Milky Way streams are the Sagittarius Stream, the Virgo Stream, and the Gaia-Enceladus Sausage, each of which is associated with a merger between the Milky Way and a disrupted satellite galaxy [Helmi, 2020]. The European Space Organization's Gaia space telescope, which launched in 2013 [Gaia Collaboration, 2018], led to an order-

[5] https://www.rug.nl/research/kapteyn/

[6] https://www.esa.int/Science_Exploration/Space_Science/Hipparcos_overview

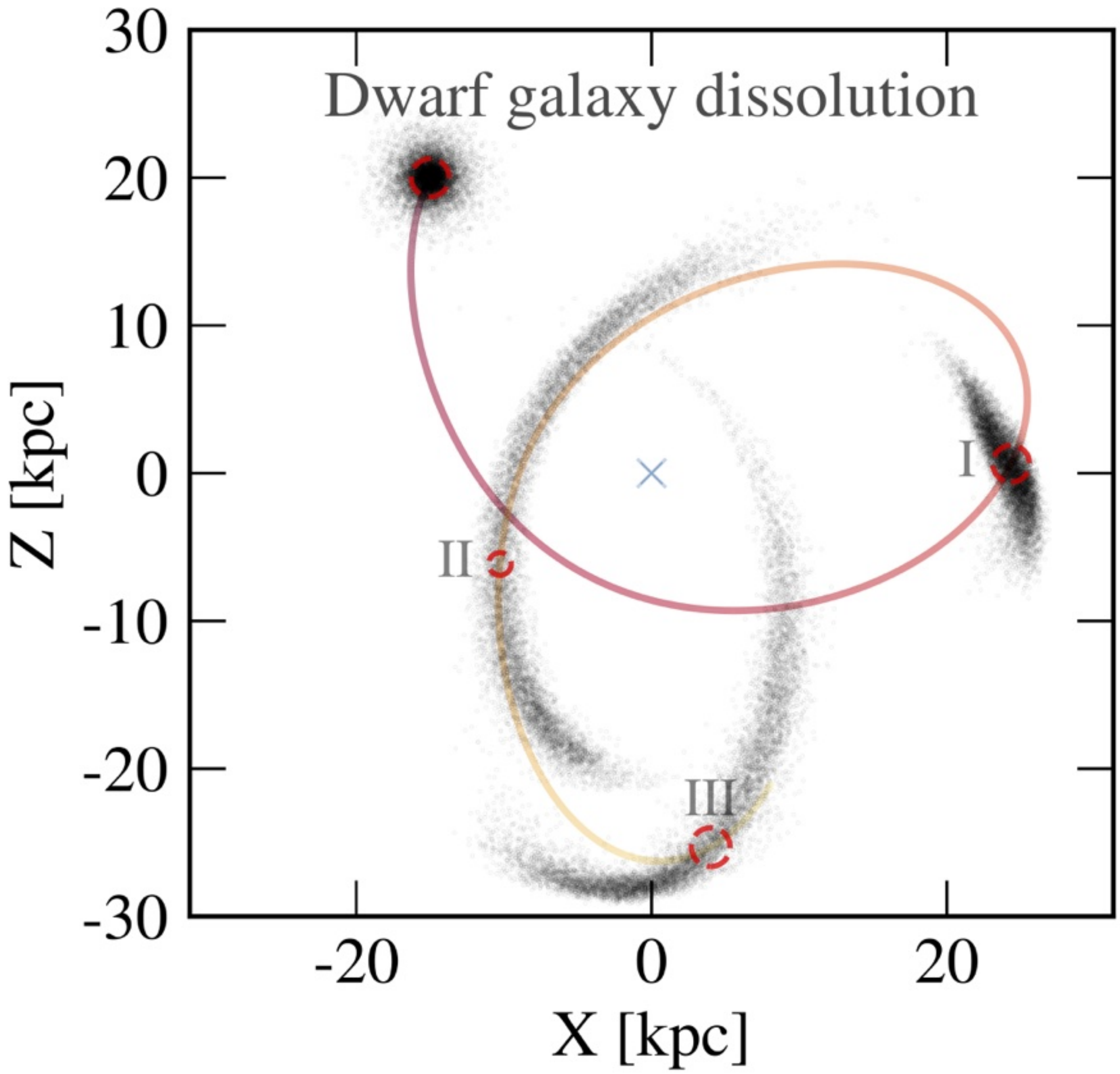

Figure 1.2: First 1 billion years of simulated stream formation due to the accretion of a $10^8 M_\odot$ satellite galaxy in a Milky Way potential. The Galactic center is marked by a blue cross. The first label (I) indicates the first pericenter pass after 300 Myr. During the beginning of stream formation, the satellite galaxy may develop short tails leading and trailing behind the orbit. After 300 Myr (II) the tidal tail is longer, and more aligned with the orbit of the progenitor. After 900 Myr (III) long tails have formed and become misaligned with the orbit of the progenitor. Adapted from Bonaca and Price-Whelan [2025].

of-magnitude increase in known stellar streams in the Milky Way due to its capability of measuring stellar positions and velocities for more than a billion stars [Bonaca and Price-Whelan, 2025].

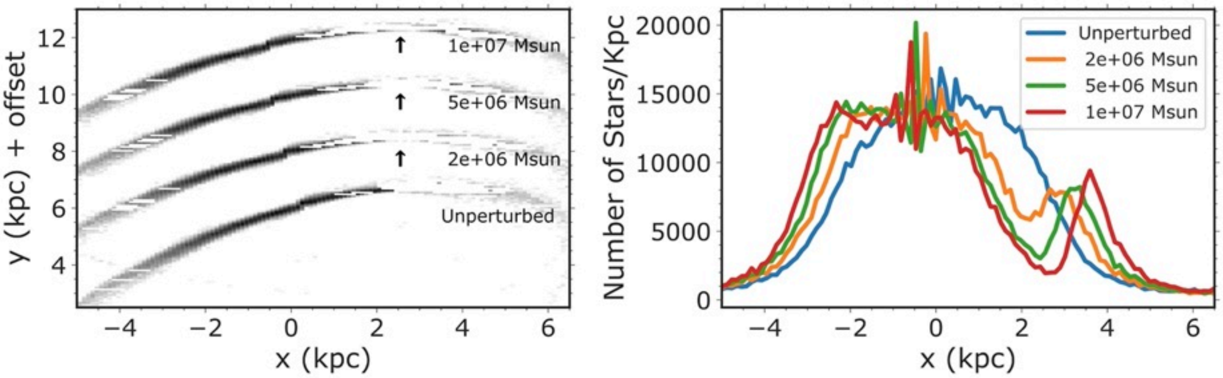

Figure 1.3: **Left:** results of numerical simulation showing the stream gaps resulting from a $5 \times 10^4 M_\odot$ stream colliding with a dark matter subhalo with mass varied between 2×10^6 and $10^7 M_\odot$. The streams are offset in the y direction. The lowest stream is unperturbed for comparison.The gap's size increases with the subhalo's mass. **Right:** linear density of each stream along the x-direction. Perturbations cause predictable changes in the stream's density distribution. Adapted from Aganze et al. [2024].

Still, identifying stellar streams is an observational feat. Astronomers locate Milky Way streams using kinematic data in combination with metallicity and other processing techniques. Stream finding relies on filtering stream stars out of the noisy foreground Milky Way or selecting datasets with a higher representation of stream members. One such example of a selection is by metallicity.

Metallicity refers to the concentration of heavy metals in a star or stellar population, formed from generations of stellar lifecycles. Streams with dwarf galaxy progenitors have lower metallicities than the host on average, due to their shallower potential. Filtering for blue horizontal branch stars, which are more common in old, metal-poor populations, will enhance the signals associated with a stellar stream while suppressing noise from the host galaxy [Bonaca and Price-Whelan, 2025]. For example, Price-Whelan and Bonaca [2018] apply a photometric filter on a field of stars to reveal the GD-1 stream from an otherwise noisy field. The metallicity distribution of a group of stars can be used to locate streams or to confirm the membership of individual stars after a stream has been identified [Sanderson et al., 2017]. This distribution can also differentiate progenitors: streams originating from globular clusters will have a narrower range of velocities and metallicities than their dwarf galaxy counterparts [Helmi, 2020].

However, high quality 6D velocity observations (meaning 3D position and velocity vectors) are not available for galaxies other than our Milky Way. As a result there have been many algorithms developed for finding streams from flat imaging fields. For a review of techniques, see Malhan and Ibata [2018]. Of particular interest are the "The Hough Stream Spotter" [Pearson et al., 2022a], "STREAMFINDER" [Malhan and Ibata, 2018], and "Via Machinae" [Shih et al., 2021]. The Hough Stream Spotter employs a mathematical tool known as the Hough transform [Hough, 1962] to detect linear patterns in noisy data. This

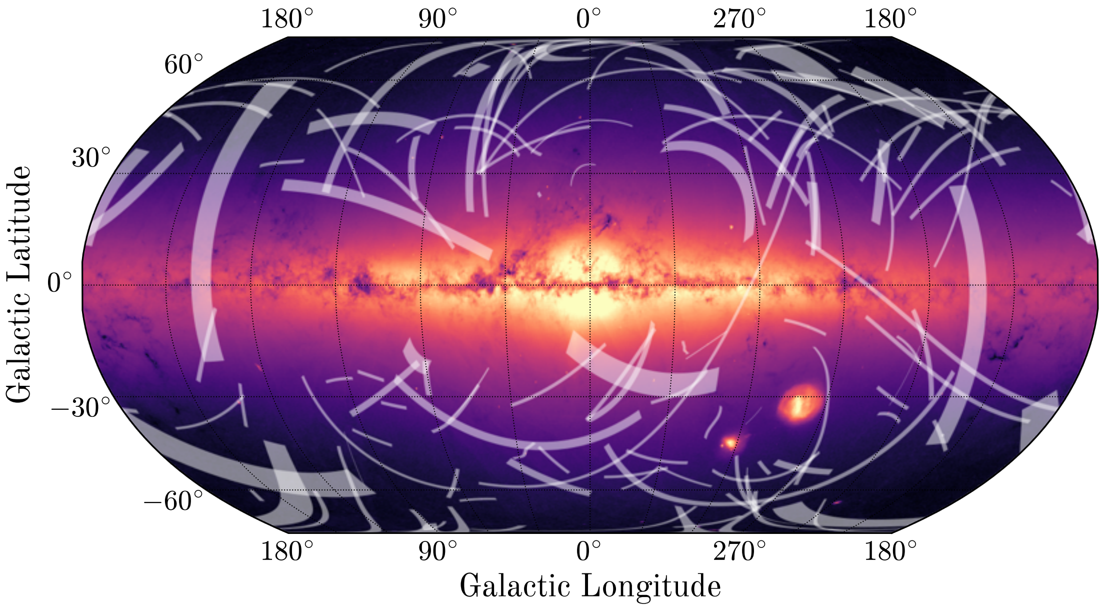

Figure 1.4: Map known stellar streams in the Milky Way. These streams, shown in Galactic coordinates, originate from galaxy mergers and have been discovered by kinematic and chemical evidence in the last few decades. The background represents the stellar number density from Gaia's third data release. This figure was created using the `galstreams` Python package [Mateu, 2023], and adapted from Bonaca and Price-Whelan [2025].

allows the extraction of stream-like structures from only a 2D (image) field of stars. Another algorithm, STREAMFINDER, integrates the orbits of a select number of particles to determine whether they remain in a fixed hypertube in phase space. Via Machinae also leverages the Hough transform, but uses unsupervised machine learning to perform density estimation and subsequently identify stellar streams. These algorithms, in combination with complimentary detection methods, have revealed dozens of stellar streams within the Milky Way [Bonaca and Price-Whelan, 2025].

Using computational stream finding methods, processing can reveal dim and diffuse stellar streams more easily, particularly when combined with improved data (e.g. Gaia). However, more disrupted streams require more in-depth analysis to uncover. A stream may wrap around the host galaxy many times, and this phase mixing makes it difficult to identify its true progenitor [Panithanpaisal et al., 2021]. To track these chaotic disrupted streams, kinematic transformations like action angle variables (see Section 2.1.3) allow researchers to identify groups of stars with similar trajectories. These stellar orbits around the galaxy change much slower than their physical position in space [Shipp et al., 2023]. For example, Borsato et al. [2020] and Wu et al. [2021] use action angles to cluster stars and identify stellar streams using clustering algorithms such as EnLink [Sharma and Johnston, 2009]. Combining all these methods can reveal the scale, composition, and dynamics of clusters which can constrain the galaxy's merger history, as well as dark matter distribution and properties [Walder et al., 2024].

In the past decade, **many authors have analyzed stellar streams in observational data** [Malhan and Ibata, 2018, Pearson et al., 2022a, Shih et al., 2021], **and in high resolution "zoom-in" simulations** [Wu et al., 2021, Panithanpaisal et al., 2021, Borsato et al., 2020]. **In contrast, my work investigates stellar streams in the FIREbox cosmological simulation**, which contains a statistical population of interacting galaxies [Feldmann et al., 2023]. **This approach enables the identification of more streams than a single high-resolution simulation, while drawing from realistic galaxy formation processes to improve predictions of stream properties in anticipation of upcoming observational data**.

