Senior thesis in Computer Science, Cosmology

# EVOLUTION_MACHINE: SATELLITE_GALAXIES

Thesis Subject:

יצירת רשת למידה ללימוד וחיזוי ההתפתחות הגלקטית בגלקסיות הלוויין

*Conducted in the "Alpha" program of the Edmond J. Safra campus of the Hebrew University*


Submitted by: Adam Beili

I.D: 213868789

Address: HaShofet Haim Cohen, Jerusalem

Email: adam.v.beili@gmail.com

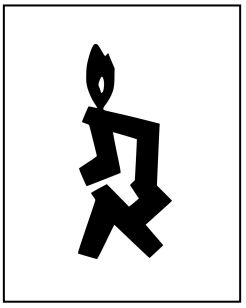
Phone Number: 050-682-6646

School: The Hebrew University Secondary School

School Number: 140061

Grade: 12th


*The thesis was conducted at the Edmond J. Safra campus of the Hebrew University, under Tomer Nussbaum and Or Sharir, in the laboratories of Avishai Dekel and Amnon Shashua*

# Acknowledgments

I have learned so much during my participation in the Alpha program. It is difficult to express my gratitude to all the people who made this a reality, but I'll sure as hell try.

First and foremost, I'd like to thank my mentor, Tomer Nussbaum. I have learned so much from my time in our lab, and none of it would be possible without him. Tomer has created the warmest, most challenging, and inspiring environment and has taught me so much, and not just in the scientific field. In the nearly two and a half years in our lab, I have learned so much about problem-solving, hard work, and self-confidence. I really can not explain how much I owe him.

Thanks to my lab-mates, Zohar Milman, Raphael Buzaglo, Naftali Deutsch, and Noam Chouchena. No matter what the problem was, Naftali, Zohar, and Buzaglo were always there to offer help and support me, especially in the beginning, when I felt like nothing was going my way.

I'd like to thank the organizers and supporters of the Alpha program for giving me this opportunity. To Shira Hirsh, Alon Oppenheim, and my alpha group guide Ori: I don't think any of you know how important this project has been for me. To say that alpha has been the highlight of my life for the past two years is an understatement, and it had provided me with so much of what I lacked when I first made Aliyah to Israel. Thank you.

Thank you to my closest and dearest friends, Maximilian Pochapski and Sophi Rochlin. I don't think you two know how much you inspire me to work harder and be better every day and how lucky I feel to have known you.

And last but not least, Thank you to my parents, Stella and Vadim, who have created opportunities and fought for me every step of my life; I am forever indebted to you.

# Preface

> "The cosmos is within us. We are made of star-stuff. We are a way for the universe to know itself." – Carl Sagan.

I entered the Alpha program in the spring of 2018, on a last-second recommendation by a friend. I don't know how I passed the exam without understanding a single word in Hebrew, but fate was on my side. I enrolled in Tomer's cosmology lab later that fall. I truly had no idea how important that opportunity would be for my personal evolution (We haven't even gotten to the galaxies yet) - I can't even begin to summarize the extent to which Tomer, my lab-mates, and the program at large have affected me. This was, without a doubt, the best decision I have made, and I am sure that what I have learned will accompany me for the rest of my life.

# Abstract

שלב הפסקת ייצור הכוכבים (quenching) בגלקסיות הוא שלב משמעותי בהתפתחות הגלקסיה. גלקסיות לווייניות (Satellite galaxies ~ SGs) הן גלקסיות המקיפות גלקסיה מרכזית (Central galaxy ~ CG), הגדולה יותר מהן בכמה סדרי גודל. התכונות והאינטראקציות של ה-SGs עם סביבתן הופכות אותן לנושא חשוב למחקר, הן מבחינת הבנתינו של מבנה הגלקסיה והן מבחינת הבנת הסביבה.
בעבודה זו נציג חיזוי של התפתחות ה-SGs ותכונותיהן העיקריות לכך, על פי מידע התחלתי של SGs לפני כניסתן לשטח ההשפעה של ה-CG. הניתוח בוצע על ידי מדגם איכותי של 118 SGs מהסימולציה הקוסמולוגית בעלת רזולוציה גבוהה - VELA - הכוללת מערך נתונים ייחודי של SGs. בעזרת למידת מכונה (ML) אנו מוצאים 3 תכונות הדרושות לחיזוי quenching ב-SGs: אקסצנטריות המסלול, יחס מסות בין SG לבין CG, ורדיוס SG, כאשר תכונה קורלטיבית וחשובה נוספת היא צורת הרכיב הכוכבי של SGs. נוסף על כך, אנו מוצאים כי התפתחות הצורה של SGs שונה באופן משמעותי מהתפתחות הצורה של CGs - הבחנה חשובה הדורשת מחקר נוסף.

The cessation of star formation (quenching) in galaxies is an essential stage in the evolution of a galaxy. Satellite galaxies (SGs) are galaxies in orbit around a central galaxy (CG), larger in size by several orders of magnitude. The properties and interactions of SGs with their environment make them an important topic for learning, both in terms of our understanding of the galaxy's structure and understanding the environment.

In this work we present a prediction of the quenching of SGs and their main features, according to initial information of satellite galaxies before entering the sphere of influence of the CG. The analysis was performed by a qualitative sample of 118 satellite galaxies from the high-resolution cosmological simulation VELA, which includes a unique data set of satellite galaxies. Using machine learning (ML), we find 3 features needed for SG quenching prediction: orbital eccentricity, SG/CG mass ratio, and SG radius, with another correlative and important feature being SG stellar component shape. Additionally we find the shape evolution of SGs varies dramatically from CG evolution - an important distinction that requires further research.

# Research Goals

To build up the VELA_v2 SG catalog, by performing feature engineering . To study the shape of SGs using the ShapeTensor algorithm, as a possible influential feature.

Using Machine Learning, identify the key features, and create new model for prediction of quenching of SGs, and starburst events at peri-centers.

# 1 - Introduction

## 1.1 - Cosmology

Cosmology is the study of the Universe and its properties on the enormous scale possible. As cosmologists, we seek to answer the questions: Where did the Universe come from? How has it evolved? And how will it continue evolving?

The currently accepted theoretical model of our Universe is based upon the Big Bang theory, which states that the Universe began from a singularity - a single point of infinite mass and temperature - that expanded exponentially during a period of time known as the inflation epoch. During this period, the fundamental forces of the universe were established. The Universe at this stage was largely homogenous and uniformly distributed, with energy density fluctuations (in the order of 1 part in 100,000) becoming the basis for large-scale structure distribution we see in the present day universe (*Allday, 2002)*.

The Universe continued to expand and cool, for approximately 370,000 years, when the first atoms could form. These first atoms emitted electromagnetic radiation in the form of photons as they reached their energy ground state. These photons are now detectable as the cosmic microwave background (CMB) *(Carroll, 2007).*

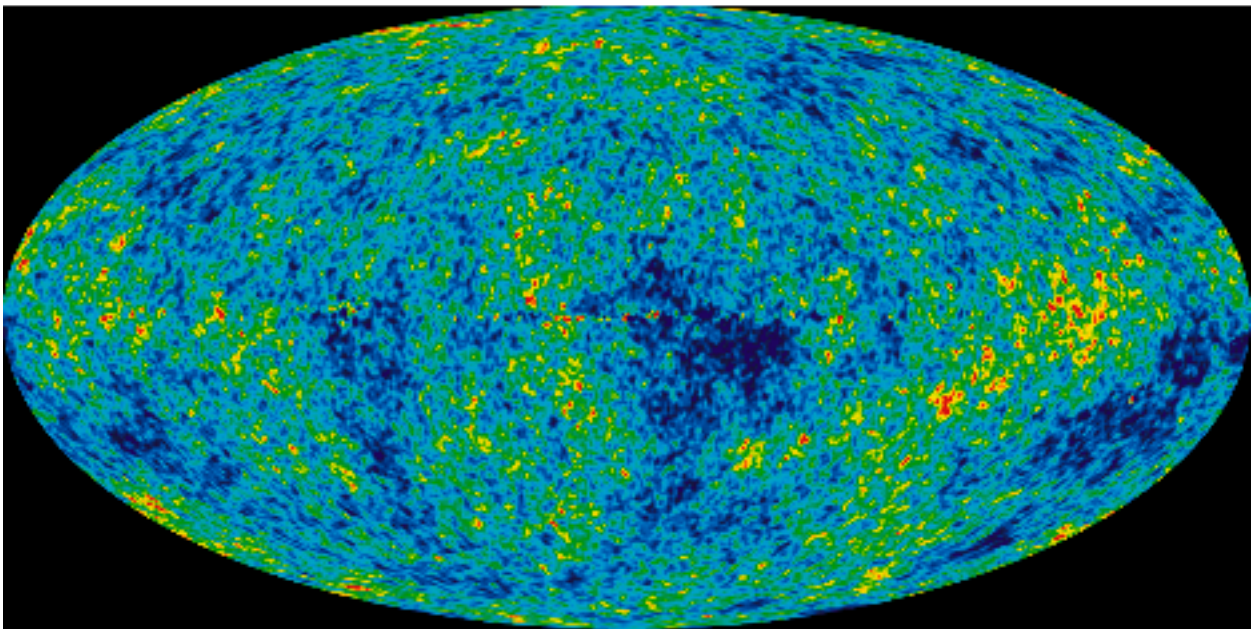


Figure 1: The Cosmic Microwave Background (CMB), as observed by 2009. Fluctuations in energy densities are thought to correspond to large scale distribution of the present universe.

Many unknowns still litter our latest cosmological theories. The Universe is still ever-expanding, at an accelerating rate. We observe missing mass from galaxies required to account for their rotational speed, gravitational lensing, and more. The Universe itself seems to be arranged by large-scale structures known as galactic filaments: web-like structures along which galaxies are held.

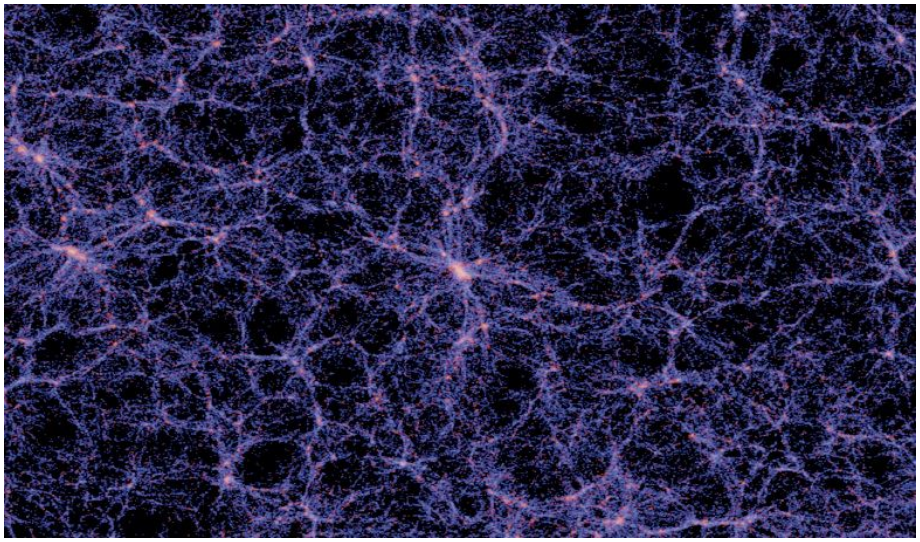


Figure 2: Galactic Filaments - 'The Cosmic Web'.