Chapter 2

Mechanics of Stellar Streams

2.1 The Orbits of Stars

2.1.1 Orbits in a Spherical Potential

To understand how the orbits of many stars evolve over time, we begin by understanding how a star moves in a spherically symmetric potential. This approach is clearly limited, since real galaxies are neither spherical, nor static. The Milky Way, for example, is currently growing through mergers and accretion [Santistevan et al., 2023]. Still, considering the dynamics for a simple potential will illustrate useful concepts to analyze a more complex potential.

Although a star feels gravitational forces simultaneously from every individual object in the galaxy (and from other galaxies as well), the low collisional cross section of each star motivates a different way of thinking about gravity. Instead of trying to calculate the exact gravitational potential as a sum of potentials created by each object, we will treat the entire galaxy as having a potential coming only from the large-scale mass distribution. Since this is made up of billions of stars, the distribution will be smooth on large scales, and we can ignore irregularities or close-encounters between stars. Since the objects move in a gravitational field, the mass of our particle is also irrelevant. All quantities will be calculated with unit mass, which we can then adjust accordingly.

The Lagrangian $\mathcal{L}$ for a particle in a spherically symmetric potential is

$$\mathcal{L} = \frac{1}{2}\left[\dot{r}^2 + (r\dot{\psi})^2\right] - \Phi(r), \tag{2.1}$$

where Φ is the spherically symmetric gravitational potential and ψ is the azimuthal angle whose direction is defined by the orbital plane. The corresponding (Euler-Lagrange) equations of motion are

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{r}} - \frac{\partial \mathcal{L}}{\partial r} = \ddot{r} - r\dot{\psi}^2 + \frac{\mathrm{d}\Phi}{\mathrm{d}r} = 0, \tag{2.2}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\psi}} - \frac{\partial \mathcal{L}}{\partial \psi} = \frac{\mathrm{d}}{\mathrm{d}t}\left(r^2\dot{\psi}\right) = 0. \tag{2.3}$$

The second of these equations (2.3) defines a conserved quantity

$$r^2\dot{\psi} = \text{constant} \equiv \mathbf{L}, \tag{2.4}$$

which is the angular momentum vector **L**. The conservation of the angular momentum vector demands that motion is restricted to an orbital plane for a spherically symmetric potential. One special case of a spherical potential is a Kepler potential, like that of the solar system. Since the solar system's gravitational potential is dominated by the sun acting as a point mass in the center, planets in the solar system remain in relatively fixed orbital planes. This conservation is deeply related to the rotational symmetry of the system, which according to Emmy Noether's theorem maps onto a corresponding conserved quantity Helliwell and Sahakian [2020]. These conserved quantities greatly simplify our analysis, and allows us constrain the motion of a star in a potential.

2.1.2 Conserved Quantities

One prediction of Hamiltonian dynamics states that, if the Lagrangian $\mathcal{L}$ of a system does not depend explicitly on time, then the Hamiltonian is conserved. Another prediction is that if we can construct a coordinate transformation which does not depend explicitly on time, then the Hamiltonian is equivalent to the mechanical energy ($H = T + U$). The Lagrangian for our system satisfies both properties so that the energy $H = T+U = \text{constant}$ is conserved for any given initial conditions. Since the momenta are given by

$$\begin{aligned} p_r &= \frac{\partial \mathcal{L}}{\partial \dot{r}} = \dot{r} \text{ and} \\ p_\psi &= \frac{\partial \mathcal{L}}{\partial \dot{\psi}} = r^2\dot{\psi}, \end{aligned} \tag{2.5}$$

we can calculate the Hamiltonian H per unit mass as

$$\begin{aligned} H\left(r, p_r, p_\psi\right) &= p_r\dot{r} + p_\psi\dot{\psi} - \mathcal{L} \\ &= \frac{1}{2}\left(p_r^2 + \frac{p_\psi^2}{r^2}\right) + \Phi(r) \\ &= \frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + \frac{1}{2}\left(r\frac{\mathrm{d}\psi}{\mathrm{d}t}\right)^2 + \Phi(r). \end{aligned} \tag{2.6}$$

From these conserved quantities $H = E$ and **L** we can combine with Equation 2.4 to solve for the motion as a function of time for any bound orbit:

$$\left(\frac{dr}{dt}\right)^2 = 2[E - \Phi(r)] - \frac{L^2}{r^2}. \tag{2.7}$$

For a detailed derivation see Binney and Tremaine [2008].

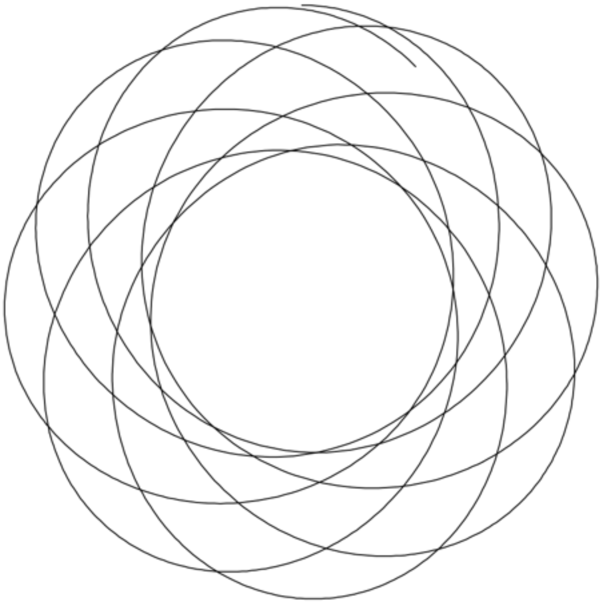

Figure 2.1: An orbit in a spherical potential sweeps out a rosette over time. Energy and angular momentum are conserved, but the potential does not guarantee that the orbits are closed. Instead, they precess about the center. A closed Kepler orbit is a special case of the spherical potential, for $\Phi(r) = -\alpha/r$. Reproduced from Binney and Tremaine [2008].

The orbit oscillates between an inner radius, the pericenter, and an outer radius, called the apocenter. For a circular orbit, the pericenter and apocenter are equal. When the apocenter is nearly equal to the pericenter, the orbit has small eccentricity. If the apocenter is much larger than the pericenter, the eccentricity approaches 1. From equation (2.7) we can see that orbits resemble rosettes as in Figure 2.1, where the radius oscillates between some minimum and maximum radius and precesses around one focus of the ellipse.

2.1.3 Integrals of Motion

Any orbit traces a path in a six dimensional space with coordinates of position and velocity $(\mathbf{x}, \mathbf{v})$. This is known as phase space. A **constant of motion** is any function $C(\mathbf{x}, \mathbf{v}, t)$ which does not change over the course of an orbit. In a $2n$-dimensional phase space, there are always $2n$ constants of motion. For example, in our 6D phase space the initial coordinates $(\mathbf{x}_0, \mathbf{v}_0)$ can be regarded as the six constants of motion. If the position and velocity are given by x(t) and v(t) = dx/dt, then

$$C[\mathbf{x}(t_1), \mathbf{v}(t_1); t_1] = C[\mathbf{x}(t_2), \mathbf{v}(t_2); t_2]. \tag{2.8}$$

An **integral of motion** $I(\mathbf{x}, \mathbf{v})$ is any function of the phase-space coordinates alone that is constant along an orbit:

$$I[\mathbf{x}(t_1), \mathbf{v}(t_1)] = I[\mathbf{x}(t_2), \mathbf{v}(t_2)]. \tag{2.9}$$

While every integral of motion is a constant of motion, the reverse is not necessarily true. An integral of motion may not depend explicitly on time. Orbits can have anywhere from zero to five integrals of motion. Some of these integrals can be written down easily: in any static potential $\Phi(x)$, the Hamiltonian $H(\mathbf{x}, \mathbf{v}) = \frac{1}{2}v^2 + \Phi(x)$ is an integral of motion. If a potential $\Phi(R, z, t)$ is axisymmetric about the z axis, the z-component of the angular momentum L_z is also an integral of motion. However, even when these integrals exist they are not guaranteed to have analytical solutions. These invariants are nonetheless useful in analyzing the geometry of orbits, and particularly in reducing the dimension of the orbit in phase-space.

We can think about this in the case of a spherically symmetric potential: In this case, the phase space has six dimensions. However, **every isolating integral of motion constrains the orbit to a $2n - 1 = 5$ degree hypersurface in phase space**. The integral $H(\mathbf{x}, \mathbf{v}) =$ constant confines the orbit to a five-dimensional subspace. Conservation of angular momentum $\mathbf{L}(\mathbf{x}, \mathbf{v}) =$ constant adds three further constraints, restricting the orbit to a two-dimensional surface. Through the equation $\Phi_0(\mathbf{x}, \mathbf{v}) =$ constant this fifth integral confines the orbit to a one-dimensional curve on this surface. Figure 2.1 is a projection of this curve.

2.2 Action Angle Coordinates

In Section 2.1, we found that spherical potentials admitted at least four integrals of motion (I): the Hamiltonian (energy) and the three components of angular momentum $\mathbf{L}$. Next we will focus on a set of coordinates called **action-angle variables**, in which the three canonical momenta ($\mathbf{p_q}$) are integrals of motion (called "actions") and their corresponding coordinates ($\mathbf{q}$) are called "angles". These coordinates are useful precisely because actions are integrals of motion in a given potential, so long as the potential changes slowly with respect to the timescale of orbital evolution.

For stars in a galaxy, their collision cross section is much smaller than their average spacing. This means that coordinates of integrals of motion, which are generally conserved for undisturbed orbits, are useful in identifying stars with shared progenitors.

Action-angle variables arrive through a **canonical transformation** of the Hamiltonian for a given potential, defined such that the new coordinates preserve the form of Hamilton's equations:

$$\begin{aligned} \dot{p} &= -\frac{\partial H}{\partial q}; \\ \dot{q} &= \frac{\partial H}{\partial p}. \end{aligned} \tag{2.10}$$

After the transformation $(p_i, q_i) \rightarrow (J_i, \theta)$ we have two equivalent coordinates which

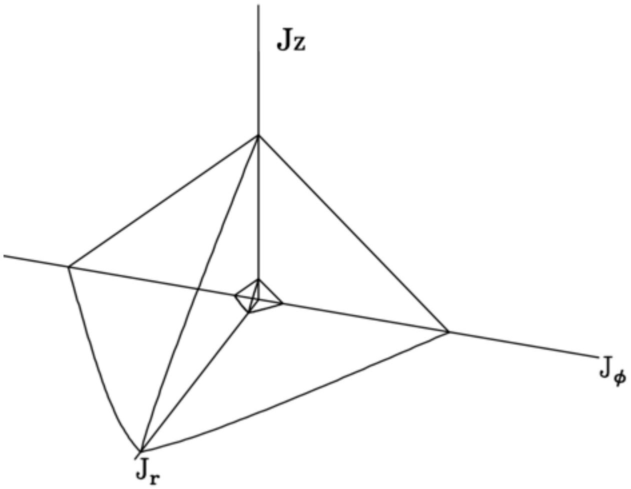

Figure 2.2: The three-dimensional action space for stars in a given potential. Points on the axes represent orbits for which only one action is non-zero. These are closed orbits, in the sense that they do not precess around the potential. The origin represents a stationary orbit at the center of the potential. Orbits of constant energy represent planes in each octant. Reproduced from Binney and Tremaine [2008].

satisfy both Equations 2.10 and the following equation for time evolution:

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \frac{\partial H(\mathbf{J})}{\partial \mathbf{J}} = \text{constant} = \boldsymbol{\Omega}, \\ \dot{\mathbf{J}} &= -\frac{\partial H(\mathbf{J})}{\partial \boldsymbol{\theta}} = 0. \end{aligned} \tag{2.11}$$

A full derivation can be found in Jo Bovy's *Dynamics and Astrophysics of Galaxies*[1] [Bovy, 2026].

2.3 Tidal Streams

Consider a satellite with mass m orbiting around a much larger host galaxy with much larger mass M. The satellite will be stripped of its outermost stars by tidal forces, and those stars will continue to orbit on a similar trajectory, with their orbit dominated by the potential associated with the host. On these orbits, we know from Equation 2.11 that the **actions J are constant** and that θ **increases at constant rate** $\dot{\theta} = \boldsymbol{\Omega}$. These stripped stars will have a distribution of actions and angles $\mathbf{J}_0 \pm \Delta\mathbf{J}, \theta_0 \pm \Delta\theta$. During pericenter (closest approach) the average actions of the stripped stars will be nearly the same as that of the satellite, since the tidal forces are symmetric.

[1] https://galaxiesbook.org

The spread of the actions $\Delta\mathbf{J}$ comes from the stars' initial spatial and velocity dispersion. Since both are related to the radius and mass of the satellite galaxy, the action dispersion will scale with the mass

$$\frac{\Delta\mathbf{J}_i}{\mathbf{J}_i} \propto \left(\frac{m}{M}\right)^{1/3}, \tag{2.12}$$

so that the satellites will have a smaller action dispersion

Galaxies are built up from smaller satellites falling into the center. Since these satellites are less massive, the stars belonging to the satellite occupy a small volume in action space (eq. 2.12). The actions of these infalling stars change very slowly over time, so long as the gravitational potential varies slowly compared to their orbital timescale. This is shown in Binney and Tremaine [2008], owing to the conservation of volume in phase space by Liouville's theorem [Helliwell and Sahakian, 2020]. Another consequence of the time evolution of action angles is the geometry of the resulting structure. In radially symmetric potentials, the stripped stars will spread along the path of their (similar) orbits, creating a filament structure. These are stellar streams, and form during the hierarchical accretion events which build today's galaxies. This process results in a web of stellar streams, the analysis of which has only recently begun. These streams can be studied as remnants of past mergers, as well as probes for dark matter distribution in a galaxy. We will use action space coordinates to study the current orbits of stars from galaxy mergers because the remnants of galactic mergers stay coherent in action space even after they have become mixed beyond recognition based only on their locations.

Chapter 3

Computational Methods

To find stellar streams around simulated galaxies in FIREbox, **I develop a pipeline in Python which assembles computational tools to process data into the final product**. This chapter describes the computational methods used in the process. Section 3.1 describes specifications, features, and limits of FIREbox cosmological simulations with respect to this project. Section 3.2 discusses computational approaches to numerical action finding. Section 3.3 discusses my approach to potential fitting, which is necessary to resolve accurate estimations for the stellar actions. Section 3.4 introduces the concept of clustering algorithms as a method of unsupervised machine learning, and outlines the process of my chosen clustering algorithm, HDBSCAN. Section 3.5 describes a catalog of stellar stream candidates built from simulation data.

3.1 The FIREbox Cosmological Simulation

I analyze stellar streams in a sample of galaxies from FIREbox, a state-of-the-art cosmological volume simulation [Feldmann et al., 2023]. The "Feedback in Realistic Environments" (FIRE) collaboration has successfully studied the formation and evolution of simulated galaxies at a variety of mass resolutions, resulting in an improved understanding of star and galaxy formation [Hopkins, 2015, Hopkins et al., 2022]. FIREbox applies the physically-motivated FIRE-2 physics model [Hopkins et al., 2018] to a large cosmological volume ($\mathrm{V}_{\mathrm{box}} \approx (22.1\ \mathrm{Mpc})^3$ at z = 0), which makes it the best simulation to study the relationships between galactic substructures like stellar streams, and properties of the host galaxy. FIREbox strikes a balance between volume, resolution, and realism. Figure 3.1 demonstrates its ability to resolve structure at cosmological and subgalactic scales. FIREbox has the highest dynamic range among many recent simulations, investigating scales from 20 pc to 20 Mpc (Figure 3.2). Another advantage is the resolution of the multiphase interstellar medium, which impacts star formation and feedback mechanisms.

By searching through a catalog of FIREbox galaxies created by Moreno et al. [2022], we have selected galaxies with masses similar to the Milky Way ($7 \times 10^{11} M_\odot \leq M_{\mathrm{vir}} \leq 3.2 \times 10^{12} M_\odot$). The virial mass ($M_{\mathrm{vir}}$) measures the mass contained within a gravitationally

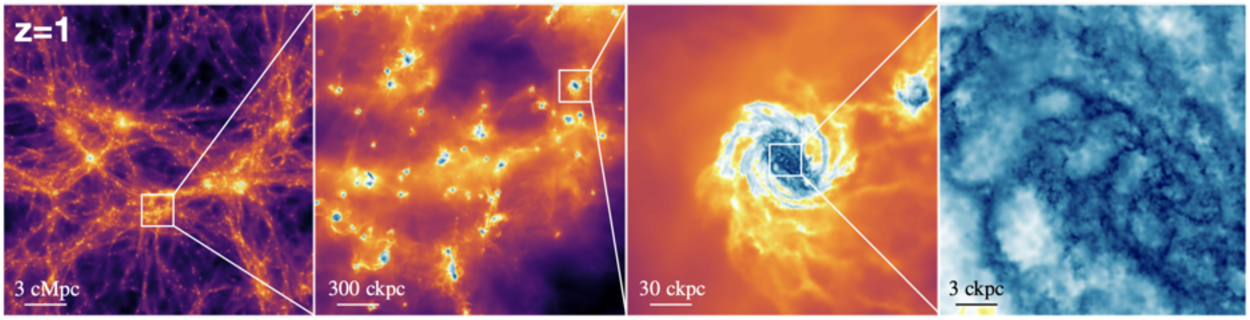

Figure 3.1: Gas distribution in `FIREbox`, over cosmological and subgalactic scales, from Feldmann et al. [2023].

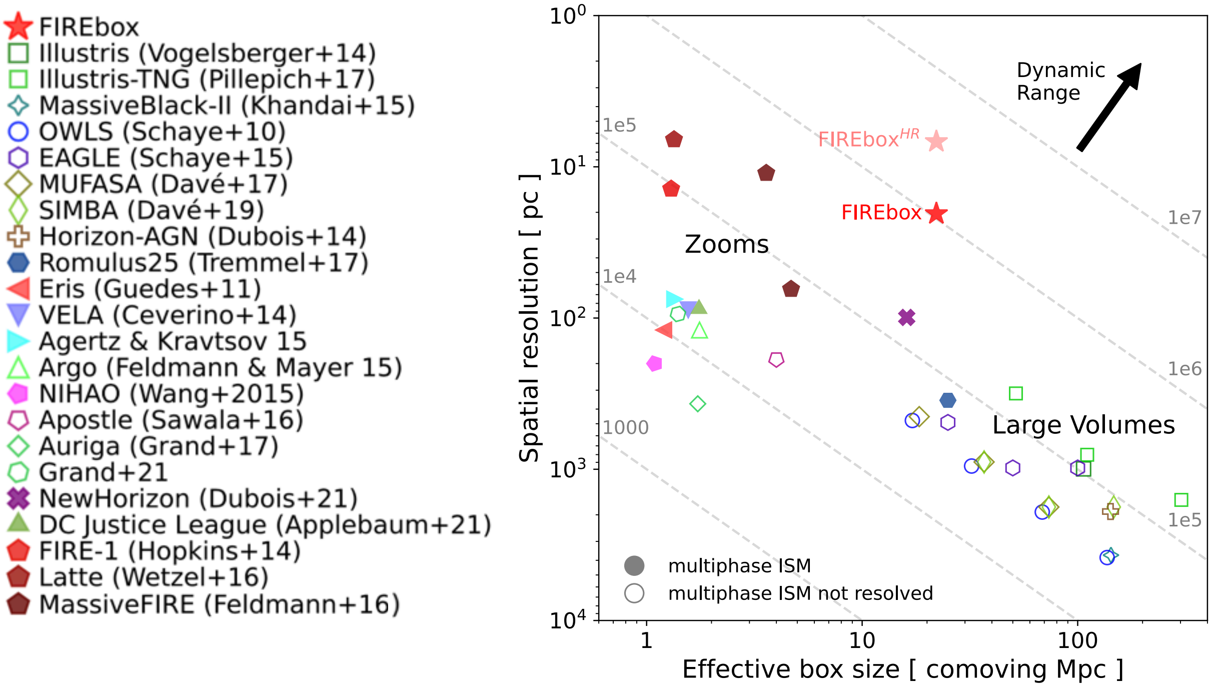

Figure 3.2: The spatial resolution and volume of `FIREbox` (filled red star) vs other simulations. Adapted from Feldmann et al. [2023]. FIREbox occupies a unique position in the context of cosmological computing.

bound system according to the virial theorem, and in our case is calculated based on the overdensity definition from Bryan and Norman [1998]. Choosing this subset ensures that the galaxies are well-resolved in FIREbox, relatively abundant, and comparable to our own. There are 28 such galaxies in FIREbox at $z = 0$.

One caveat of investigating stellar streams in FIREbox is the mass resolution. With a baryonic mass resolution on the order of $10^4 M_{\odot}$, the FIREbox simulation will not resolve the smallest stellar streams. The smallest known stellar streams in the Milky Way are on the order of $10^2 M_\odot$ [Bonaca and Price-Whelan, 2025]. However, the FIREbox simulation resolves hundreds of low-mass galaxies [Feldmann et al., 2023] whose dynamical and chemical properties are preserved upon their accretion onto a more massive host. I aim to identify these structures in my thesis.

3.1.1 FIRE-2 Physics

The FIREbox simulation uses a number of physically motivated processes to accurately and self-consistently simulate processes within the cosmological volume. Specifically, FIREbox uses the "FIRE-2" [Hopkins et al., 2018] implementation of physical processes such as star formation, stellar feedback, and gas dynamics [Feldmann et al., 2023]. The simulation runs in the `GIZMO`[1] code [Hopkins, 2015], which calculates gravitational forces between particles and models hydrodynamic effects. The FIRE-2 simulations do not include active galactic nuclei (AGN) feedback during galaxy formation. Active galactic nuclei are massive blackholes which accrete nearby gas and stellar matter. They also generate radiation pressure and outflows whose imparted momentum can suppress star formation. Although AGN feedback will predominantly impact the center of galaxies, it is possible that its inclusion could impact stream statistics, particularly in the case of suppressing star formation in dwarf galaxies. Although there is evidence for AGNs in dwarf galaxies [Reines, 2022], it is unclear whether they are the dominant form of feedback in this mass range. The connection between AGN feedback and low mass galaxies in FIREbox is left for future work.

3.2 Action Finding

Galaxies are complicated collections of stars and dark matter particles moving in orbits determined by the superposition of potentials created by all other particles. However, simplifications such as the assumption of axisymmetry guarantee that these orbits possess three integrals of motion [Sanders and Binney, 2016]. These integrals of motion allow the calculation of action angles, which are useful in analyzing the trajectories of subgroups of particles, like stellar streams. Despite this, actions cannot be computed analytically for many potentials. Therefore, numerical methods approximate the action-angle coordinates, relying on analytical solutions for similar conditions.

3.2.1 Stäckel Fudge Method

The Stäckel fudge is a numerical method for estimating the actions and angles given positions, velocities, and some separable potential, also called a Stäckel potential $(\mathbf{x}, \mathbf{v}, \Phi(\mathbf{x}))$. Potentials of this form, represented in ellipsoidal coordinates, allow for an explicit solution to the Hamilton-Jacobi equation through separation of variables. The separation process yields two constants of motion in addition to the energy, as well as an expression for momentum as

[1] http://www.tapir.caltech.edu/~phopkins/Site/GIZMO.html

a function of the corresponding coordinate and the constants of motion $p_i(q_i; E, I_2, I_3)$. The actions are given by

$$J_i = \frac{1}{2\pi}\int p_i dq_i \tag{3.1}$$

for each coordinate-momentum pair (p_i, q_i).

The implementation of the Stäckel fudge approximation in the `Galpy` Python library [Bovy, 2015] solves the integral given particle positions and velocities in any axisymmetric potential provided. Assuming the potential is locally separable, the algorithm finds the magnitude of oscillations in spheroidal coordinates, and computes the integrals corresponding to J_r, J_ϕ, and J_z. The error in this approximation is $\leq 10\%$, even for highly eccentric orbits [Vasiliev, 2018]. Moreover, the process is well-optimized in its `C` implementation, and fully parallelizable for scaling to large data sets.

3.3 Potential Fitting

The Stäckel method requires a known potential to estimate actions. As in Panithanpaisal et al. [2021], **I use the `AGAMA` (All-purpose Galaxy Modeling Architecture) python package [Vasiliev, 2018] to fit a physically-motivated axisymmetric compound potential for each Milky Way mass galaxy in FIREbox.**

The galaxies consist of three components: stars, gas, and dark matter. These components have different morphologies. Dark matter is effectively collisionless, while hot gas remains extended and pressure-supported [Stevens et al., 2017], so they both take on a generally spherical distribution. Meanwhile, gas which can radiate its heat away will lose its support and collapse toward the center. Conservation of angular momentum causes the cooled gas to spin up and settle into a thin, rotationally supported disk. The difference in distribution requires two different basis function expansions to approximate the two morphological components. As in Arora et al. [2022], I group the dark matter and hot gas together, and fit using an expansion in spherical harmonics with $\ell \leq 4$. Meanwhile, the stars and cold gas is approximated by an azimuthal harmonic expansion with $m \leq 4$. I chose the gas cut-off temperature to be 1000K based on the gas temperature distribution in the galaxy.