Recent advances in observations and data have given us the tools to precisely calculate fundamental properties of our Universe, such as the energy distribution. Out of the energy density within the known Universe, naught 5% is held within ordinary Baryonic Matter, the matter that makes up all stars, planets, and us. 95% of the energy of the Universe is actually held within two phenomena that - as of yet - we can only perceive through their effects. They are Dark Matter and Dark Energy. *(Carroll, 2007).*

Dark matter (DM) - an unseen matter that interacts solely through the gravitational field. DM is the currently accepted theoretical explanation for otherwise impossible effects seen within galaxies and the large scale structure of the Universe.

Dark energy -a vacuum energy that permeates throughout the fabric of space. Dark Energy would explain the accelerating expansion of space. The Cosmological constant - Lambda - is the frontrunner for Dark Energy. In this scenario, Dark Energy is a property of space itself, and is therefore uniform and homogenous across space.

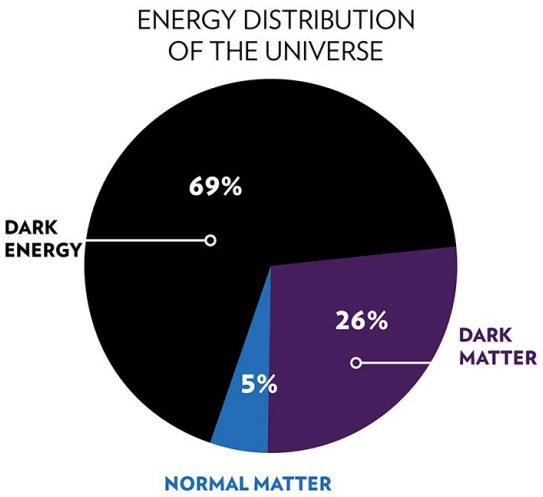



Figure 3: The energy distribution of the Universe. *Image: Chandra X-ray Observatory, 2008).*

The cosmological model that incorporates all these aspects is known as the Lambda-Cold Dark Matter (LCDM) model, and is the leading theoretical parametrization of the Universe. Lambda is the value associated with dark energy, and cold DM is a hypothesized variety of slow-moving weakly-interacting DM, where the only interactions that take place are gravitational. One ramification of this is the notion that DM cannot form solids, and can pass through itself and other matter undetected.

## 1.2 - Concerning Galaxies

A Galaxy is a gravitationally bound cosmological body consisting of gas, stars, and DM. Galaxies are observed and studied with great detail in order to further and improve our understanding of the development and the origin of the Universe.

### 1.2.1 - Galaxy Structure and Evolution

Galaxies, like all other aspects of the Universe, are ever-changing and evolving. Under the LCDM model, galaxies form using a hierarchical bottom-up process. In such a process, the development of galaxies begins along galactic filaments, where through clustering, clumps of DM - known as haloes - along with intergalactic gas are pulled together over time. As the density of cold-gas within the proto-galaxies increases, the star-formation process ignites. Further inputs of cold-gas feed the galaxies as they grow larger, accumulating more mass through clustering and merger events with other galaxies.

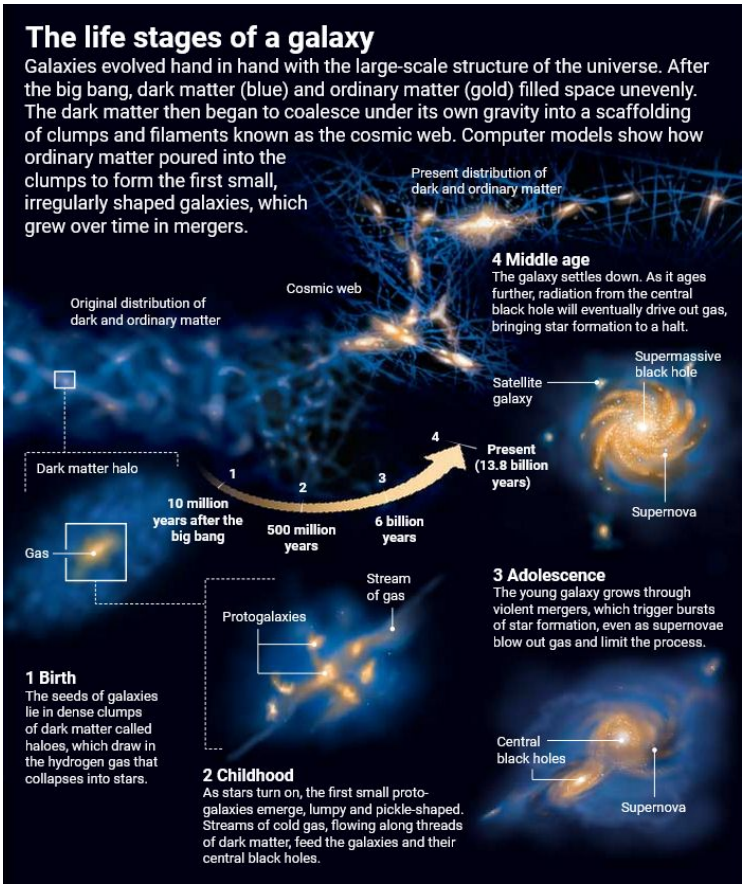



Figure 4: Evolution of a galaxy. *Image: Bickel, in Science (2018).*

As galaxies age, the central concentration of star-forming gasses gets depleted, which leads to a halt in the star formation rate (SFR) of the galaxy. This phenomenon is known as quenching.

Compaction events, where a massive inflow of gasses into the galaxy - caused by mergers with other smaller galaxies or external counter-rotating streams of gas around the DM Halo - reignite the SFR process. Eventually, as field galaxies reach a critical mass of at least $10^{10.5}$ stellar masses, they may fully quench; any inflow event of gasses can no longer restart the SFR process, and no new stars are formed in the galaxy *(Dekel et al., n.d.: Tacchella et al., 2016)*.

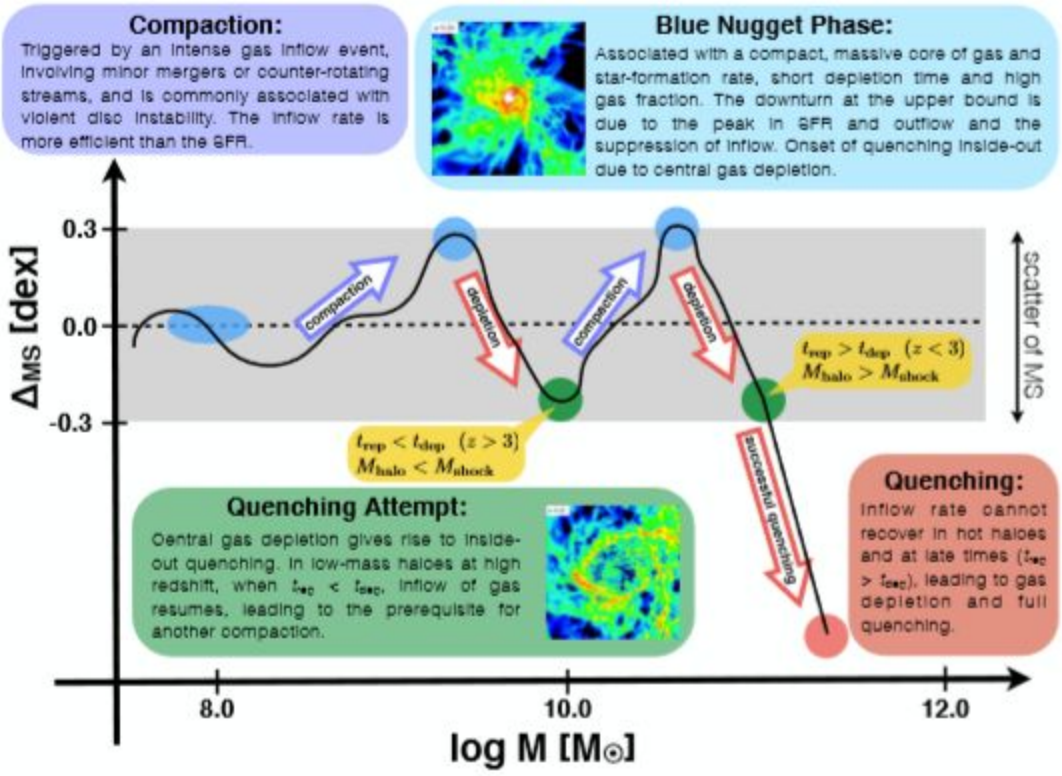



Figure 5: SFR-Stellar Mass Diagram of field galaxies. *Image: Tacchella et al., (2016)*: General evolution path of star forming field galaxies until quenching. As shown here, before reaching the $10^{10.5}$M critical mass, the galaxies become increasingly volatile, having multiple quenching attempts and compaction events until gas inflow is not enough to recover SFR, leading to full quenching.

In this paper we use the *specific star-formation rate* (sSFR), defined as the SFR per unit galaxy stellar mass. sSFR shows additional relations between galactic mass and SFR. $[sSFR] = [10^8/Gyr]$

For large scale galaxies, a typical structure can be seen as:

- A DM halo, a spherical DM component with a radius on the order of 10 times larger than the visible galactic component.

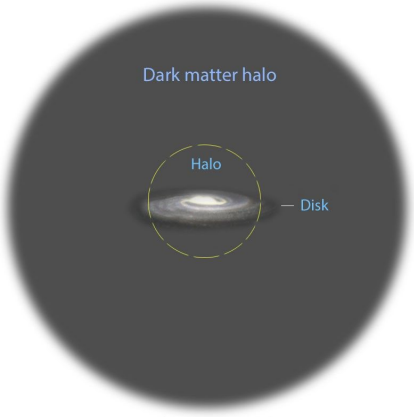



Figure 6: Visualization of the DM halo component, relative to the central, visible galactic component. *Image: The University of Zurich*

- Gas, of which cold-gas coalesces to form stars, and the remainder forms the rest of the structures within the galaxy.

Galaxies form in large super-clusters, and as has been found using satellite observational data and simulations, large galaxies accumulate companion satellite galaxies. These satellite galaxies undergo far more tumultuous lives than their central hosts.

## 1.3 - Satellite Galaxies

A rapidly emerging subfield of study in cosmology is that of satellite galaxies (SG) - galaxies that are captured by the massive gravitational well of a central galaxy (CG), and orbit within the DM halo of their CG host. We define satellite galaxies as a galactic body with a mass less than a quarter the size of the central *(Nussbaum, 2018)*.

The creation, evolution and quenching/merging of SGs are crucial factors of the large-scale evolution of the Universe *(Press & Schechter, 1974).* SGs undergo varied and tumultuous changes in their uncertain lives. Interacting with other satellites, their host CG, the DM halo of the CG, and more. They do not follow the evolution path found for larger central galaxies, and their varied transformations make them a valuable study subject to understand the evolution of galaxies and the Universe as a whole.