3.4 Clustering Algorithms

One central hypothesis of phase space clustering is that objects with similar origins will reside in similar locations in some phase space. That is, the distance will be lower between members of the group than to non-members. However, there are multiple ways to define "distance" in arbitrary dimensions. For simplicity, we will use Euclidean distance in n-dimensional space given by:

$$d = \sqrt{\sum q_n^2}, \tag{3.2}$$

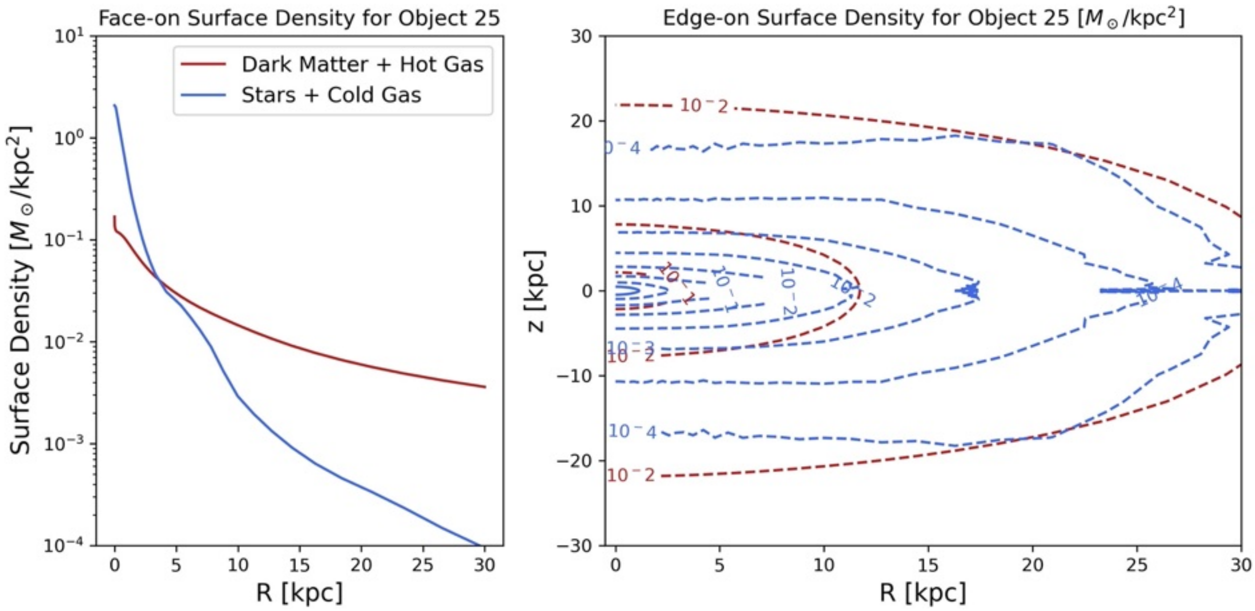

Figure 3.3: Face-on and edge-on surface density for components of the best-fit potential. These potentials may be interpreted as the mass contained within a radius R (on the left). The "cold" component dominates at $R \leq 5$ kpc, but falls off quickly while the diffuse "hot" component maintains a greater density out to hundreds of kpc. Compare to Figure 1.1.

where q_n represents one of the n orthogonal coordinates q which define the space.

Even now that we have identified a distance metric, it is non-trivial to find groups of data in this space. The task, known as clustering, should be accomplished by an algorithm which takes as an input a list of points and their associated coordinates, and returns a cluster map, identifying each point with one of the clusters according to its distance from the cluster. The goal of this thesis is to use a clustering algorithm to identify potential streams around MW-mass galaxies in FIREbox by finding local overdensities of star particles in action space.

3.4.1 HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm which attempts to solve the problem by combining hierarchical and density based clustering. The algorithm was developed for a diverse range of applications with very few assumptions about the dimension, distribution, or classes of data [Campello et al., 2013]. The algorithm assumes that the data contains both noise and clusters. As a demonstration, the following data was generated using `scikit-learn` to contain evident clusters as well as noise (Figure 3.4, left). An algorithm like k-means clustering will perform poorly on this toy data because of the complicated spatial distribution. The clusters are visible by eye because of their density, in contrast to the sparse, noisy background.

The first step in the HDBSCAN algorithm is to build a matrix that determines the dis-

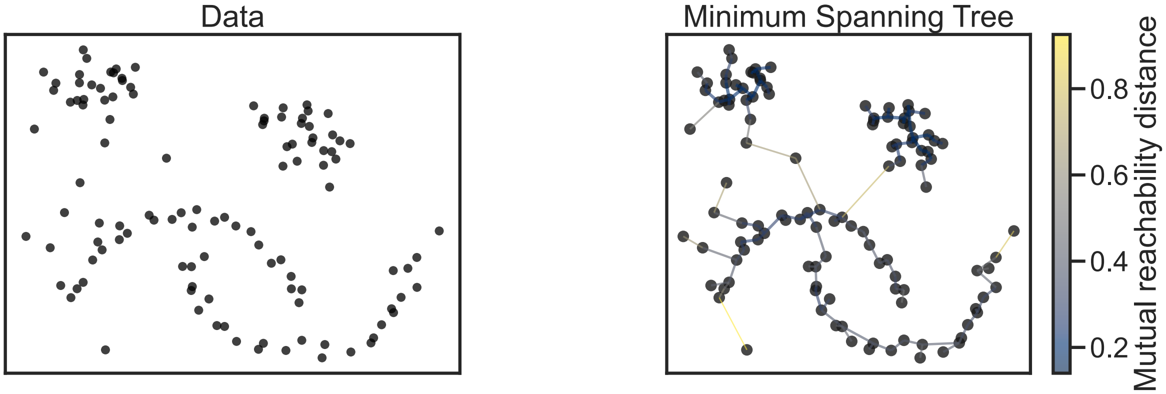

Figure 3.4: **Left:** This toy data was generated to have well-defined centers, with random noise added to demonstrate the clustering algorithm. The data is 2 dimensional, but the algorithm is generalizable to arbitrary dimensions. Axes represent arbitrary orthogonal coordinates. **Right:** The minimum spanning tree of mutual reachability distances constructed by the clustering algorithm.

tance between each point. The HDBSCAN clustering algorithm is based on single linkage clustering. To reduce the impact of noise, the algorithm implements a density-based transform to distinguish between regions of dense and sparse data. This is achieved by finding the kth nearest neighbor using the distance matrix. We define this distance as the **core distance** between a point x and its kth nearest neighbor as $\mathrm{core}_k(x)$. Then, for each point pair we define their **mutual reachability distance** as the maximum of their core distances:

$$d_{\mathrm{mreach}-k}(a,b) = \max\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}. \tag{3.3}$$

This distance metric has the effect of **spreading out points in low density regions** (which have high mutual reachability distances). This benefits the algorithm by spreading low-density noise points further from the high-density clusters. To find the clusters, the algorithm builds a tree. The mutual reachability distance matrix $\boldsymbol{M}_{\mathrm{ut}}$ represents a weighted graph, with vertices weighted corresponding to the mutual reachability distance between the points. Starting with some high threshold value, and lowering it while removing all edges with weight above that threshold will result in islands of connected clusters appearing. Eventually the method will reveal a skeleton of connected data, with some data well connected and some completely isolated at different threshold levels. In reality, this is a computationally expensive task— there are n^2 edges to check for every iteration of the threshold value. Instead, Prim's algorithm [Prim, 1957] provides an efficient method to compute the **minimum spanning tree** corresponding to the graph with the minimum number of edges and total distance such that all points are connected (Figure 3.4, right).

Next, the algorithm can build a **connectivity dendrogram** (Figure 3.5, left) to merge clusters based on their distance to the nearest cluster. This is created by sorting the edges of the tree by their length. First, the closest points (corresponding to the shortest edges) are clustered together. They now represent a proto-cluster. Then the algorithm iterates through again, merging through until the furthest distance has been reached, and all the points have been sorted into a single cluster.

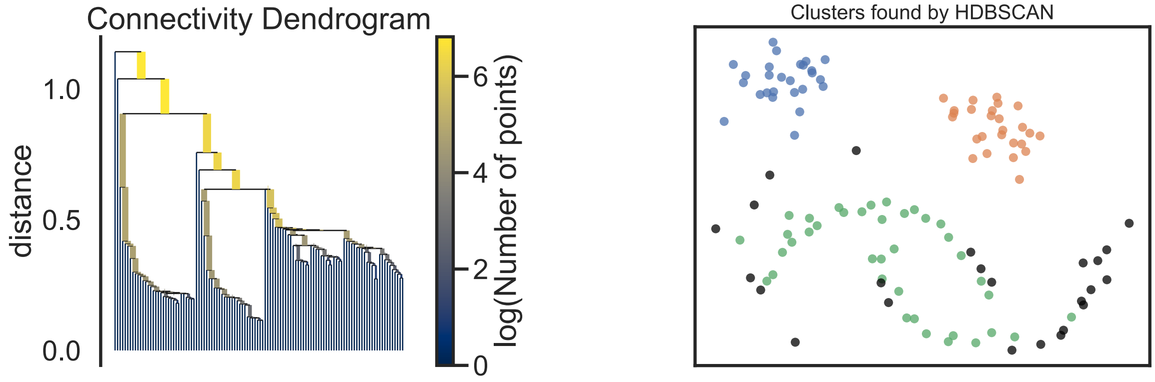

Figure 3.5: **Left:** Connectivity dendrogram for hierarchical clustering with HDBSCAN, showing distance relationships between data points. Logarithmic scale indicates cluster sizes. The vertical axis (distance) measures the distance between clusters and points. The lower a split occurs, the closer the points in the branch are. At the bottom, each point is individual, in its own cluster. Moving upward, clusters merge into larger groups until they form a single structure. **Right:** The clusters are categorized based on hierarchical and density analysis. Points are colored by their identified cluster identity. Points classified by the algorithm as noise are plotted in dark gray.

This is where single linkage clustering (DBSCAN) ends. It is possible to define clusters by drawing a horizontal line through an arbitrary distance on the merger tree, effectively setting a maximum mutual reachability distance. This would also set a limit on the cluster density. HDBSCAN attempts to group points through the concept of persistent clusters. Instead of two clusters merging into one, it is more helpful to think about a single cluster (with some minimum size) fragmenting as the distance threshold decreases. This will result in a simpler condensed graph, which contains useful information about the persistence of the clusters across different thresholds.

The most persistent clusters are most desirable to identify, since they are less likely to be the product of noise. To measure the persistence, we define $\lambda = 1/d_{\text{mreach-k}}$. For each point p in a given cluster, the value λ_p is the lambda value at which the point “fell out of the cluster”. For the cluster, λ_{birth} is the lambda at which the cluster separated from a larger cluster. The **stability** for this cluster is given by

$$\sum_p (\lambda_p - \lambda_{\text{birth}}). \tag{3.4}$$

The algorithm determines clusters by comparing the stability of the children to that of the parent. If the parent has a smaller stability than the sum of the children, the algorithm sets the stability of the parent to that of the sum of of the children. If the parent has a greater stability than the sum of its children, that cluster is selected. Once the root node is reached, the currently selected clusters are chosen, and all other points are classified as noise. This guarantees that the algorithm finds the most stable cluster across different cutting thresholds. The results of HDBSCAN on our toy data are shown in Figure 3.5.

3.4.2 Applying HDBSCAN

To optimize clustering on stellar particle data, I calculate the smallest radius from the galactic center which contains 75% of the stellar mass, and filter out stars within that radius. This filtering reduces the number of star particles, which decreases the computation time drastically due to the poor scaling associated with the HDBSCAN clustering algorithm. The calculated radius, called r_{75}, is generally an order of magnitude less than the virial radius ($r_{75} < R_{\text{vir}}/10$). As such, we do not expect to lose crucial kinematic information about prominent stellar streams by ignoring the dense inner region of the galaxy. Recently accreted satellites may have components crossing the center of the galaxy, but likely have recognizable extended tails at $r \leq r_{75}$.

3.5 Extracting Streams from FIREbox

It is useful to have a metric by which we may optimize clustering and understand the performance of action space clustering. We accomplish this by finding stream candidates from earlier snapshots of the simulation. This stream "ancestor" catalog will be made available to the collaboration for use in future stream finding projects. **We define a stream progenitor as any gravitationally self-bound collection of stars which was a satellite of another host galaxy at a redshift of $\mathbf{z = 0.5}$**. Tracking those satellite particle IDs forward in time, we can easily see the time-evolution of these streams. This catalog will be made available to the FIREbox collaboration for future use. We chose a threshold of $z = 0.5$ based on work by Wu et al. [2021], who predict that earlier-accreted satellites are less likely to be clustered in action space due to the non-adiabatic evolution of the host galaxy's potential. Although we track particles identified at $z = 0.5$, the code will find earlier candidates if directed to any previous snapshot. Figure 4.2 shows a selected object with an active and recent merger history in physical (x, y) space and in action space. Star particles are colored according to the ID of the most recent progenitor. All identified streams were gravitationally self-bound at $z = 0.5$. Figure 4.3 shows the same information for a galaxy with fewer and less massive identified streams. Appendix A contains more visualizations of streams in three dimensions (Figure A.1) and in action space (Figure A.2).