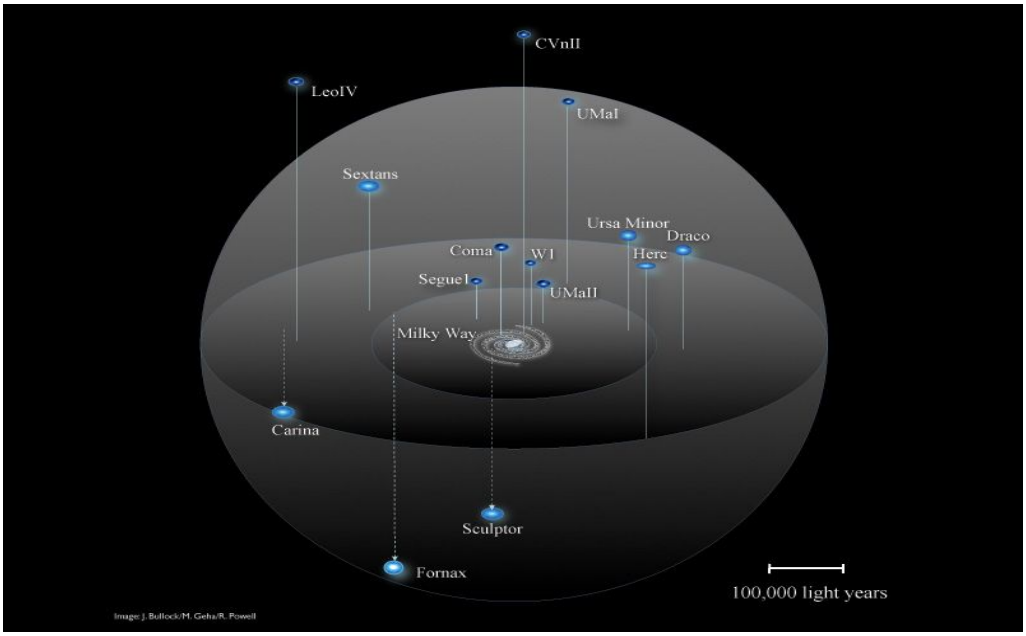



Figure 7: A map of the Milky Way and its observable SGs. *Image: Bullock J., Geha M., Powell R..*

### 1.3.2 - SG Quenching

SGs quenching can be split into 5 distinct evolution paths *(Wuyts et al., 2011; jie Peng te al., 2010; van den Bosch et al., 2008; Woo et al., 2017)*:

**Gas Depletion** - Dwarf SGs that do not attain a large enough mass to prevent early quenching, before entering CG DM Halo. Due to ram pressure forces within the CG DM Halo, they cannot attain further cold-gas, and therefore remain quenched.

**Mass Quenching** - large field galaxies that quenched by mass before entering CG DM Halo, but may still become a satellite of an even larger galaxy.

**Fast Quenching** - As the SG reaches its first peri-center, ram pressure and tidal forces increase, expelling all the gas from the SG. Here, ram pressure is the dominant force. Tidal forces at later orbits begin stripping the stellar mass of the SG.

**Slow Quenching** - In this scenario, ram pressure and tidal forces are small in relation to the self-gravity of the SG, and may also be opposite to each other. This causes compaction of SG gas at the first peri-center, extending the life of the SG. It is not until the 2nd or 3rd peri-center events that the SG quenches, when ram pressure and tidal forces can dominate over self-gravity. Slow Quenching is also characterized by dark-matter reduction in the SG.

**No Quenching** - A small subset of SGs avoid quenching within the DM halo of their respective CG.

**Minor Merger** - The vast majority of SGs orbiting their CG end up merging with it, as their orbit deteriorates over time. Many SGs are rather short lived, as they may merge with the CG before even completing a single orbit.

While Pre-halo quenching SG's are easily identifiable, predicting whether a SG will go through fast or slow quenching, or whether it will survive around its central, is currently an unknown in the field of cosmology. Finding these features can allow us to get a full description and model of the quenching phenomena.

Current predictive models for the prediction of SG evolution are unable to account for most evolution paths.

### 1.3.3 - SG Swansong Events

*Nussbaum, 2018* found that a fraction of SGs in the VELA simulation experience spikes in the sSFR (the rate at which new stars are created in a galaxy) near peri-center around CG. These unusual events were labeled as swansongs, the SGs last burst of stars before being swiftly quenched. This is in accordance with other papers in the field. We have yet to fully understand the physical causes behind this. An example of a swansong galaxy is located in Appendix A.

## 1.4 - VELA_v2 Cosmological Simulation

Galactic evolution takes place over a timeframe of billions of years, which leads to the ultimate problem of observational cosmology: it is impossible to see the evolution in real-time of any single galaxy. In order to solve this issue, physicists make use of cosmological simulations: computer simulations that produce a model of particles making up galaxies - incremented over time in datasets known as snapshots - of parts of the Universe, over a predetermined timeframe.

Cosmological simulations make use of the LCDM model and the basics of what we know about physics to make that happen, and allow us to study the long term evolution paths of galaxies, find correlations and causations, discover new physics phenomena, and further our understanding of the cosmos.

One method of creating a cosmological simulation is the Adaptive Refinement Tree (ART) algorithm *(Kravtsov et al., 1997, Ceverino & Klypin, 2009)*. The ART algorithm is a numerical process that creates a 3-dimensional grid of several large cubes filled with particles. The simulation can increase the resolution constantly by subdividing cubes in specific areas when data needs to be calculated in greater detail. The beauty of the ART algorithm lies in the fact that it allows the computer to save areas without much data at a lower resolution, speeding up computing time, while still retaining great detail in information-dense regions.

The VELA v2 cosmological simulation suite uses an algorithm called an adaptive mesh refinement tree (AMR), a variation of the ART algorithm, which simulates and records the evolution of cosmic structures under various environmental influences. VELA is a suite of 34 simulated environments, each one containing a central galaxy and a number of satellite galaxies, saved as dozens of 'snapshots', each snapshot a slice of recorded particle data from a certain interval in time. The simulation assumes an expanding universe with the standard §CDM cosmology, with the WMAP5 cosmological parameters: $\Omega_\Lambda = 0.73$, $\Omega_m = 0.27$, $\Omega_b = 0.045$, $h = 0.7$ and $\sigma_8 = 0.82$ *(Komatsu et al., 2009)*. VELA simulated galaxies have a redshift range of up to 1, which is half the age of the universes *(Dekel et al., n.d.; Mandelker et al., 2016)*. VELA includes mechanisms for the following physical process: (1) Gravity, (2) Gas Hydrodynamics, (3) Stochastic star-formation, (4) Stellar mass loss, (5) Gas cooling, (6) Photo-ionization heating, (7) Gas recycling and metal enrichment, (8) Supernovae thermal feedback, (9) Radiation pressure feedback.

For this paper, we used a subset of VELA central and satellite galaxies, which had been identified by a Merger Tree *(Nussbaum, 2018*). The result is 34 Central galaxies and 118 satellite galaxies.

Figure 8: VELA v2 - 34 simulated central galaxies imaged using the 2D imaging tool Sunrise. *Image: Simons et al., 2019.* Satellite galaxies are visible as small specks surrounding each galaxy.

From VELA, using various python modules such as YT, we have managed to extract many parameters, and save them into Pandas data frames. This paper builds upon previous catalogs from *Nussbaum, (2018)*.

The VELA suite of 34 simulated galaxies

| Sim id | $R_{vir}$ $[kpc]$ | $M_{vir}$ $[M_\odot]$ | $M_\star$ $[M_\odot]$ | $M_{\rm gas}$ $[M_\odot]$ | SFR $[M_\odot {\rm yr}^{-1}]$ | $R_{{\rm eff},M_\star}$ $[kpc]$ | $cell_{min}$ $[pc]$ | $a_{\rm fin}$ | $z_{\rm fin}$ | #components |
|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 58.25 | 11.2 | 9.31 | 9.17 | 2.64 | 0.93 | 18 | 0.5 | 1.0 | 55 |
| 02 | 54.5 | 11.11 | 9.21 | 9.07 | 1.43 | 1.81 | 36 | 0.5 | 1.0 | 22 |
| 03 | 55.5 | 11.14 | 9.58 | 8.95 | 3.67 | 1.41 | 18 | 0.5 | 1.0 | 63 |
| 04 | 53.5 | 11.09 | 8.91 | 8.9 | 0.45 | 1.73 | 36 | 0.5 | 1.0 | 24 |
| 05 | 44.5 | 10.85 | 8.86 | 8.71 | 0.38 | 1.81 | 18 | 0.5 | 1.0 | 25 |
| 06 | 88.25 | 11.74 | 10.33 | 9.51 | 20.6 | 1.05 | 36 | 0.37 | 1.7 | 131 |
| 07 | 104.25 | 11.96 | 10.76 | 9.9 | 18.14 | 2.85 | 36 | 0.54 | 0.85 | 247 |
| 08 | 70.5 | 11.45 | 9.54 | 9.17 | 5.7 | 0.74 | 36 | 0.57 | 0.75 | 142 |
| 09 | 70.5 | 11.44 | 10.01 | 9.46 | 3.57 | 1.74 | 36 | 0.4 | 1.5 | 137 |
| 10 | 55.25 | 11.12 | 9.78 | 9.11 | 3.2 | 0.46 | 36 | 0.56 | 0.78 | 113 |
| 11 | 69.5 | 11.43 | 9.88 | 9.52 | 8.94 | 2.14 | 36 | 0.46 | 1.17 | 49 |
| 12 | 69.5 | 11.42 | 10.29 | 9.3 | 2.7 | 1.13 | 36 | 0.44 | 1.27 | 47 |
| 13 | 72.5 | 11.5 | 9.76 | 9.55 | 4.48 | 2.48 | 18 | 0.4 | 1.5 | 79 |
| 14 | 76.5 | 11.56 | 10.1 | 9.64 | 23.32 | 0.32 | 36 | 0.41 | 1.44 | 46 |
| 15 | 53.25 | 11.08 | 9.71 | 8.92 | 1.35 | 1.07 | 36 | 0.56 | 0.79 | 34 |
| 16* | 62.75 | 11.7 | 10.61 | 9.7 | 18.47 | 0.61 | 26 | 0.24 | 3.17 | 103 |
| 17* | 105.75 | 12.05 | 10.93 | 10.04 | 61.4 | 1.36 | 34 | 0.31 | 2.23 | 202 |
| 19* | 91.25 | 11.94 | 10.65 | 9.76 | 40.47 | 1.22 | 32 | 0.29 | 2.44 | 94 |
| 20 | 87.5 | 11.73 | 10.56 | 9.55 | 5.55 | 1.72 | 18 | 0.44 | 1.27 | 259 |
| 21 | 92.25 | 11.8 | 10.61 | 9.64 | 7.89 | 1.73 | 18 | 0.5 | 1.0 | 325 |
| 22 | 85.5 | 11.7 | 10.64 | 9.45 | 12.0 | 1.31 | 18 | 0.5 | 1.0 | 60 |
| 23 | 57.0 | 11.17 | 9.88 | 9.12 | 3.06 | 1.16 | 18 | 0.5 | 1.0 | 64 |
| 24 | 70.25 | 11.44 | 9.94 | 9.41 | 3.88 | 1.68 | 18 | 0.48 | 1.08 | 257 |
| 25 | 65.0 | 11.34 | 9.84 | 8.93 | 2.29 | 0.73 | 18 | 0.5 | 1.0 | 349 |
| 26 | 76.75 | 11.55 | 10.2 | 9.44 | 9.36 | 0.74 | 18 | 0.5 | 1.0 | 57 |
| 27 | 75.5 | 11.54 | 9.85 | 9.48 | 6.1 | 1.98 | 18 | 0.5 | 1.0 | 53 |
| 28 | 63.5 | 11.3 | 9.27 | 9.32 | 5.54 | 2.32 | 18 | 0.5 | 1.0 | 42 |
| 29 | 89.25 | 11.72 | 10.36 | 9.55 | 16.82 | 1.89 | 19 | 0.5 | 1.0 | 335 |
| 30 | 73.25 | 11.49 | 10.2 | 9.37 | 2.97 | 1.43 | 18 | 0.34 | 1.94 | 296 |
| 31* | 38.5 | 11.37 | 9.89 | 9.1 | 15.27 | 0.43 | 21 | 0.19 | 4.26 | 37 |
| 32 | 90.5 | 11.77 | 10.42 | 9.64 | 14.86 | 2.58 | 18 | 0.33 | 2.03 | 199 |
| 33 | 101.25 | 11.92 | 10.68 | 9.7 | 32.68 | 1.23 | 18 | 0.39 | 1.56 | 117 |
| 34 | 86.5 | 11.72 | 10.19 | 9.66 | 14.47 | 1.84 | 18 | 0.35 | 1.86 | 248 |
| 35* | 44.5 | 11.35 | 9.75 | 9.39 | 22.93 | 0.33 | 24 | 0.22 | 3.54 | 92 |

Figure 9: VELA Simulation Major Quantities Table: The virial radius - $R_{vir}$, total virial mass - $M_{vir}$, stellar mass - $M_*$, gas mass - $M_{gas}$, star-formation rate - $SFR$ and half stellar mass-radius - $R_{eff,M_*}$, for the 34 VELA simulations. When $M_{vir}, M_*, M_{gas}$, $SFR$ and $R_{eff,M_*}$ are quoted within 0.1 $R_{vir}$ and all of the masses $M_{vir}, M_*, M_{gas}$ are shown in their $log_{10}$ values. Also listed are the minimum cell size $cell_{min}$ in the snapshot, the final simulation scale factor - $a_{fin}$, and redshift - $z_{fin}$, as well as the number of stellar components throughout the simulation $\#_{components}$. All physical properties and $cell_{min}$ are quoted at $z = 2$, except for the five cases marked $^*$, where they are quoted as the final simulation output, $z_{fin} > 2$. Looking at $\#_{components}$, there are numerous galaxies in the simulation with a diverse population of the central halos in its mass range.

# 2 - Machine Learning

The field of Machine Learning (ML) focuses on the use of computer algorithms created to perform specific tasks, as derived from sample data *(Mitchell, 1997)*. ML algorithms are first taught using a set of *training* data, and then tested using a separate set of *testing* data. Machine Learning is heavily intertwined with statistics and probability theory. ML algorithms are most commonly divided into two main approaches:

**Supervised Learning**: A supervised learning algorithm is provided with a matrix of training data, consisting of a feature vector that represents each training sample's data input, as well as the sample's associated label/additional dependent feature value *(Alpaydin, 2010)*. Supervised learning may be used for tasks such as *(Alpaydin, 2010)*:

*Classification*: A pattern recognition algorithm that attempts to predict a discrete feature label of a new sample. In this case, the label serves as a grouping mechanism to separate individual samples into classes. An example of classification would be classifying an animal as a dog vs cat, based on labeled training data, for each picture if it is a cat or a dog.

*Regression*: A algorithm that searches for a relationship between the dependent feature and the independent feature vector. The dependent feature is a continuous variable. One use of regression is stock price prediction based on past trends.

**Unsupervised Learning:**

Unsupervised learning algorithms are provided with the same matrix of training data as in supervised learning, but without the associated label.

Unsupervised learning may be used for *(Alpaydin, 2010)*:

*Cluster Analysis* - The task of searching for homogenous groupings in a given dataset. Used for pattern recognition, information retrieval, etc. Cluster analysis is the unsupervised equivalent of classification. An example of cluster analysis would be training an algorithm to group pictures of cats and dogs into two classes, without being told which is which.

*Anomaly detection* - The identification of outlier events or observations in a given dataset. A use of anomaly detection could be to identify anomalous occurrences in a server's traffic data, which could be a dangerous intrusion by a third party.

## 2.1 - Decision Trees and Random Forests

Decision trees (DTs) are a statistical supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple conditional control statements (decision rules) inferred from the data sample. DTs have a flow chart structure, used to determine probability of an outcome *(Rokach & Maimon, 2005)*. Decision trees split data into groupings called leaf nodes based on the homogeneity parameter *purity* of the groupings *(Quinlan, 1986)*.

For example, if our dataset consists of 20 samples, 10 negative and 10 positive. Each sample has other independent features. The goal of the DT (in classification context) is to split the data based on a decision rule (feature value => x, True/False), to get two leaf nodes, each containing only samples from one class.

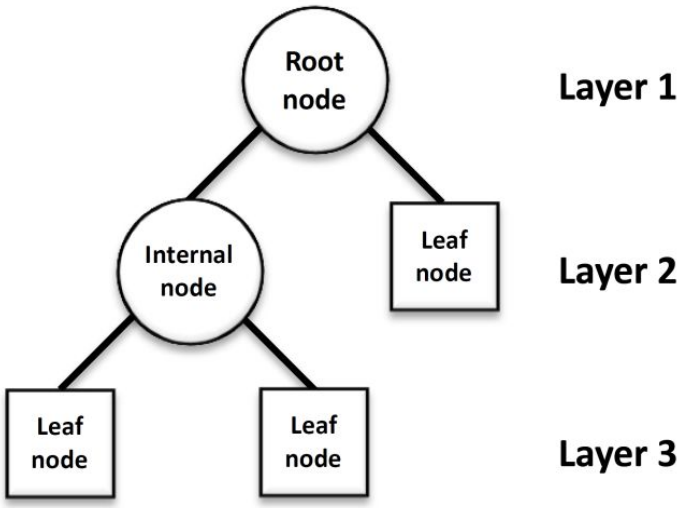



Figure 10: A decision tree flow chart example. *Image: Mei-Hung Chiu et al., 2016.* The root node is the starting layer 1, where the first split to the sample data occurs. A node can split into further internal nodes, or into a final leaf node. Splits are chosen based on increase in purity of grouping in at least one leaf or internal node.

A decision tree is made up of a root node and sequence of internal nodes, with leaf nodes being specific outcomes *(Quinlan, 1985)*. The purity criterion of a decision tree often used is the *Gini criterion ((Rokach & Maimon, 2005; Breiman et al., 1984)* , defined as:

$$Gini = 1 - \sum_{j} p_j^2$$

Where $p_j$ is the probability of element $j$ being classified for a distinct class.

Decision trees are often used in classification tasks, as they are inexpensive computationally to train. However, single decision trees are often not enough for highly convoluted data, and they are prone to overfitting - a problem where the decision tree is highly accurate when tested on the training data, but makes mistakes when testing on new data.

A Random Forest (RF) algorithm is a type of ensemble ML model consisting of many differently-structured decision trees *(Breiman, 2001)*. The final prediction of a RF model is based on a voting system by each decision tree: the class with the most votes is the final prediction.

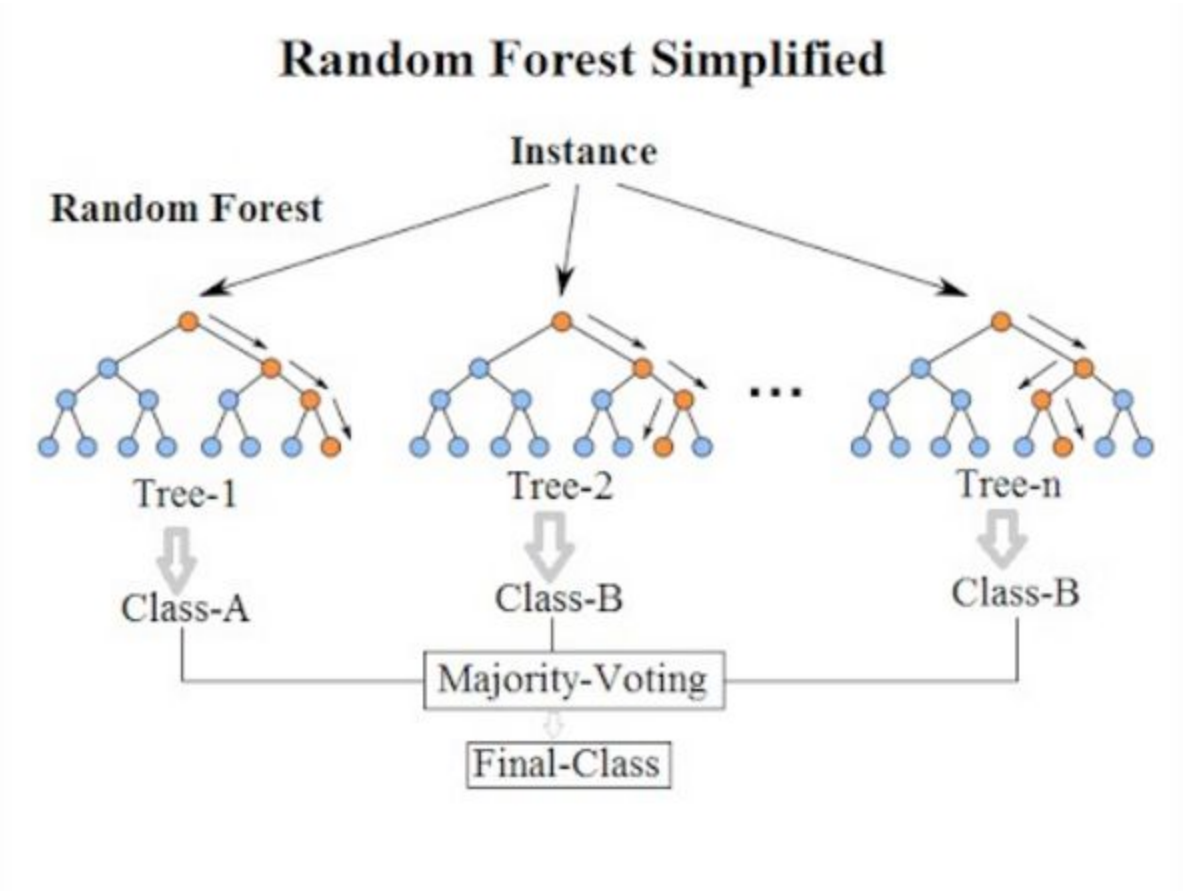



Figure 11: A Random Forest visualization. *Image: Dimitriadis Stavros et al., 2013.* Each tree is structured differently, and therefore results in a different vote (result). The final result is based on the majority vote of the random forest.

## 2.2 - ML Evaluation for Classification Algorithms

How does one evaluate a ML Classification model? There are a range of metrics that we can use as data scientists.

*Classification Accuracy*:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ all\ predictions}$$

Classification accuracy is a pure percentage showing the correct prediction out of all predictions made by the model. While applicable in some cases, datasets containing 'Class imbalances' (i.e, Significantly more instances of Class A over Class B), a high accuracy can be attained even by having the algorithm guess Class A every time.

We can further separate predictions into a *confusion matrix*, a 2 dimensional table of the model's predictions against the actual label.

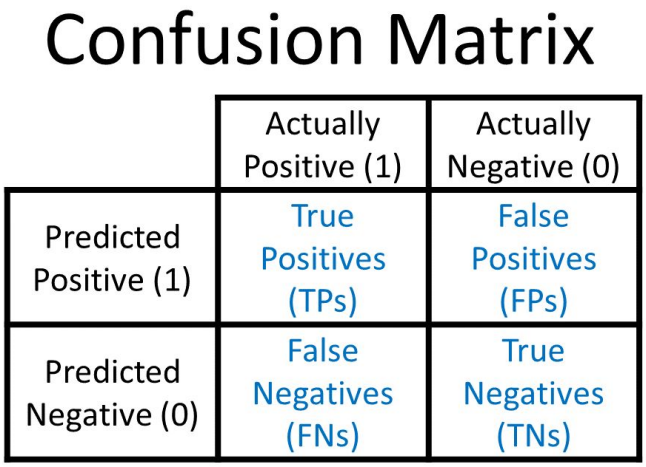



Figure 12: 2D Confusion Matrix, consisting of *Positive* and *Negative* class labels. On the top row, we see Actual labels in testing dataset (not provided to the ML), on the side, we see the predicted labels by the ML model. *Image: Draelos R., 2019.*

Using the values from the confusion matrix, we can attain the following evaluation metrics *(Powers, 2011)*:

*Precision*: A measurement of positive (1) prediction accuracy; or the first row of the confusion matrix:

1) $$Precision = \frac{TP}{TP + FP}$$

*Recall*: A measurement of positive (1) prediction accuracy when looking at the *actual* (1) column. Recall is also known as True Positive Rate (TPR):

2) $$Recall = \frac{TP}{TP + FN}$$

In order to increase precision, FP rate has to drop, which could be done by lowering the probability that a model predicts a sample to be *positive*. In order to increase recall, the

opposite has to happen; the model must predict all actual *positive* samples as *positive*: it must raise the probability of *positive* being the predicted label. In other words, precision and recall often have opposite correlation: higher precision = lower recall, and vice versa.
To offset this, we can use an *F1 score*:

3) $F1\ score\ =\ 2\frac{Precision * Recall}{Precision + Recall}$

The F1 score takes into account both FPs and FNs, making it useful for class imbalances within a dataset. For this paper we targeted an $F1\ score \geq 0.9$ for all classes.