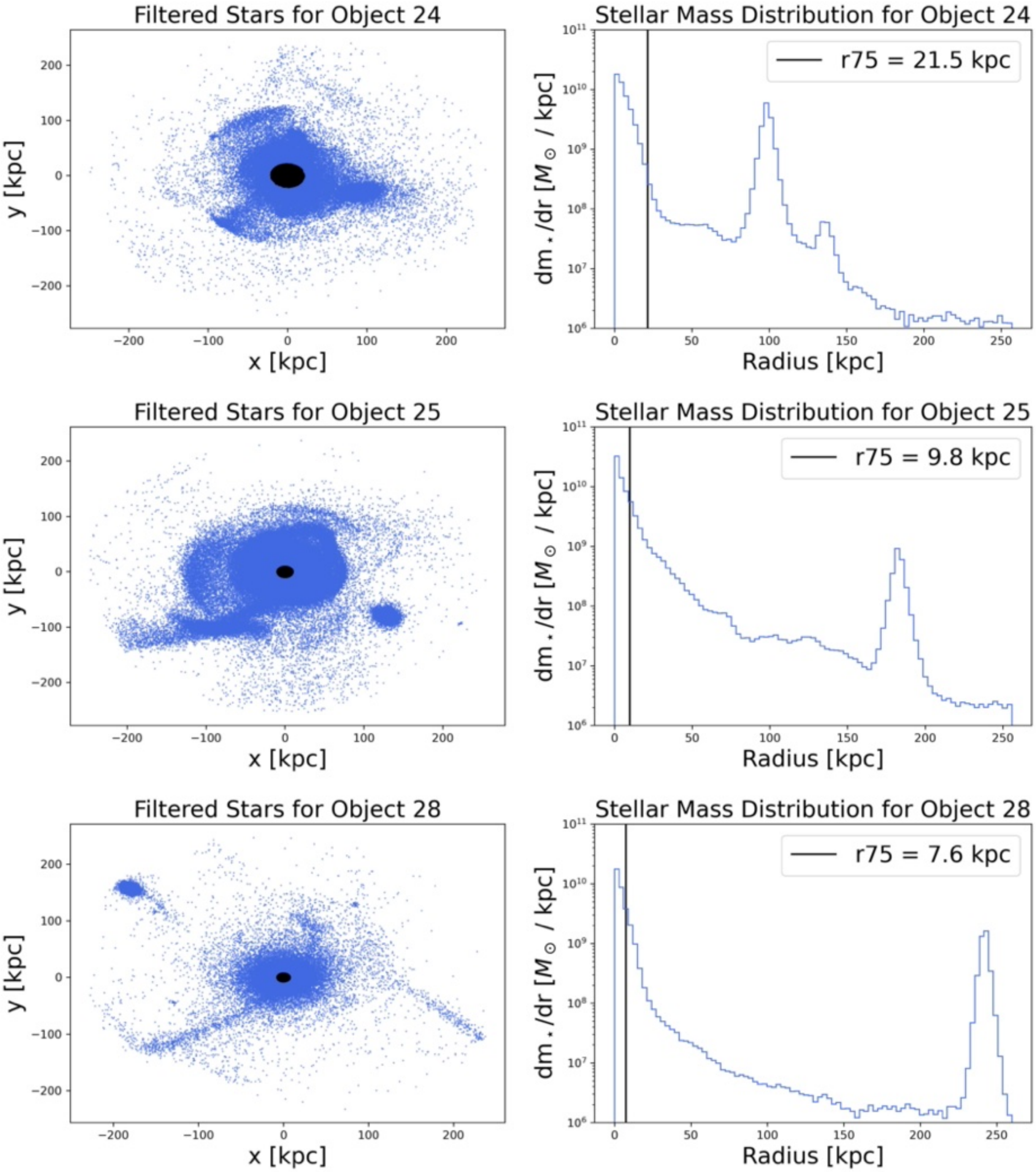

Figure 3.6: Visualizing results of filtering the innermost stellar particles representing 75% of the total stellar mass. On the left, the stellar particles in black have $r \leq r_{75}$ while the particles in blue have $r \geq r_{75}$. Only the blue particles will be used for action finding and clustering analysis. On the right, the stellar mass distribution is compared with r_{75}. The dense central region has many stars, but limited kinematic information. These galaxies are all Milky Way mass, and sampled to demonstrate diverse morphologies.

Chapter 4

Results

4.1 Streamfinding Using Clustering Methods

Clustering methods have been thoroughly investigated for their use in finding streams and other stellar substructures. Wu et al. [2021] use the clustering algorithm Enlink [Sharma and Johnston, 2009] in action space to recover streams in FIRE zoom simulations. They are able to recover only the most recent and prominent streams, which have not been entirely phase mixed or disrupted by subsequent merger events. Myeong et al. [2018] also use Gaia kinematic data to identify Milky Way streams in action space, using metallicity as confirmation. They transform the estimated actions into semilog action space $[\log(J_R), J_z, \log(J_\phi)]$ because of the high variance of action values in these dimensions. Clustering in unprocessed action space results in a bias toward the J_r and J_z axes, as the distance between points is greater in these dimensions.

4.1.1 Coordinate Normalization

In an effort to address the problems associated with clustering in dimensions with different distributions, I employ the `scikit-learn` Standard Scaler, which subtracts the mean and divides by the standard deviation for each data set. The centering and scaling happen for each axis independently, so that units and variance within each axis are comparable to the other axes. This is important because clustering algorithms are sensitive to the distance between points in all dimensions, so the distribution along a particular axis acts as a weighting applied to the data. Normalizing the data this way also allows for arbitrary combination of dimensions, such as actions, position, velocities, energies, and even metallicity.

4.1.2 Parameter Optimization

Optimizing the parameters for a clustering algorithm is a difficult pursuit. Using **HDBSCAN**, there are three parameters to optimize over:

1. `min_cluster_size`

   The minimum cluster size is the smallest size the algorithm will consider to be a cluster. A higher cluster size will reduce the number of returned clusters, merging potential groups. This comes from HDBSCAN optimizing for the most stable groups depending on what may be considered a cluster.

2. `min_samples`

   The minimum number of samples sets how conservative the algorithm is in deciding to merge clusters and classify noise points. Higher values of `min_samples` will result in more noise points and smaller more local clusters.

3. `cluster_selection_epsilon`

   The value of `cluster_selection_epsilon` determines the minimum distance at which two clusters may separate. Setting this to a high value will result in large clusters, and a low value may result in microclusters.

Optimizing over the parameters to return realistic clusters representing stellar streams is a nontrivial computational task. One potential approach is gradient descent, where an algorithm would take small discrete steps in parameter space depending on which clustering resulted in minimizing a loss/error function. For example, given a starting set of parameters $(\texttt{min\_cluster\_size}, \texttt{min\_samples}, \texttt{cluster\_selection\_epsilon}) = (350, 5, 0.05)$ the algorithm would compare six separate clusterings resulting from the parameters $(350 \pm \delta, 5 \pm \delta, 0.05 \pm \delta)$ and evaluate the error function for each of the clusterings, then take a step toward the parameters which resulted in the lowest error. In a stochastic implementation of gradient descent, only two clusterings would be computed at each step in either direction of a randomly chosen parameter. In our case the error function would be a comparison between the clustering and some known classification of stellar streams (see Section 4.1.3). Gradient descent methods are typically performed in several trials with different initial conditions to find the global minimum, rather than just the local minimum.

However, gradient descent and other nonlinear rootfinding methods are not feasible for the task because of the computation time required for each iteration. For 2×10^5 particles, the clustering requires 12 minutes of computing time for a given set of parameters. My computer is only able to perform 5 steps per hour, too slow to implement a complex optimization scheme. Downsampling further is ineffective because the parameters are sensitive to the density and precise distribution of data. These parameters should also be general, so that they do not have to be tuned either by hand or algorithmically depending on the galaxy. For this thesis, I fine-tuned two parameters (`min_samples`, `cluster_selection_epsilon`) according to the output cluster distribution and ran a binary search over values of `min_cluster_size` to return a given number of clusters. However, given more computational resources, my clustering code can be modified to optimize over several parameters.

4.1.3 Performance and Accuracy

To evaluate the performance of the cluster algorithm, I have chosen to use the **Rand index** [Rand, 1971], which calculates a similarity score between the predicted and true clusterings. The score is close to 0.0 for random labels, and 1.0 for identical labels (or a permutation). The Rand index is also bounded below zero by −0.5 for highly incompatible clusterings.

Given a set of n elements S (stellar particles in our case) and two partitions $X = X_1, X_2, ...X_n$ and $Y = Y_1, Y_2, ...Y_n$ (clusterings), define the following:

- a: The number of pairs of elements in S that are in the **same** subset in X and Y
- b: The number of pairs that are in **different** subsets in both X and Y
- c: The number of pairs that are in the **same** subset in X and **different** subsets in Y
- d: The number of pairs that are in **different** subsets in X and the **same** subset in Y

In simpler terms, the term $a + b$ is the **number of agreements** between the partitions, and the term $c + d$ is the **number of disagreements**. The Rand index is defined as

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}. \tag{4.1}$$

The Rand index is bounded between $[0, +1]$, and is intuitive for comparing partitions in data. However, it does not account well for partitions with drastically different numbers or sizes of subsets. Because of this, I chose the **adjusted Rand index** as described by Hubert and Arabie [1985], which adjusts the Rand index for random chance in categorizing elements. Due to this change, **the adjusted Rand index is bounded by** $[\mathbf{-0.5}, \mathbf{1.0}]$. In either case, **a high Rand index indicates a similar clustering**, while **a low Rand index indicates a dissimilar clustering**.

Using a wide range of clustering parameters, I was unable to achieve a high Rand index when compared to the ground truth stream candidates (compare Figures 4.1, 4.2). When clustering in 6D space (x, y, z, jr, jz, jp), some clusters do resemble visible features in the extracted streams. However, there are too many overlaps and separations to resolve a vast majority of the stream candidates. The adjusted Rand indices were near zero ($\approx -.008$) for a wide range of clustering parameters. Further improvements may be possible by employing filters based on metallicity, or a clustering algorithm more optimized for astronomical data.

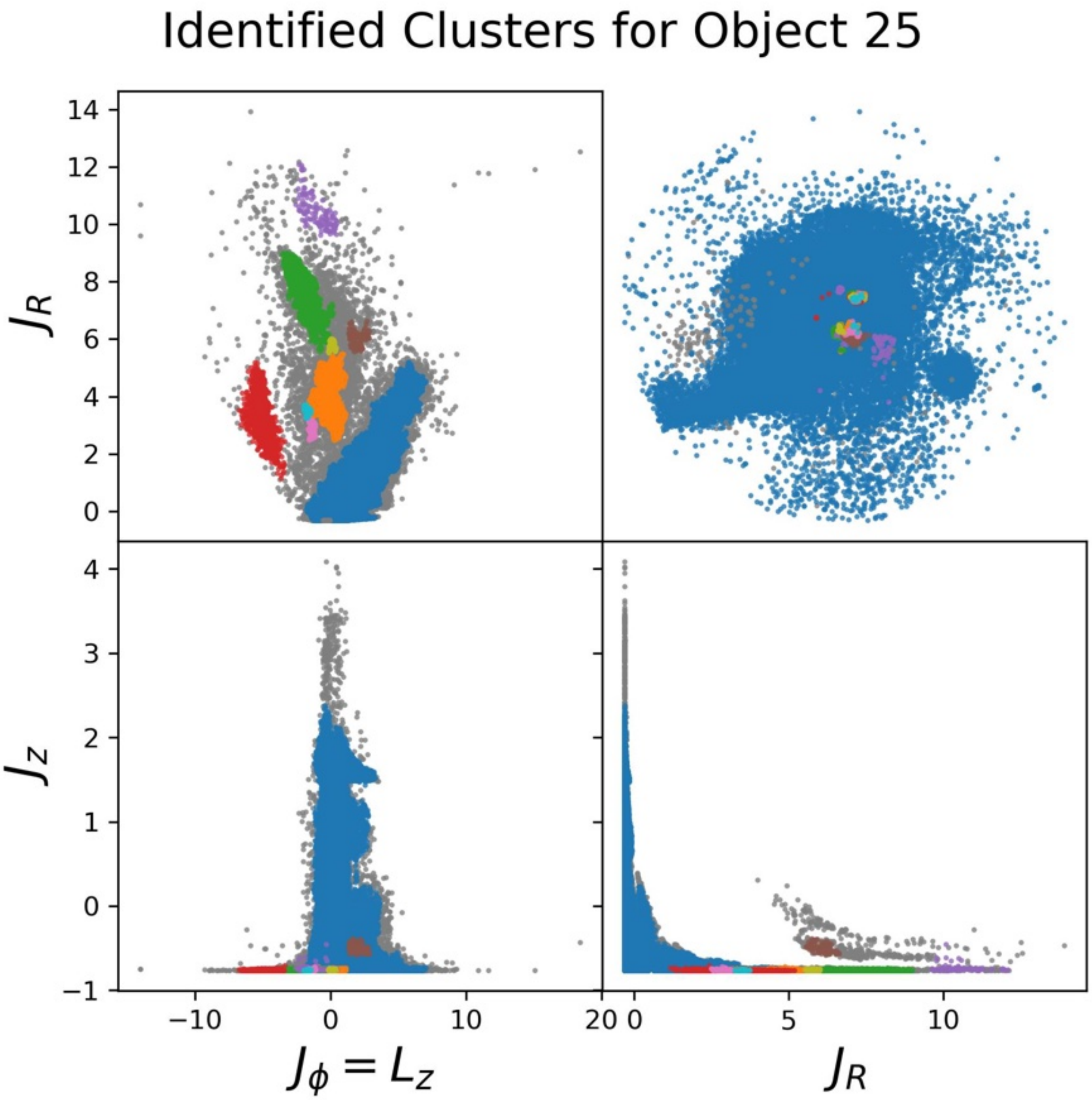

Figure 4.1: Results of HDBscan clustering over a normalized 6D space including the three actions as well as cartesian coordinates $(x, y, z, J_R, J_Z, J_\phi)$. The top-left, bottom-left, and bottom-right panels show projections of the 3D action space, with identified streams indicated by color. Noise points are plotted in gray (Section 3.4.1). The top right panel shows a face-on (x, y) projection of galaxy 25. The Rand score, which compares the clustering to the extracted streams from FIREbox on a scale of $[-0.5, 1.0]$ was $R = -0.0086$ corresponding to a mostly random partition with respect to the stream candidates. The clustering parameters were tuned to identify 10 distinct clusters. The parameters were: `minimum_cluster_size` $= 43$, `cluster_selection_epsilon` $= 0.1$, and the `min_samples` $= 6$. The data has been scaled and non-dimensionalized by the clustering algorithm.