It is important to note that confusion matrices can have more than 2 classes, and each class will have its own precision, recall, and f1 score.

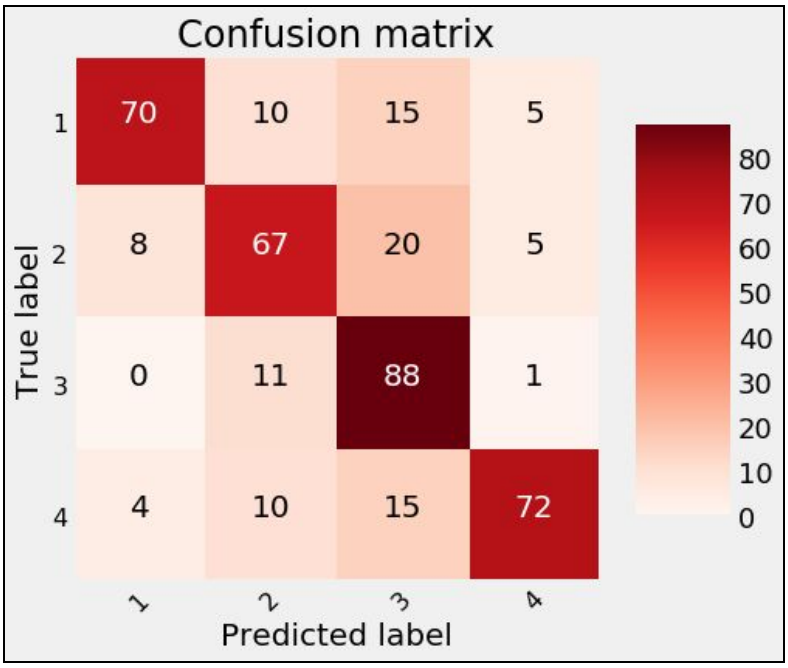



Figure 13: Confusion Matrix with multiple classes. Cell intersections between class names are true positives for each class. Anything within a single predicted row that is not a TP can be classified as FP, and along the column as FN.

# 3 - Methods

For the purposes of this research paper, we expanded the SG dataset, providing new features and calculations. In addition, we detail here the ML methods used during the research.

## 3.1 - Tools

*Astro Cluster* - The Astrophysics Cluster at the Hebrew University of Jerusalem is a computer networking, analysis, parallel computing, and data storage server, consisting of 10 analysis nodes, 96 computation nodes, 2 gpu nodes and a main storage server. We connect to the Astro Cluster in order to access the VELA_v2 simulation data, do data preprocessing and additional calculations, and run machine learning algorithms.

- *Python*: A programming language with significant data computation and data science capabilities. Python has extensive libraries of prebuilt functions designed for a variety of purposes. We will be using the following libraries:

  - *NumPy*: A library for efficient vector and matrix calculations in python.
  - *pandas*: A library offering tools for creating and handling data sets, tables, and time series' (a collection of data points indexed over time).
  - *Seaborn and Matplotlib*: Graphing and plotting libraries.
  - *YT*: the YT library is used for the efficient and easy processing of simulation data. It is an effective tool for processing cosmological simulations and is widely used by researchers *(Turk, 2010)* for analyzing and making calculations on volumetric particle data.
  - *Sci-kit learn*: A library built for Machine Learning in Python. It features algorithms for purposes such as Classification, Regression, Clustering, Model Selection, Data Pre-processing, and more.
  - *Graphviz*, *DTreeViz*: Python libraries built on top of the *DOT* language for graph creation. DTreeViz is a specialized library using graphviz, specifically for visualizing decision trees out of sk-learn.
  - *VIVID*: A library built for 3D visualization of particle simulations in python. Includes several useful features, including ellipsoids, arrow vectors, and animations through time.

- *IPython:* Interactive-Python is a command shell for operating with Python, allowing for cloud-based in-browser coding, support for data-visualization, interpreters, and used for scientific research. This gives us the use of the Astro Cluster.

## 3.2 - SG Dataset from VELA

In order to understand the influential factors on quenching of SGs, we first have to select and collect the data from relevant SGs. Filtering by SGs that have been identified in more than 10 snapshots, with at least 5 in the halo of CG, and have at least one pericenter and apocenter, we get our selected 118 long living SGs.

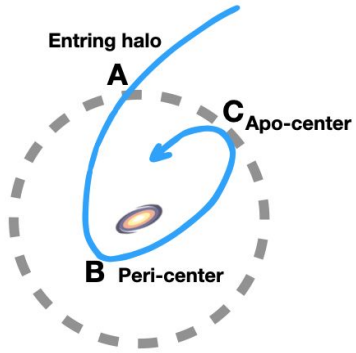



Figure 14: A sketch of an SGs orbit around a CG. 3 important points are represented. A(Entering halo) - the first point at which an SG enters the virial DM halo of the CG. B(peri-center(1)) - the first peri-center or nearest approach along the first orbit around the CG. C(apo-center(1)) - the first apo-center or furthest point along the first orbit around the CG, before continuing onto the next orbit.

With such a small sample size, we chose to expand our dataset by taking SG data from 3 snapshots, and defining each as a separate sample.

1. **Outside_Rvir** = Snapshot before Entering_Rvir_Outside
2. **Entering_Rvir_Outside** = The last snapshot before SG appears inside Halo of CG
3. **Entering_Rvir_Inside** = First snapshot where SG appears within Halo of CG

Using this method, we triple our sample size, with a modest decrease in variance and increase to model robustness due to the inherent fluctuation within the SG life. This dataset we will use for ML training and testing on SG quenching and starburst event prediction.

### 3.2.1 - Classifying Quenching points of SG's

We scripted an algorithm to search when $sSFR(0.5kpc)$ of an SG was less than $10^{-4}\ [1/Gyr]$ for a period of three consecutive snapshots or longer, in which case we would consider the first snapshot in this bin as the quenching moment. In certain scenarios, such as a galaxy with 2 consecutive quenched snapshots also being the last 2 of the satellite galaxy in the catalogue, we would consider the second to last snapshot as the quenching moment.

We then compared each satellite galaxy's quenching moment against its first peri-center event, and assigned each galaxy a quenching label - Qlabel.

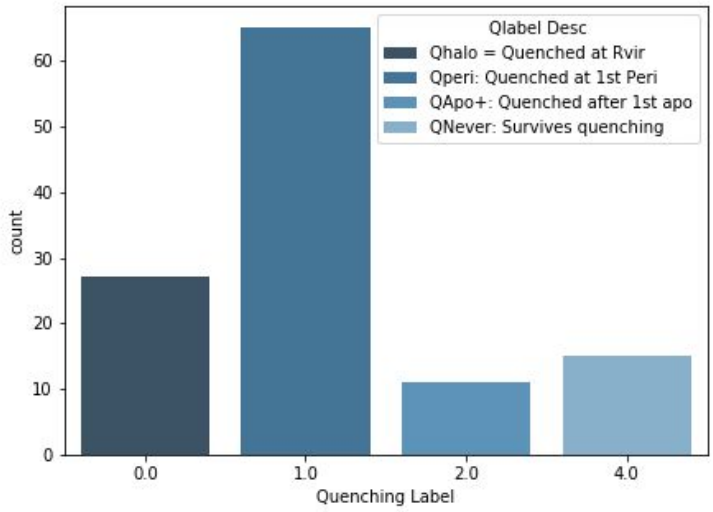



Figure 15: Histogram of SG Qlabels. We can see here that of the 118 SG's, the majority 55% are Q1 quenched at first pericenter galaxies.

As found in *Nussbaum, (2018)*, long living satellite galaxies are grouped into these 4 classes:

- **Qhalo**: these are long living SG's that due to known factors quench on their own, before encountering a CG.
- **QPeri**: These SG's are characterized as fast-quenching galaxies, where due to sudden tidal forces from the CG at the first pericenter event of the SG's orbit, and ram pressure forces while moving through the Central DM Halo, the SG is stripped of its cold gas.
- **QApo+**: These SG's are found to be slow-quenching galaxies, and the quenching event occurs during one of the consecutive pericenters. Slow-quenching is characterized additionally by dark-matter reduction in the SG.
- **QNever**: These SG's survive despite the forces on them by the host CG, managing to continue growing and retain sSFR throughout its lifespan.

Populations:
*QHalo*: Quenched at entering halo = 27 SGs
*QPeri*: Quenched at 1st peri-center = 65 SGs
*QApo+*: Quenched after 1st apo-center = 11 SGs
*QNever*: Survives quenching = 15 SGs

### 3.2.2 - Swansong Events

Similarly to 3.2.1, we created an algorithm to search for a rise in $sSFR(0.5kpc)$ of the SG, after 1 or more consecutive snapshots of $sSFR(0.5kpc) <= 10^{-4}\ [1/Gyr]$ (quenching attempt), and which were followed by another 3 consecutive snapshots of $sSFR(0.5kpc) <= 10^{-4}\ [1/Gyr]$ (full quenching). Filtered by $sSFR(0.5kpc)$ spikes within 1 snapshot of peri-center.

We found the following populations:
*Swansong Events* = 11 SGs
*Non-Swansong* = 107 SGs

## 3.3 - Feature Engineering and preprocessing

As shown by *Nussbaum, 2018*, the current model of SG quenching cannot explain the range of quenching behavior in SGs. In addition, research on the causes of starburst events at peri-centers is unclear. As such, we built up a new catalog of features on the VELA simulation.

### 3.3.1 - Orbital Eccentricity

Satellite galaxies follow decaying elliptical orbits, as they form outside the halo of their host CG and travel under the influence of the CG gravitational force and dynamical friction. The point in the elliptical orbit nearest to the CG - which exists as one of the ellipses focuses - is called the *peri-center*. Similarly, the furthest point in the orbit is the *apo-center*.

Eccentricity *e* is a dimensionless measure of an orbit's irregularity from circularity. The value is derived from the following equation, where $r_a$ is the distance at the apocenter, and $r_p$ is the distance at the peri-center. Eccentricity was calculated using the first peri-and-apo centers.

$$e = \frac{r_a - r_p}{r_a + r_p} = 1 - \frac{2}{\frac{r_a}{r_p} + 1}$$

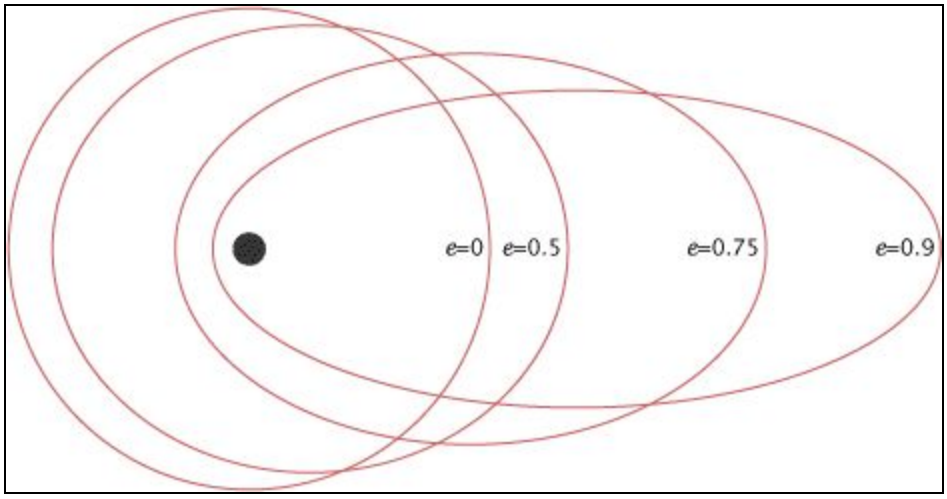



Figure 16: Change in eccentricity on elliptical orbits. *Image: Simon R., NASA Earth Observatory (2009).* The larger the eccentricity, the more elongated the orbit. *e=0* is a completely circular orbit, e=1 would be an escape parabolic orbit.

### 3.3.2 - SG Dynamical Effects

As an SG orbits around its CG, there are 4 dominant dynamical forces acting on the SG, As taken from *Nussbaum, 2018.*

**Ram Pressure** - The pressure exerted on the SG as it moves through the gas and DM of the CG halo. Ram pressure calculated as described at *(Simpson et al., 2018; Gunn & J. Richard, 1972)*:

$$1)\ f_{ram}(r) = \frac{P_{ram}}{\Sigma coldgas(r)} \cdot \hat{v} = \frac{P_{gas,\,environment} \cdot V^2_{orbit}}{\Sigma coldgas(r)} \cdot \hat{v}$$

$p_{gas,\,environment}$ is the gas in front of the satellite when $p_{gas,\,environment} = \frac{\Delta M_{gas}}{\Delta V}$ of a shell in the halo of the central galaxy between $R_{orbit} \pm R_{sat}$.

**Tidal Forces** - The gravitational force inflicted on the SG by the CG halo gravitational potential well. Tidal force per one solar mass is defined as *(Dekel, Devor, & Hetzroni, 2003)*:

$$2)\ f_{tidal}(r) = (\alpha - 1)\ \frac{GM_{(cen,\,R_{orbit})}}{R_{orbit}^3} \cdot \bar{r}$$

Where $M_{(cen,\,R_{orbit})}$ is the total mass encapsulated in a sphere with radius $R_{orbit}$ around the CG center, $R_{orbit}$ is the distance between the SG and CG, and $r$ is a chosen radius of the satellite galaxy. $\alpha$ is measured as the average density slope of the CG halo: $\alpha(r) = -\frac{dIn\,\bar{p}}{dIn\,r}$ ; $\bar{p} = \frac{M_{cen(r)}}{\frac{4\pi}{3}r^3}$ and $G$ is the gravitational constant.

**Self-Gravity** - The self-gravity of the SG itself. Calculated using Newton's gravitational force as:

$$3)\ f_{self-gravity}(r) = \frac{GM_{(sat,\,r)}}{r^2} \cdot \overline{\hat{r}}$$

Where $r$ is the radial distance to the center of the satellite center, $M_{(sat,r)}$ is the total mass of the satellite in a $r$ sphere, and $G$ is the gravitational constant.

**Dynamical Friction** - SGs experience a drag force known as dynamical friction as they travel through the DM halo of the CG. This dynamical friction causes orbital decay, which we expressed through the eccentricity feature.

### 3.3.3 - Galactic ShapeTensor

From the previous research done, including recently created visualizations of SGs using VIVID, we hypothesized that an important missing feature was the shape of the galaxy.

-

We modified an iterative algorithm developed by and used in *Tomassetti et al. 2016* to calculate the structure of the SG. The spheroid, used to define the shape of the galaxy, is a 3-dimensional object containing 3 radii along each axis of the spheroid: a,b,c.

The shape tensor of the spheroid is therefore defined as:

$$S_{i,j} = \frac{1}{M}\sum_k m_k (r_k)_i (r_k)_j \qquad \mathcal{S} = \frac{1}{\alpha}\begin{bmatrix} a^2 & 0 & 0 \\ 0 & b^2 & 0 \\ 0 & 0 & c^2 \end{bmatrix}$$

Where $m_k$ is the mass of the k-th particle, $(r_k)_{i,j}$ are the distances from the center of SG along the axes i,j, and M is the total mass. The eigenvalues of S are proportional to the squares of the radii (a > b > c) of the ellipsoid that describes the spatial distribution of the particles that

constitute the system, and corresponding eigenvectors mark the orientations of these principal axes. The final shape of an ellipsoid can be described with $p = b/a,\ q = c/b$.

The algorithm begins with spherical ellipsoid a=b=c=R, then iteratively calculates new S for all particles within the previous iteration ellipsoid, rescales, and rotates the ellipsoid such that a=R, repeating until the difference between ellipsoid iterations $tol \leq 0.05$.

$$tol = max\left(\frac{|p_{old} - p_{new}|}{p_{old} + p_{new}}, \frac{|q_{old} - q_{new}|}{q_{old} + q_{new}}\right)$$ [Where p=b/a, q=c/b.]

If the total amount of particles in the ellipsoid drops to <50, or the number of iterations gets over 15, the ellipsoid can't converge, so we select the latest possible shape.

We ran the algorithm on combinations of radii and matter-type from the following:

- By SG radii: $R_{sat}[kpc],\ 1[kpc],\ 0.5[kpc],\ R_{0.9coldgas}[kpc]$. Values taken from *(Nussbaum, 2018)*. $R_{sat}$ being the spherical radius encapsulating 90% of the mass of the SG stellar component, in line with *Tweed et al., 2009; More et al. 2011*. $R_{0.9coldgas}$ is similarly defined as the spherical radius containing 90% of the cold gas mass within the stellar radius $R_{sat}$.
- By matter type: Gas, Stars, Dark Matter (DM)

The ShapeTensor algorithm was successful in identifying the majority (>99%) of SG matter-radius combinations where there were enough particles to attempt a calculation.

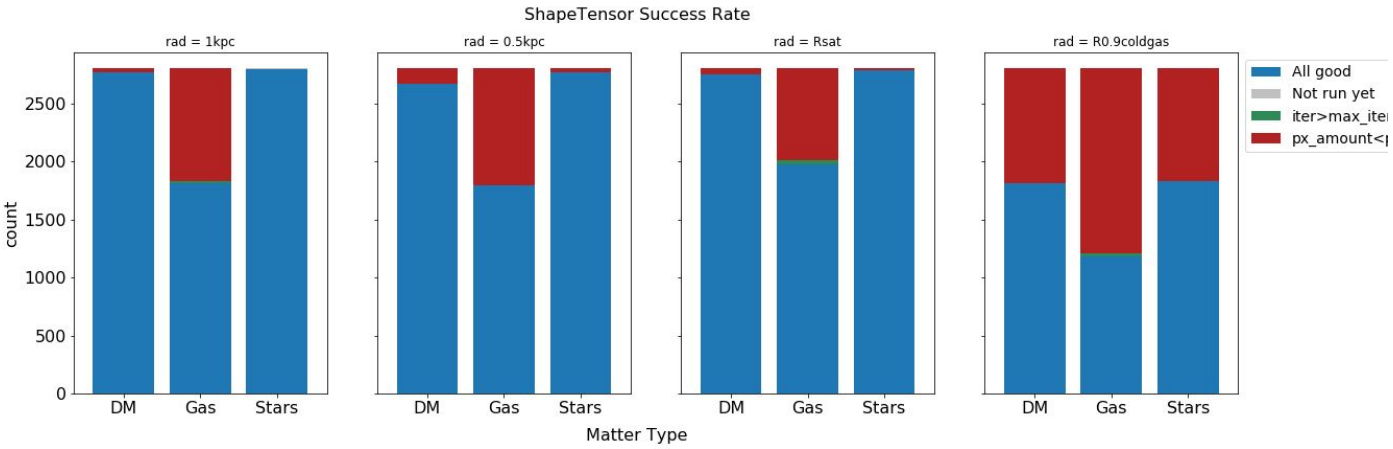



Figure 17: Histogram plot showing a success rate of Shape Tensor algorithm over the SG catalog (all snapshots). SG Snapshots where the ShapeTensor fails are those where there is a lack of any gas particles, which also explains the high amount of shapes missing from *rad=0.9Coldgas[kpc]* combinations. We noted a minority of outliers within the Gas-Rsat combination that failed due to *max_iter* being reached. Further investigation using VIVID found these to be complex and unordered gas distributions due to dynamical effects surrounding the SG.

### 3.3.4 - Intersection of Plane and Ellipsoid

We hypothesized that a possible influential feature in the quenching of SGs was the placement of its gas and star mass relative to a plane dividing the center of the SG. The more gas and stars that trailed behind, the higher chance the galaxy would get quenched by the first pericenter. To calculate this, we first took the general equation for an ellipsoid:

$$1)\ \frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z2}{c^2}=1$$

Where $a,b,c$ are the radii of the ellipsoid, $x,y,z$ are the eigenvectors of the radii.

A plane can be defined with the position vectors of a point and a normalized vector passing through the point, for which we used the center of the SG and its velocity.

$\widehat{n_x}=\widehat{v}\cdot\widehat{x}$ , $\widehat{n_y}=\widehat{v}\cdot\widehat{y}$ , $\widehat{n_z}=\widehat{v}\cdot\widehat{z}$ Where $\widehat{v}$ = normalized velocity of SG, $\widehat{x},\widehat{y},\widehat{z}$ = eigenvectors.

$\widehat{n_x}+\widehat{n_y}+\widehat{n_z}=0$ Is the defining equation for the plane.

We can therefore define the area of the intersection of the center plane and the ellipsoid shape as *(Ferguson, 1979)*:

$$3)Ellips_{area}=\pi\frac{abc}{\sqrt{a^2\widehat{n_x}^2+b^2\widehat{n_y}^2+c^2\widehat{n_z}^2}}$$

To find the relative material behind the ellipsoid is a simple task of taking the volume of the ellipsoid and divide by the area of the plane intersection:

$$4)\ Volume_{ell}=\frac{4}{3}abc$$

$$5)\ Ellips_{depth}=\frac{Volume}{Area}=\frac{4}{3}\sqrt{a^2\widehat{n_x}^2+b^2\widehat{n_y}^2+c^2\widehat{n_z}^2}$$

We calculated $Ellips_{area}$ and $Ellips_{depth}$ for gas and stars over 3 radii: $0.5kpc,\ R_{0.9coldgas},R_{sat}$ .

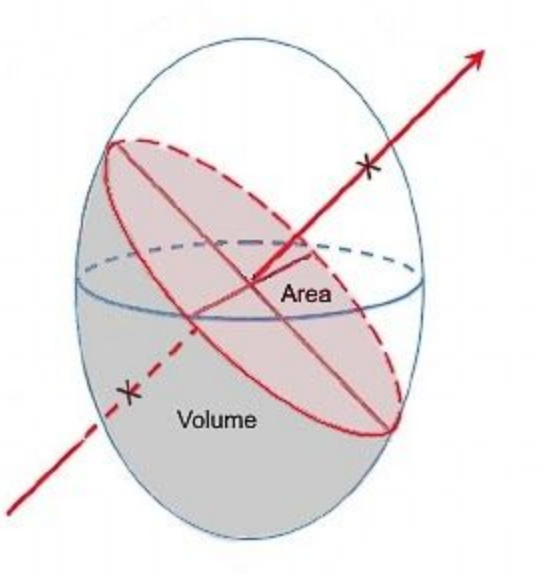



Figure 18: Ellipsoid with plane intersection and volume.