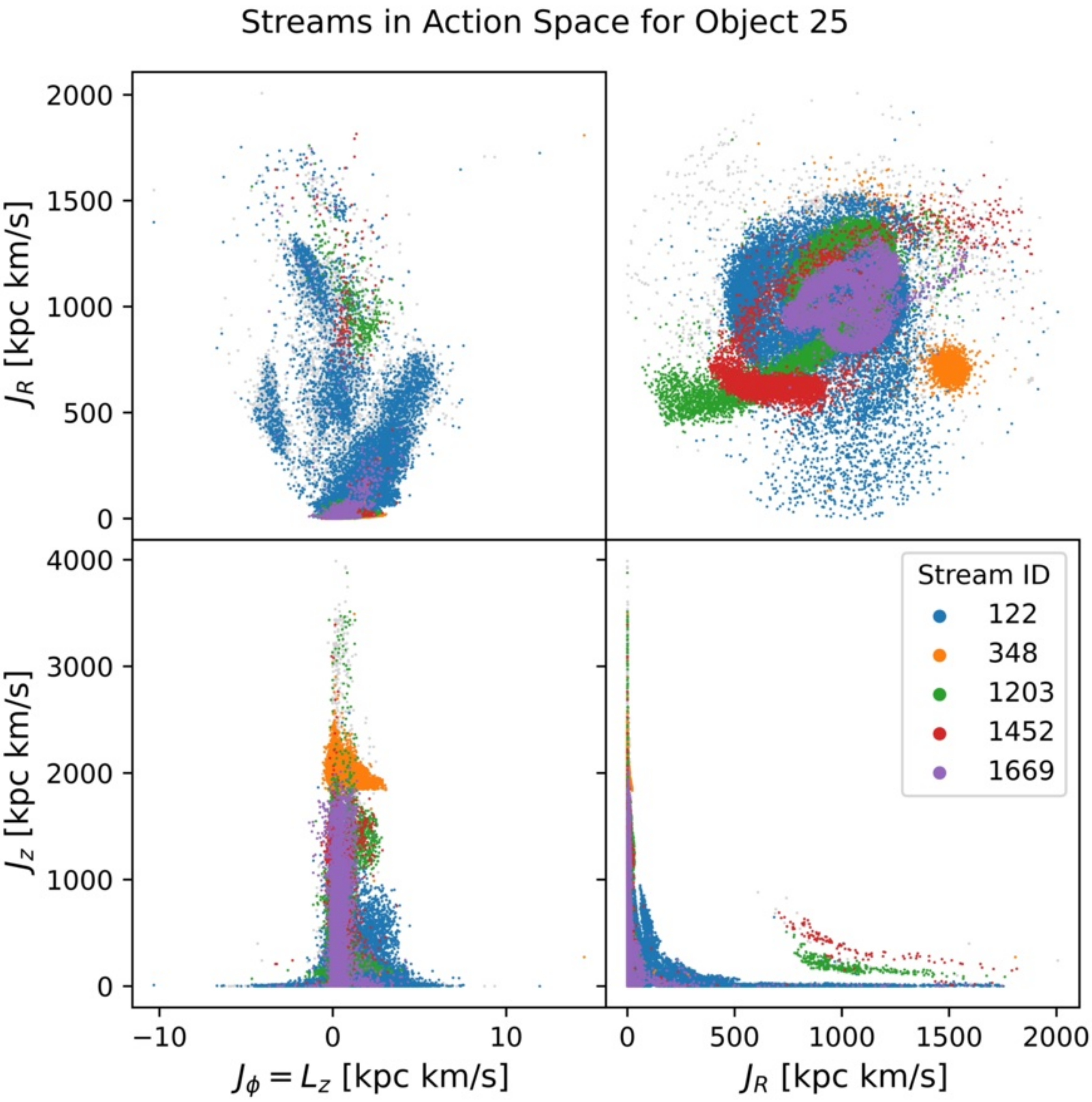

Figure 4.2: Stellar stream candidates found by identifying particles in satellites of the host galaxy with ID 25 from $z = 0.5$. Streams remain clustered in action space, even after wrapping many times around the host. In particular, Stream 384 (orange) has not been significantly disrupted by the host, and is still recognizable as a satellite.

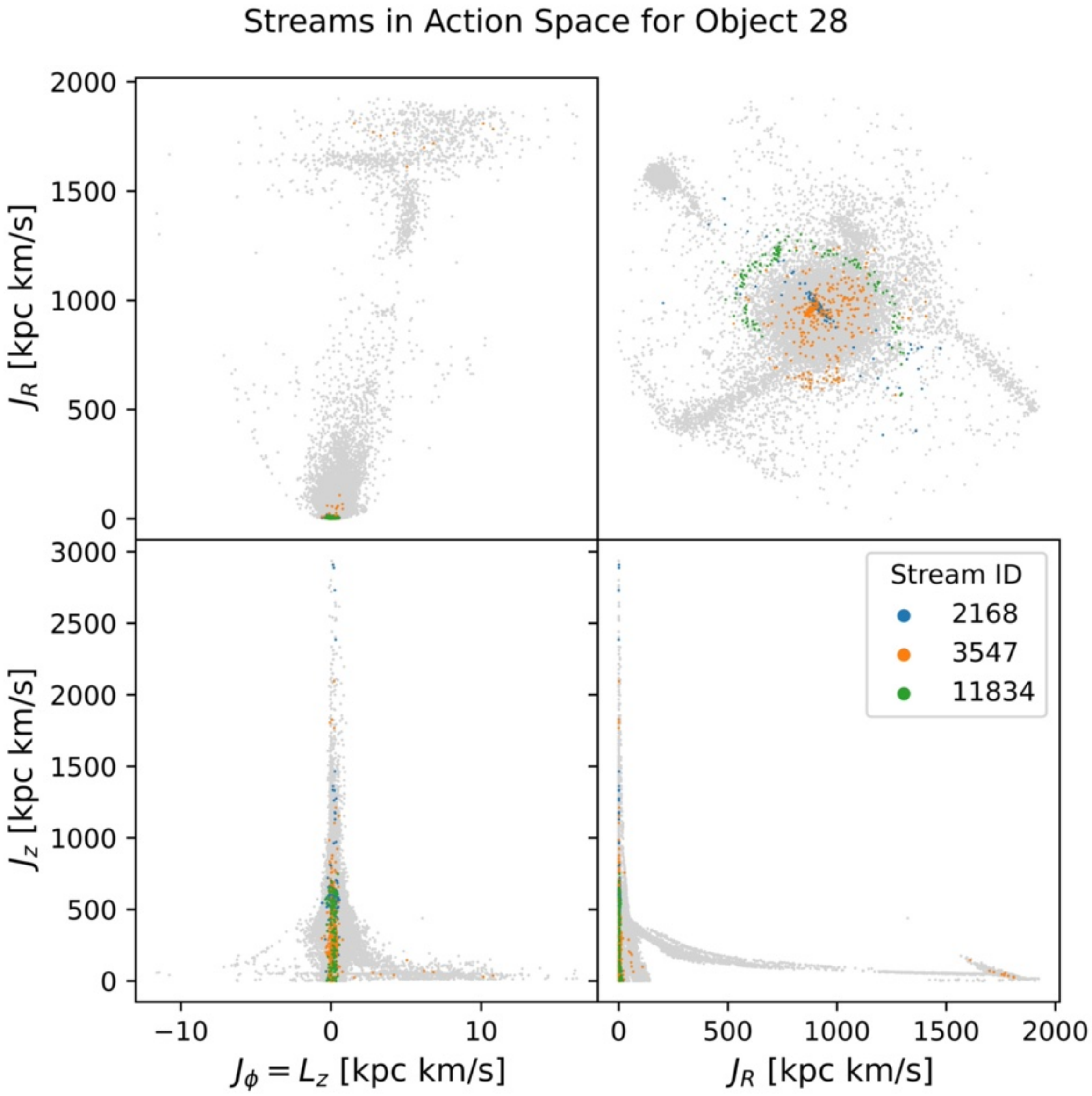

Figure 4.3: Stellar stream candidates found by identifying particles in satellites of the host galaxy with ID 28 from $z = 0.5$. Although Stream 11834 (green) has been stretched out by the tidal force of the host, it remains coherent in action space.

Chapter 5

Conclusion

I set out to identify and analyze stellar streams in FIREbox. I processed and filtered simulation data, converted kinematic data to phase space information, and optimized a clustering algorithm to find overdensities in action space. Then, I compared the results to stellar stream candidates as extracted from previous snapshots of the simulation. **My thesis consists of three main scientific products:**

1. an **action finding pipeline**,
2. a **catalog of stellar stream candidates in FIREbox**, and
3. an **analysis of HDBSCAN clustering for stream finding.**

The **action finding pipeline** will be made publicly available within the collaboration, and will be particularly useful to those interested in investigating the kinematic properties of galaxies in FIREbox or a comparable simulation. Code and data for this project are available in the `stellar_streams` repository on Github[1]. I developed the pipeline to calculate the action integrals for FIREbox galaxies based on a compound best-fit gravitational potential. In future work, these potentials could be tested against different models including time-dependent and higher-order approximations as in Arora et al. [2022]. When combined with the stellar stream catalog extracted from FIREbox, it becomes possible to visualize the streams in action space as in Figures 4.2, 4.3, A.2, and A.3. These resources together create the opportunity to study the dynamics of stellar streams in FIREbox, improving predictions about future stream observations.

The **catalog of stellar stream candidates in FIREbox** will also be made publicly available within the collaboration, and will serve as a valuable resource for studying stellar streams around FIREbox galaxies. For example, those interested in testing stream finding algorithms in preparation for extragalactic stream observations from Vera Rubin, Euclid, ARRAKIHS, and the Nancy Grace Roman Space Telescope in the coming decade [Walder et al., 2024] will be able to quickly compare their results with the galaxy's merger history.

[1] https://github.com/benhanf/stellar_streams

Another approach involves developing a stream finding algorithm using the stream catalog to create mock observations and validate its performance. Other researchers may want to investigate the properties of stellar streams in FIREbox, such as frequency, dimension, metallicity, and structure. These properties could provide predictions in advance of observations to test our theories of dark matter and galactic evolution. See Figures A.1 for a selected visualizations of stellar streams in FIREbox.

Clustering in phase space is a challenging problem, and must be combined with other information (such as metallicity) to provide accurate identification of stellar substructure. Other approaches to streamfinding make use of explicit potential models, integrating the orbits of stars to extract more dimensions of data from kinematic observations [Malhan and Ibata, 2018]. Others utilize metallicity filters to reduce the noise associated with foreground stars [Sanderson et al., 2017], adopt different clustering algorithms [Wu et al., 2021], or cluster in different phase spaces, such as energy and angular momentum [Bonaca and Price-Whelan, 2025]. I found that **the optimization process to use DBSCAN to find stellar streams in action space data is nontrivial and computationally intensive**. The algorithm's parameters did not converge quickly on an accurate classification, nor did the parameters generalize well across a sample of similar datasets. However, phase space clustering has proved to be a useful tool in some situations, as in the case of Enlink [Sharma and Johnston, 2009].

Convolutional neural networks are an increasingly popular machine learning approach to astronomical data classification. They are trained iteratively on labeled data to identify features like spiral tails, shells, and streams in other galaxies [Gordon et al., 2024, Vera-Casanova et al., 2025, Huertas-Company and Lanusse, 2023]. There are limitations to the interpretability of deep learning algorithms, as well as open questions about their optimization and scaling. Still, machine learning techniques in astronomy are still in their early stages, and will only improve with access to a greater quantity and quality of observations. I am optimistic about the future of astronomy, and hope that technological advancements will enhance our understanding of the Universe.

Appendix A

Stellar Stream Catalog Visualizations

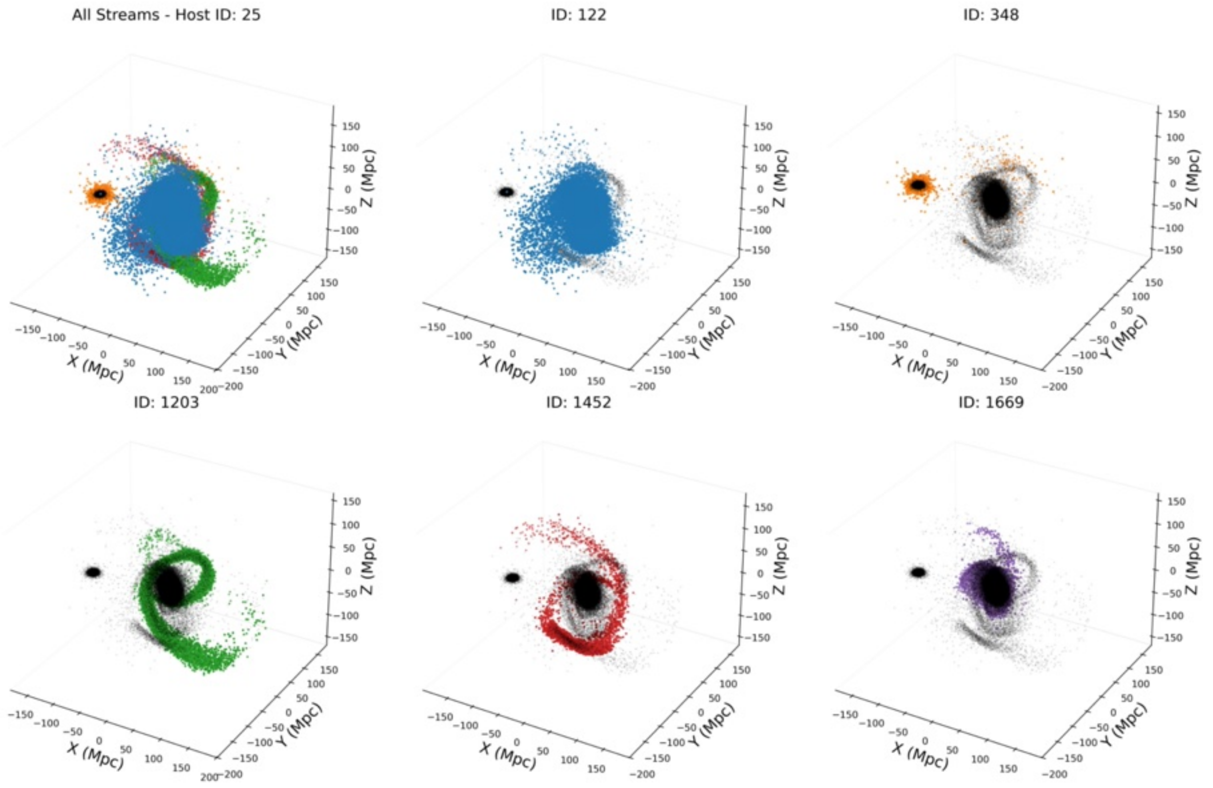

Figure A.1: Sample of stellar stream candidates found by identifying particles in satellites of the host galaxy from $z = 0.5$.

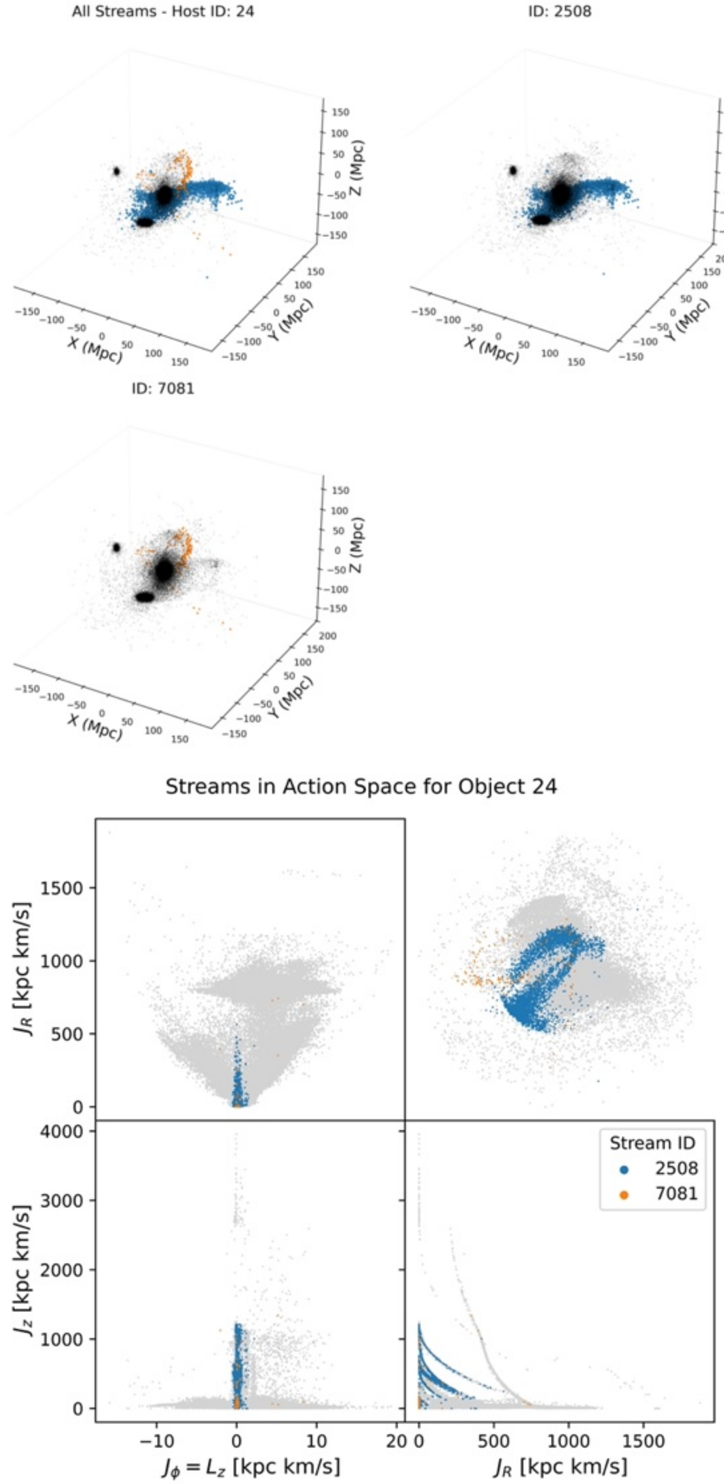

Figure A.2: Stellar stream candidates found by identifying particles in satellites of the host galaxy with ID 24 from $z = 0.5$. The action finding pipeline calculates their actions based on the current ($z = 0$) locations, velocites, and potential.

Isolated Streams in Action Space, Object 25

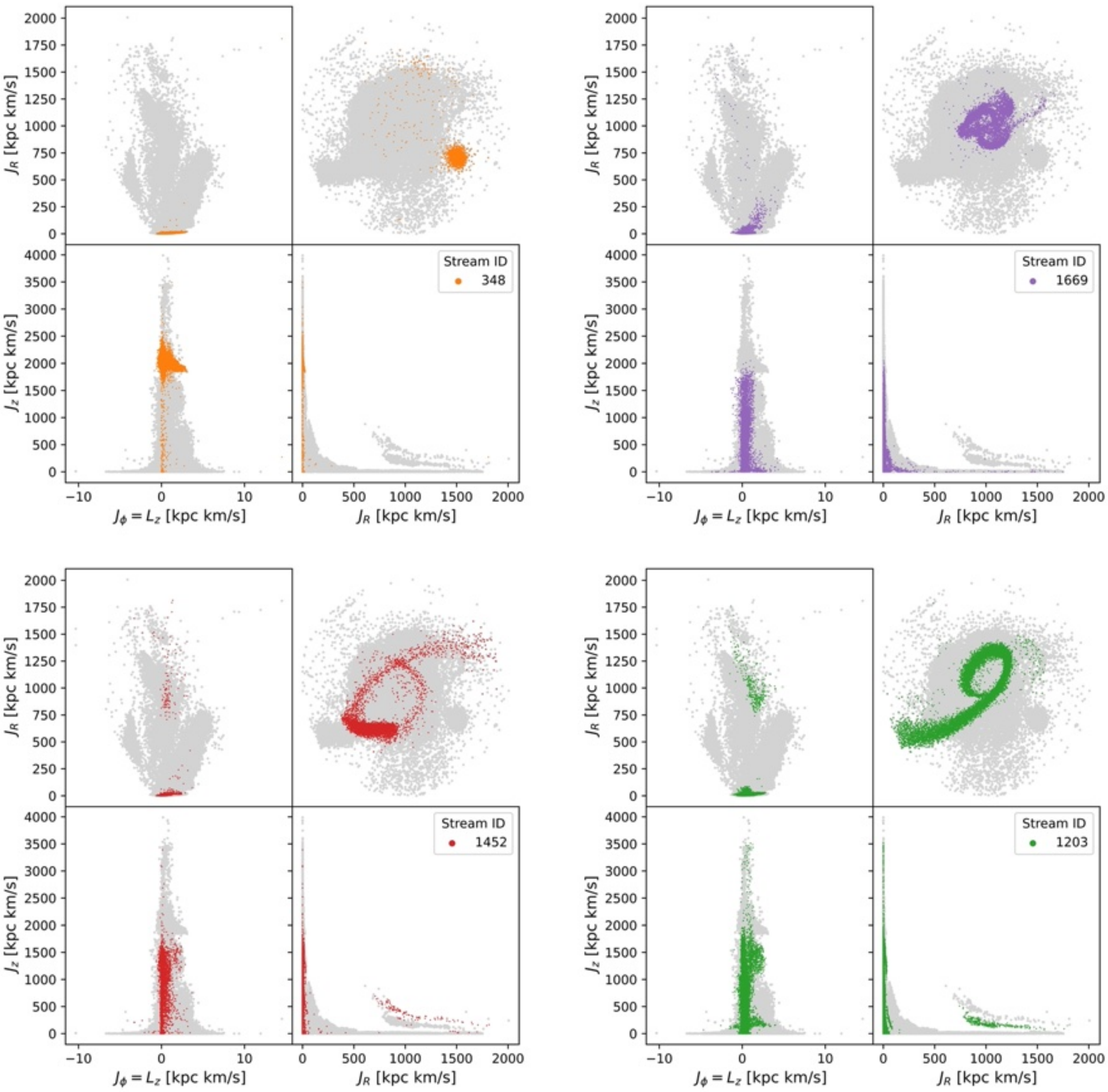

Figure A.3: Four selected streams of host 25 in 2D projected (x, y) and action space.

Appendix B

Code and Computational Resources

This thesis made use of the following Python packages:

- `Galpy` [Bovy, 2015]
- `Agama` [Vasiliev, 2018]
- `Scikit-learn` [Pedregosa et al., 2011]
- `Matplotlib` [Hunter, 2007]
- `Numpy` [Harris et al., 2020]

Over the course of this thesis I benefited from access to the Green Planet Computing Cluster at UC Irvine via Professor James Bullock and his research group. My work also relied on FIREbox data prepared by Professor Jorge Moreno.

Code developed for this thesis, as well as notebooks and data for generating figures, is hosted on https://github.com/benhanf/stellar_streams.

Bibliography

Christian Aganze, Sarah Pearson, Tjitske Starkenburg, Gabriella Contardo, Kathryn V. Johnston, Kiyan Tavangar, Adrian M. Price-Whelan, and Adam J. Burgasser. Prospects for detecting gaps in globular cluster stellar streams in external galaxies with the nancy grace roman space telescope. *The Astrophysical Journal*, 962(2):151, February 2024. ISSN 1538-4357. doi: 10.3847/1538-4357/ad159c. URL http://dx.doi.org/10.3847/1538-4357/ad159c.

Arpit Arora, Robyn E. Sanderson, Nondh Panithanpaisal, Emily C. Cunningham, Andrew Wetzel, and Nicolás Garavito-Camargo. On the Stability of Tidal Streams in Action Space. *ApJ*, 939(1):2, November 2022. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ac93fb. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac93fb.

James Binney and Scott Tremaine. *Galactic Dynamics: Second Edition*. Princeton University Press, rev - revised, 2 edition, 2008. ISBN 9780691130262. URL http://www.jstor.org/stable/j.ctvc778ff.

Joss Bland-Hawthorn and Ortwin Gerhard. The galaxy in context: Structural, kinematic, and integrated properties. *Annual Review of Astronomy and Astrophysics*, 54(1):529–596, September 2016. ISSN 1545-4282. doi: 10.1146/annurev-astro-081915-023441. URL http://dx.doi.org/10.1146/annurev-astro-081915-023441.

Ana Bonaca and Adrian M. Price-Whelan. Stellar streams in the gaia era. *New Astronomy Reviews*, 100:101713, 2025. ISSN 1387-6473. doi: https://doi.org/10.1016/j.newar.2024.101713. URL https://www.sciencedirect.com/science/article/pii/S1387647324000204.

Nicholas W Borsato, Sarah L Martell, and Jeffrey D Simpson. Identifying stellar streams in *Gaia* DR2 with data mining techniques. *Monthly Notices of the Royal Astronomical Society*, 492(1):1370–1384, February 2020. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stz3479. URL https://academic.oup.com/mnras/article/492/1/1370/5681406.

Jo Bovy. galpy: A python library for galactic dynamics. *The Astrophysical Journal Supplement Series*, 216(2):29, February 2015. ISSN 1538-4365. doi: 10.1088/0067-0049/216/2/29. URL http://dx.doi.org/10.1088/0067-0049/216/2/29.

Jo Bovy. *Dynamics and Astrophysics of Galaxies.* Princeton University Press, 1 edition, 2026. URL https://galaxiesbook.org.