### 3.3.5 - Angular Momentum

To calculate Angular Momentum (AM), we first create a spherical mask around SG and select particles within a maximal distance R to the point of origin of the sphere, then compute AM using the formula:

$$AM = \sum_i (R_i \times m_i V_i)$$

Where $R_i$ is the distance of the particle to the center of origin, $m_i$ is the mass of the particle, and $V_i$ is the velocity of the particle relative to the origin.

AM was computed on combinations of radii and matter-type, as in 3.3.3.

### 3.3.6 - Special Features

Building on current catalogs, we created compound features to show relationships between certain parameters to the Qlabel. These are:

**Proportionate features:**

1 ) Ratio of galactic radii:

$R_{sat}/R_{vir}$ - (Radius of the SG, Radius of CG)

2 ) Distance at peri-center as a function of CG radius:

$R_{peri_1}/R_{vir}$ - (Distance on the first peri-center event, Radius of CG)

4) Ratio of galactic masses:

$M_{sat}/M_{rvir}$ - (total Mass of SG, total Mass of CG)

4) Angle in radians between Velocity of SG and its point of entrance at the halo...

$cos^{-1}(\frac{V_{sat} \bullet R_{orbit}}{|V_{sat}||R_{orbit}|})$ - (Velocity of SG, position vector at the radius of CG.)

**Dot Product Features of SG:**

5) The distribution of SG velocity along with the short axis C.

$gas_c Vec_{R_{0.9coldgas}} \bullet V_{sat}$ - (Eigenvector along C axis of the gas ellipsoid at R=0.9coldgas, The velocity of SG)

6) The angle between the SG velocity vector and its angular momentum.

$V_{sat} \bullet AM^{Stars}_{Rsat/0.5kpc/R_{0.9coldgas}}$ - (Velocity of SG, Angular momentum of Stars at specified radii)

7) The distribution of SG angular momentum along with the short axis C.

$AM^{Stars}_{Rsat/0.5kpc/R_{0.9coldgas}} \bullet gas_c Vec_{R_{0.9coldgas}}$

8) The angle between SG angular momentum of Stars and DM.

$AM^{Stars}_{Rsat} \bullet AM^{DM}_{Rsat}$ - (Angular momentum of Stars and DM at Rsat)

**Force Features:**

As taken from Nussbaum, 2018, the following forces ratios were added:

9) Ratio of ram pressure versus tidal force against on SG:

$f_{ram}/f_{tidal}$

10) Ratio of ram pressure acting on SG against SG self-gravity:

$f_{ram}/f_{self-gravity}$

11) Ratio of tidal force acting on SG against SG self-gravity:

$f_{tidal}/f_{self-gravity}$

A list of all features in the catalog is available in Appendix B.

## 3.4 - Machine Learning

Our goal in machine learning was using ML algorithms for feature selection, to then create an interpretable accurate predictive model. For the first step, we chose to use a random forest classifier RFC with highly randomized trees, fed through a recursive feature elimination RFE algorithm. For the final predictive model, a decision tree classifier DT was our choice, for its interpretability and minimal accuracy reduction (assuming strong features).

### 3.4.1 - Random Forest Classifier

RFC's and other ensemble learning methods have several capabilities that make them more powerful and precise than many other ML algorithms, such as:

*Bootstrapping* and *Feature Randomness*: Decision trees can be structured very differently based on slight changes in training data. Bootstrapping and feature randomness is the process of giving different trees random subsets of training data and features respectively, where the number of samples and features given to each tree is less than the total amount *(Breiman, 2001)*. Bootstrapping results in decision trees with higher variance than a single decision tree, but because an RFC aggregates (averages) the decision trees, the total model now has lower variance. This is especially useful for small finite datasets, where significant errors could be made when using just a single decision tree. Feature randomness results in lower bias. Together, this makes the RFC a very accurate and powerful ML model.

For an imbalanced classification problem, we can use class weights, where the weights are inversely proportional to the class frequency: $W_i = \frac{N_{samples}}{N_i * N_{samples_i}}$, where $W_i$ is the weight per class *i*, $N_{samples}$ is the total number of samples in the dataset, $N_i$ is the number of classes, and $N_{samples_i}$ is the number of samples per class *i*.

The use of multiple trees and bootstrapping also prevents over/under-fitting, as the probability of class imbalances converges on the actual class weights, and extreme trees that consistently guess incorrectly are thrown out.

We used the `random_forest_classifier()` class in sk-learn. We ran with bootstrapping and feature randomness enabled, where `n_features` per tree is limited to $\sqrt{N_{features}}$. For the purity criterion, we used *Gini impurity***.** We limited the `max_depth` of each DT to $N_{features}/10$ levels and

`n_trees` to 100. We additionally fed class weights to the classifier using the `class_weights` parameter, setting it to *balanced*. The remaining parameters were left as default sk-learn values.

We split the dataset into training and testing sets using *train_test_split*. We chose a train/test split of 73/27, slightly above the convention of 75/25. We chose this split due to wanting to avoid drop in evaluation based on outliers in our data, as a 2% increase in testing data will show a more accurate picture of the evaluation, without hurting training data (as a 66/33 split would). We enabled stratified splitting to keep the classes split equally between training and testing datasets.

```
#Extracting class labels, establishing train/test samples
labels = np.array(data.pop('Qlabel'))
X_train, X_test, y_train, y_test = skl.train_test_split(data, test_size=0.27,
                                                        stratify=labels)


#Defining, fitting RF classifier
rfc = random_forest_classifier(X_train, y_train, n_trees=100, n_features='sqrt',
                               criterion='gini',bootstrap=True,
                          max_depth=len(features)/10,
                               class_weights = 'balanced')
rfc.fit(X_train,y_train)
```

### 3.4.2 - Decision Tree Classifier

The final model was to be trained with a DT classifier, for interpretability purposes.

We used the `DecisionTreeClassifier()` class in sk-learn. For purity, we again used *Gini impurity*. We used the sk-learn default `splitter` of *best*, which searches for the optimal split in all provided features based on the maximal drop in gini impurity. We limited the `max_depth` to $N_{features}$ levels and set `min_samples_leaf` (the minimum samples required in each leaf when splitting a node) to 2. We set `class_weight` to *balanced.* The remaining parameters were left as default sk-learn values. The use of RFC and DT is in line with *Turk et al., (2015).*

We used `train_test_split` again as before and set.

```
#Defining, fitting DT classifier
dt = DecisionTreeClassifier(criterion='gini',splitter='best',max_depth=len(features),
                                      min_samples_leaf=2,class_weight='balanced')
dt.fit(X_train,y_train)
```

### 3.4.3 - Feature Selection

For feature selection, we used a two-step process:

1) Using Pearson correlation values between features with *df.corr(),* we eliminated highly collinear features ($r > 0.8$), significantly increasing the efficiency of training and increasing the accuracy of the final model.

Pearson correlation coefficient $r$ defined as

$$r_{x,y} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x^2})}\sqrt{(\Sigma y_i^2 - n\bar{y^2})}}$$

Where $n$ is the sample size, $x_i, y_i$ are the feature values per point $i$, $\bar{x}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $y_i$ is the arithmetic mean.

```
corr_matrix = df.replace([np.inf, -np.inf], np.nan).dropna().corr().abs()
# Select upper triangle of correlation matrix
corr_triangle= corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))
# Find feature columns with correlation greater than 0.8
to_drop = [column for column in corr_triangle.columns if any(corr_triangle[column] > 0.8)]
#Final dataset, dropping selected features
data = df.drop(to_drop,axis=1)
```

2) We used `sklearn.feature_selection.RFECV()` (Recursive feature elimination with cross-validation), an algorithm that selects features by recursively considering smaller and smaller sets of features, where the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the optimal or minimum number of features to select is eventually reached. RFECV incorporates cross-validation for tuning optimal number of features, by testing the selected scorer on partitioned data.

We use a recall weighted score as the metric for feature effectiveness. Based on Nussbaum, 2018, we selected a minimum of 3 features to select. We used the sk-learn standard of 5 KFolds for cross-validation. The remaining parameters were left as default sk-learn values.

```
#Defining and fitting RFECV
rfecv = RFECV(estimator=rfc, min_features_to_select=3, step=1,
                                cv=StratifiedKFold(5),scoring='recall')
rfecv.fit(X_train, y_train)

#Returns bool mask to apply on features
if len(rfecv.support_()) <= max_features:
       chosen_features_mask = rfecv.support_()
else:
       chosen_features_mask = rfecv.support_()[:max_features]
```

If RFECV provides a greater number of chosen features than we expect in the final model, we select the $N_{max\_features}$ top features. We defined $N_{max\_features} = 10$, as any more and the model would not be interpretable to our standards. The final DT model is retrained using the chosen features.

Additionally, while creating a baseline using non-correlating features, we calculated absolute correlation as a mean of feature correlations between each other:

$$\prod_{L_i, L_j \in Feature,\, i<j}$$

$Pearson(L_i, L_j)$, while $L_i, L_j$ are labels in the feature list and $i, j$ are their equivalent indexes.

### 3.4.4- ML Evaluation

We used `sklearn.metrics.confusion_matrix()` and matplotlib to plot a confusion matrix of the fitted DT.

```
#Testing classifier (DT, RF) prediction
pred = clf.predict(X_test)
cm = confusion_matrix(y_test, pred)
```

Using `sklearn.metrics.classification_report()`, we extracted the evaluation scores of the DT.

```
report = metrics.classification_report(y_test, pred, output_dict=True)
```

To visualize the DT, we used `dtreeviz()` from DTreeViz. Additionally, we can feed a single sample *X* to view it's decision path.

```
viz = dtreeviz(dt,X_train,labels,target_name='Qlabel',class_names=['QPeri','QNever'],
      feature_names=X_train.keys().values, X = X_test.iloc[0])
```

# 4 - Results

We present an analysis of the results of our feature engineering and ML model fitting stages, including a comprehensive report on the final SG quenching model, and shape tensor findings.

As part of the research, we analyzed several visualizations of galaxy systems in the VELA Suite, created using VIVID, a new 3D visualization tool for use with particle simulation data. The models are available on our SketchFab:

https://sketchfab.com/tomer.nussbaum/models

## 4.1 - Shape Tensor

In this chapter, we analyze the results of the ShapeTensor algorithm, primarily on the shapes found at Entering halo.