Greg L. Bryan and Michael L. Norman. Statistical properties of x-ray clusters: Analytic and numerical comparisons. *The Astrophysical Journal*, 495(1):80, mar 1998. doi: 10.1086/305262. URL https://dx.doi.org/10.1086/305262.

James S. Bullock and Kathryn V. Johnston. Tracing galaxy formation with stellar halos. i. methods. *The Astrophysical Journal*, 635(2):931–949, December 2005. ISSN 1538-4357. doi: 10.1086/497422. URL http://dx.doi.org/10.1086/497422.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Robert Feldmann, Eliot Quataert, Claude-André Faucher-Giguère, Philip F Hopkins, Onur Çatmabacak, Dušan Kereš, Luigi Bassini, Mauro Bernardini, James S Bullock, Elia Cenci, Jindra Gensior, Lichen Liang, Jorge Moreno, and Andrew Wetzel. Firebox: simulating galaxies at high dynamic range in a cosmological volume. *Monthly Notices of the Royal Astronomical Society*, 522(3):3831–3860, 04 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad1205. URL https://doi.org/10.1093/mnras/stad1205.

C.S. Frenk and S.D.M. White. Dark matter and cosmic structure. *Annalen der Physik*, 524 (9–10):507–534, September 2012. ISSN 1521-3889. doi: 10.1002/andp.201200212. URL http://dx.doi.org/10.1002/andp.201200212.

Gaia Collaboration. Gaia Data Release 2. Summary of the contents and survey properties. , 616:A1, August 2018. doi: 10.1051/0004-6361/201833051. URL https://ui.adsabs.harvard.edu/abs/2018A&A...616A...1G.

Alexander J Gordon, Annette M N Ferguson, and Robert G Mann. Uncovering tidal treasures: automated classification of faint tidal features in decals data. *Monthly Notices of the Royal Astronomical Society*, 534(2):1459–1480, 09 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae2169. URL https://doi.org/10.1093/mnras/stae2169.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

T. M. Helliwell and V. V. Sahakian. *Modern Classical Mechanics*. Cambridge University Press, 2020.

Amina Helmi. Streams, Substructures, and the Early History of the Milky Way. *Annu. Rev. Astron. Astrophys.*, 58(1):205–256, August 2020. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-032620-021917. URL https://www.annualreviews.org/doi/10.1146/annurev-astro-032620-021917.

Philip F. Hopkins. A new class of accurate, mesh-free hydrodynamic simulation methods. , 450(1):53–110, June 2015. doi: 10.1093/mnras/stv195. URL https://ui.adsabs.harvard.edu/abs/2015MNRAS.450...53H.

Philip F. Hopkins, Andrew Wetzel, Dušan Kereš, Claude-André Faucher-Giguère, Eliot Quataert, Michael Boylan-Kolchin, Norman Murray, Christopher C. Hayward, Shea Garrison-Kimmel, Cameron Hummels, Robert Feldmann, Paul Torrey, Xiangcheng Ma, Daniel Anglés-Alcázar, Kung-Yi Su, Matthew Orr, Denise Schmitz, Ivanna Escala, Robyn Sanderson, Michael Y. Grudić, Zachary Hafen, Ji-Hoon Kim, Alex Fitts, James S. Bullock, Coral Wheeler, T. K. Chan, Oliver D. Elbert, and Desika Narayanan. FIRE-2 simulations: physics versus numerics in galaxy formation. , 480(1):800–863, October 2018. doi: 10.1093/mnras/sty1690. URL https://ui.adsabs.harvard.edu/abs/2018MNRAS.480..800H.

Philip F. Hopkins, Andrew Wetzel, Coral Wheeler, Robyn Sanderson, Michael Y. Grudic, Omid Sameie, Michael Boylan-Kolchin, Matthew Orr, Xiangcheng Ma, Claude-Andre Faucher-Giguere, Dusan Keres, Eliot Quataert, Kung-Yi Su, Jorge Moreno, Robert Feldmann, James S. Bullock, Sarah R. Loebman, Daniel Angles-Alcazar, Jonathan Stern, Lina Necib, and Christopher C. Hayward. FIRE-3: Updated Stellar Evolution Models, Yields, & Microphysics and Fitting Functions for Applications in Galaxy Simulations. *arXiv e-prints*, art. arXiv:2203.00040, February 2022. URL https://ui.adsabs.harvard.edu/abs/2022arXiv220300040H.

Paul V.C. Hough. Method and means for recognizing complex patterns, 1962.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985. ISSN 1432-1343. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

M. Huertas-Company and F. Lanusse. The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys. , 40:e001, January 2023. doi: 10.1017/pasa.2022.55. URL https://ui.adsabs.harvard.edu/abs/2023PASA...40....1H.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Khyati Malhan and Rodrigo A Ibata. Streamfinder – i. a new algorithm for detecting stellar streams. *Monthly Notices of the Royal Astronomical Society*, 477(3):4063–4076, 04 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty912. URL https://doi.org/10.1093/mnras/sty912.

Cecilia Mateu. galstreams: A library of milky way stellar stream footprints and tracks. *Monthly Notices of the Royal Astronomical Society*, 520(4):5225–5258, January 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad321. URL http://dx.doi.org/10.1093/mnras/stad321.

Paul Menker and Andrew Benson. Advancing stellar streams as a dark matter probe – i: Evolution of the cdm subhalo population, 2024. URL https://arxiv.org/abs/2406.11989.

Jorge Moreno, Shany Danieli, James S. Bullock, Robert Feldmann, Philip F. Hopkins, Onur Çatmabacak, Alexander Gurvich, Alexandres Lazar, Courtney Klein, Cameron B. Hummels, Zachary Hafen, Francisco J. Mercado, Sijie Yu, Fangzhou Jiang, Coral Wheeler, Andrew Wetzel, Daniel Anglés-Alcázar, Michael Boylan-Kolchin, Eliot Quataert, Claude-André Faucher-Giguère, and Dušan Kereš. Galaxies lacking dark matter produced by close encounters in a cosmological simulation. *Nature Astronomy*, 6(4):496–502, February 2022. ISSN 2397-3366. doi: 10.1038/s41550-021-01598-4. URL http://dx.doi.org/10.1038/s41550-021-01598-4.

G C Myeong, N W Evans, V Belokurov, J L Sanders, and S E Koposov. Discovery of new retrograde substructures: the shards of Centauri? *Monthly Notices of the Royal Astronomical Society*, 478(4):5449–5459, June 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1403. URL https://doi.org/10.1093/mnras/sty1403. _eprint: https://academic.oup.com/mnras/article-pdf/478/4/5449/25105825/sty1403.pdf.

Nondh Panithanpaisal, Robyn E. Sanderson, Andrew Wetzel, Emily C. Cunningham, Jeremy Bailin, and Claude-André Faucher-Giguère. The Galaxy Progenitors of Stellar Streams around Milky Way–mass Galaxies in the FIRE Cosmological Simulations. *ApJ*, 920(1): 10, October 2021. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ac1109. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac1109.

Sarah Pearson, Susan E. Clark, Alexis J. Demirjian, Kathryn V. Johnston, Melissa K. Ness, Tjitske K. Starkenburg, Benjamin F. Williams, and Rodrigo A. Ibata. The hough stream spotter: A new method for detecting linear structure in resolved stars and application to the stellar halo of m31. *The Astrophysical Journal*, 926(2):166, feb 2022a. doi: 10.3847/1538-4357/ac4496. URL https://dx.doi.org/10.3847/1538-4357/ac4496.

Sarah Pearson, Adrian M. Price-Whelan, David W. Hogg, Anil C. Seth, David J. Sand, Jason A. S. Hunt, and Denija Crnojević. Mapping dark matter with extragalactic stellar streams: The case of centaurus a. *The Astrophysical Journal*, 941(1):19, December 2022b. ISSN 1538-4357. doi: 10.3847/1538-4357/ac9bfb. URL http://dx.doi.org/10.3847/1538-4357/ac9bfb.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of*

Machine Learning Research, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Adrian M. Price-Whelan and Ana Bonaca. Off the beaten path: Gaia reveals gd-1 stars outside of the main stream. *The Astrophysical Journal Letters*, 863(2):L20, August 2018. ISSN 2041-8213. doi: 10.3847/2041-8213/aad7b5. URL http://dx.doi.org/10.3847/2041-8213/aad7b5.

R. C. Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957. doi: 10.1002/j.1538-7305.1957.tb01515.x.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2284239.

Amy E. Reines. Hunting for massive black holes in dwarf galaxies. *Nature Astronomy*, 6: 26–34, January 2022. doi: 10.1038/s41550-021-01556-0. URL https://ui.adsabs.harvard.edu/abs/2022NatAs...6...26R.

Jason L. Sanders and James Binney. A review of action estimation methods for galactic dynamics. *Mon. Not. R. Astron. Soc.*, 457(2):2107–2121, April 2016. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stw106. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw106.

Jason L. Sanders, Jo Bovy, and Denis Erkal. Dynamics of stream–subhalo interactions. *Monthly Notices of the Royal Astronomical Society*, 457(4):3817–3835, 01 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw232. URL https://doi.org/10.1093/mnras/stw232.

Robyn E. Sanderson, Andrew R. Wetzel, Sanjib Sharma, and Philip F. Hopkins. Better Galactic Mass Models through Chemistry. *Galaxies*, 5(3):43, August 2017. doi: 10.3390/galaxies5030043. URL https://ui.adsabs.harvard.edu/abs/2017Galax...5...43S.

Isaiah B Santistevan, Andrew Wetzel, Erik Tollerud, Robyn E Sanderson, Jorge Moreno, and Ekta Patel. Modelling the orbital histories of satellites of milky way-mass galaxies: testing static host potentials against cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 527(3):8841–8864, 12 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad3757. URL https://doi.org/10.1093/mnras/stad3757.

Sanjib Sharma and Kathryn V. Johnston. A Group Finding Algorithm For Multidimensional Data Sets. *ApJ*, 703(1):1061–1077, September 2009. ISSN 0004-637X, 1538-4357. doi: 10.1088/0004-637X/703/1/1061. URL https://iopscience.iop.org/article/10.1088/0004-637X/703/1/1061.

David Shih, Matthew R Buckley, Lina Necib, and John Tamanas. via machinae: Searching for stellar streams using unsupervised machine learning. *Monthly Notices of the Royal*

Astronomical Society, 509(4):5992–6007, 11 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab3372. URL https://doi.org/10.1093/mnras/stab3372.

Nora Shipp, Nondh Panithanpaisal, Lina Necib, Robyn Sanderson, Denis Erkal, Ting S. Li, Isaiah B. Santistevan, Andrew Wetzel, Lara R. Cullinane, Alexander P. Ji, Sergey E. Koposov, Kyler Kuehn, Geraint F. Lewis, Andrew B. Pace, Daniel B. Zucker, Joss Bland-Hawthorn, Emily C. Cunningham, Stacy Y. Kim, Sophia Lilleengen, Jorge Moreno, Sanjib Sharma, and S Collaboration & FIRE Collaboration. Streams on FIRE: Populations of Detectable Stellar Streams in the Milky Way and FIRE. *ApJ*, 949(2):44, June 2023. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/acc582. URL https://iopscience.iop.org/article/10.3847/1538-4357/acc582.

Adam R. H. Stevens, Claudia del P. Lagos, Sergio Contreras, Darren J. Croton, Nelson D. Padilla, Matthieu Schaller, Joop Schaye, and Tom Theuns. How to get cool in the heat: comparing analytic models of hot, cold, and cooling gas in haloes and galaxies with eagle. *Monthly Notices of the Royal Astronomical Society*, page stx243, January 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx243. URL http://dx.doi.org/10.1093/mnras/stx243.

Eugene Vasiliev. Agama: action-based galaxy modelling architecture. *Monthly Notices of the Royal Astronomical Society*, 482(2):1525–1544, October 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty2672. URL http://dx.doi.org/10.1093/mnras/sty2672.

Alex Vera-Casanova, Nicolas Monsalves Gonzalez, Facundo A. Gómez, Marcelo Jaque Arancibia., Valentina Fontirroig, Diego Pallero., Rüdiger Pakmor., Freeke van de Voort., Robert J. J. Grand., Rebekka Bieri., and Federico Marinacci. Stream automatic detection with convolutional neural network (sad-cnn), 2025. URL https://arxiv.org/abs/2503.17202. Preprint, submitted to A&A March 21, 2025.

Madison Walder, Denis Erkal, Michelle Collins, and David Martinez-Delgado. Probing the dark matter haloes of external galaxies with stellar streams, February 2024. URL http://arxiv.org/abs/2402.13314. arXiv:2402.13314 [astro-ph].

S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183(3):341–358, 07 1978. ISSN 0035-8711. doi: 10.1093/mnras/183.3.341. URL https://doi.org/10.1093/mnras/183.3.341.

Youjia Wu, Monica Valluri, Nondh Panithanpaisal, Robyn E Sanderson, Katherine Freese, Andrew Wetzel, and Sanjib Sharma. Using action space clustering to constrain the recent accretion history of Milky Way-like galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(4):5882–5901, December 2021. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stab3306. URL https://academic.oup.com/mnras/article/509/4/5882/6430175.