### 4.1.1 - Shapes at Halo

We can use a scatter plot showing a P vs Q relationship for each SG at different stages of its life. Firstly, looking at the shapes of Gas at $R_{0.9coldgas}$ at the entrance to the halo:

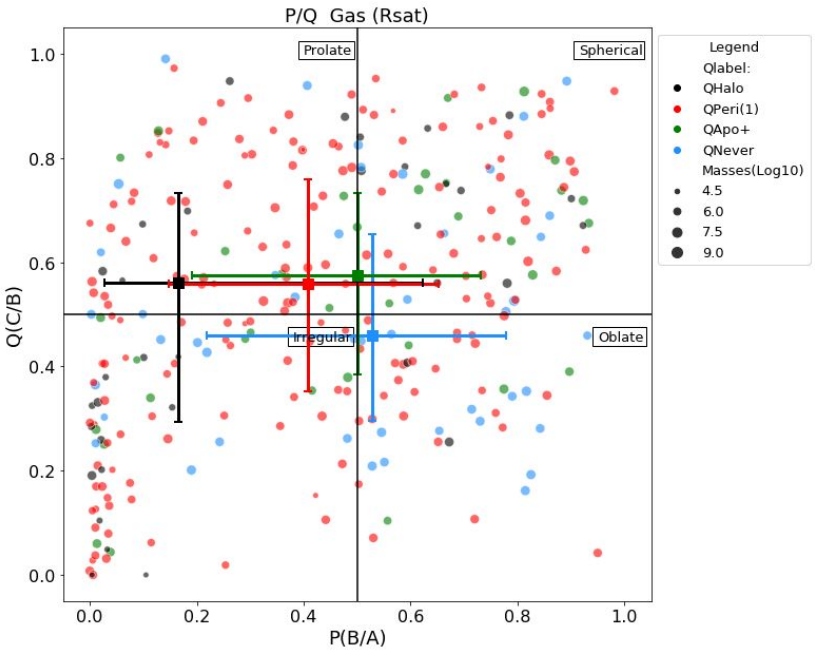



Figure 19: Scatter plot showing P(B/A) vs Q(C/B) clustering on Gas at $R_{sat}$ Color coding by SG Qlabel. Point size by component mass at selected radius, measured in solar mass $\approx 2\times10^{30}kg$. Error bars are centered around the median of Qlabel classes, and variance by 75% and 25% percentile.

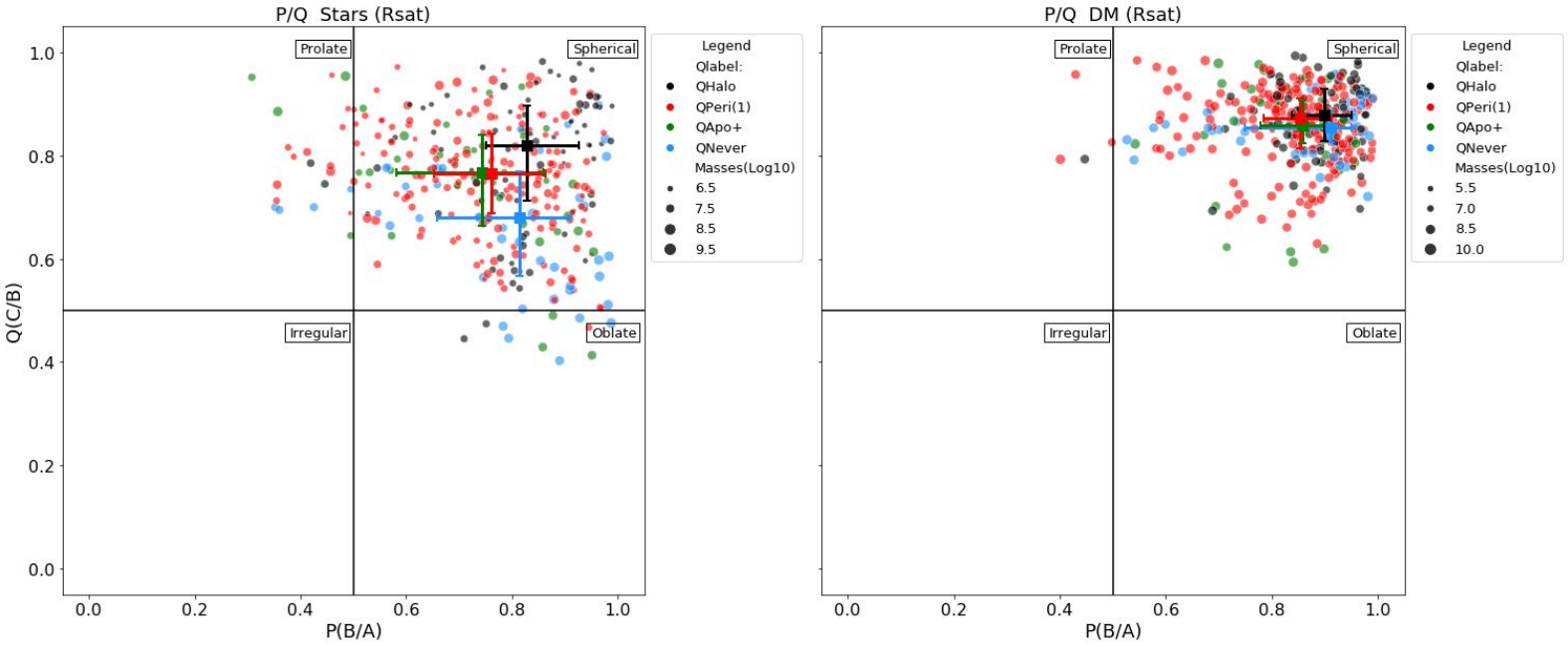



Figure 20: Scatter plot showing P(B/A) vs Q(C/B) clustering on Stars and DM at $R_{sat}$. As in Figure 19. For both Stars and DM (appendix), SG shapes mainly appear as spherical, and no there is no clustering by Qlabel or swansong.

As seen by mass in the legend, irrelevant of DM or stellar mass supremacy, we see that SG stellar and DM components are spherical over the entire sample of SGs at entrance to halo, with few outliers near prolate and oblate shapes.

We find this SG spherical shape distribution to be different to the CG shape evolution: as per *Tomasetti et al., (2018),* where stellar and DM components evolve from prolate during the DM dominated era, to oblate during stellar dominant era. The causes for this are unclear as of yet.

We find that the SG gas component at entrance to halo ranges greatly throughout all shape P/Q combinations, but with a slight tendency towards prolate shape for all Qlabels. The exception is QNever, which lands within the scope of oblate. We do not find there to be a statistical relation in between component mass and shape. In addition, similarly to *Tomasetti et al., (2018),* we find a significant population of SGs with highly triaxial shapes. However, we find the triaxiality to be more extreme, as 27.4% of all SG gas components have P values of <=0.2.

In both stellar and gas components, the shapes for the Qlabel *QNever* are slightly oblate. Similarly, for *QPeri* both components are prolate leaning. A question arises: why do stable oblate disks quench less often than elongated prolate shapes?

For *QHalo*, we see a clear grouping, as the stellar component has full stellar shape, while gas component more often than not is shapeless, as P ~ 0 is indicative of an irregular non-uniform gas distribution. This makes sense: as in CGs, quenched nearly gasless SG galaxies transform into sphericals.

Another question arising is the difference of gas component shapes between *QPeri* vs *QApo+,* where QPeri tend to have already more disorganized prolate shapes, while QApo+ sits between prolate and spherical. On the other hand, the spherical components are identical between them. Why is that?

## 4.2 - Quenching Model

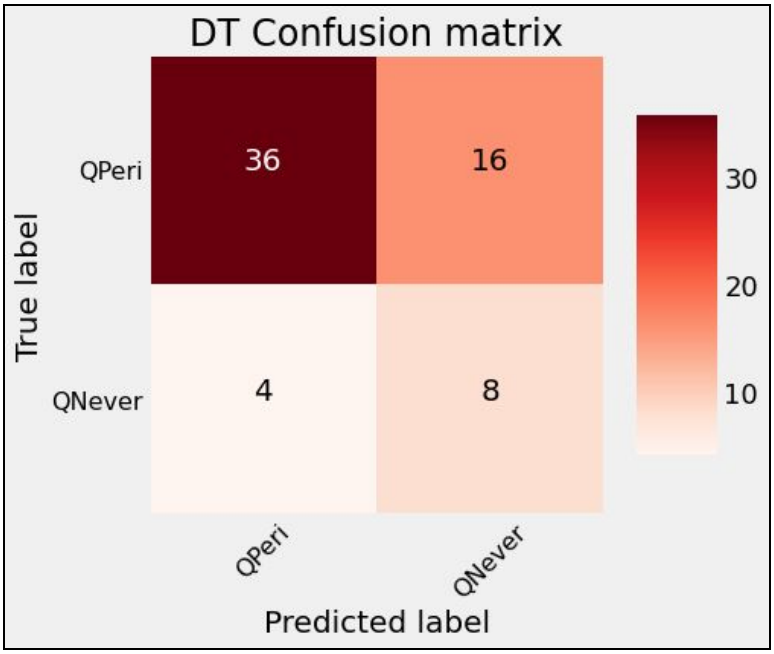
In this chapter we will provide an in-depth analysis of our ML findings, including a highly accurate predictive model for SG quenching.

### 4.2.1 - Baseline Model (Done on QPeri vs QNever)

For our baseline model, we used Pearson's correlation coefficient r, to select the 5 least correlative features. Doing so, we get most of the variance within the dataset, and theoretically, a good feature set for a decision tree.

| **Feature List:** | | | | | **Absolute Correlation** |
|---|---|---|---|---|---|
| dm_q(1kpc) | EllipsArea gas(0.5kpc)_rvir | Rsat/Rvir[kpc] | EllipsDepth stars(0.5kpc)_rvir | theta (Vorbit*Rorbit) | 0.036 |

Running the above through a DT model, we achieved our baseline confusion matrix and evaluation metric:

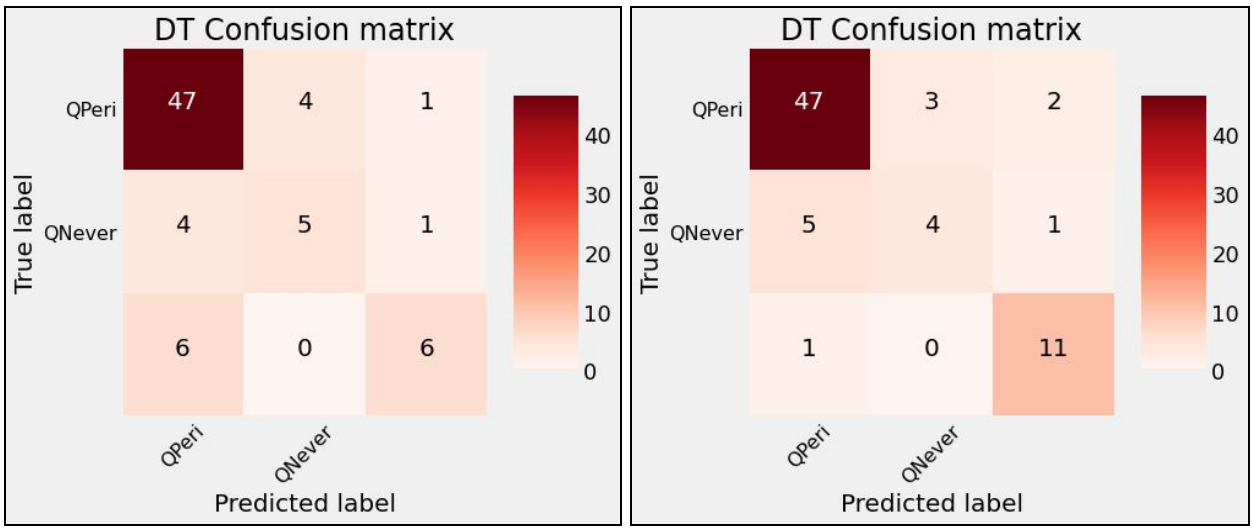



| Class | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| QPeri | 0.90 | 0.71 | 0.79 | 52 |
| QNever | 0.35 | 0.67 | 0.46 | 12 |

Figure 21: Confusion Matrix and evaluation metrics of the DT trained on uncorrelated features.

### 4.2.2 - ML Quenching Model

Our Quenching model was built on a three-class expanded dataset: *QPeri, QApo+,* and *QNever*. We found the expanded dataset to have significantly higher prediction ratings on the training and testing datasets, and no overfitting was found to occur. The first 'dry run' results, run on 75 features after Pearson correlation reduction:

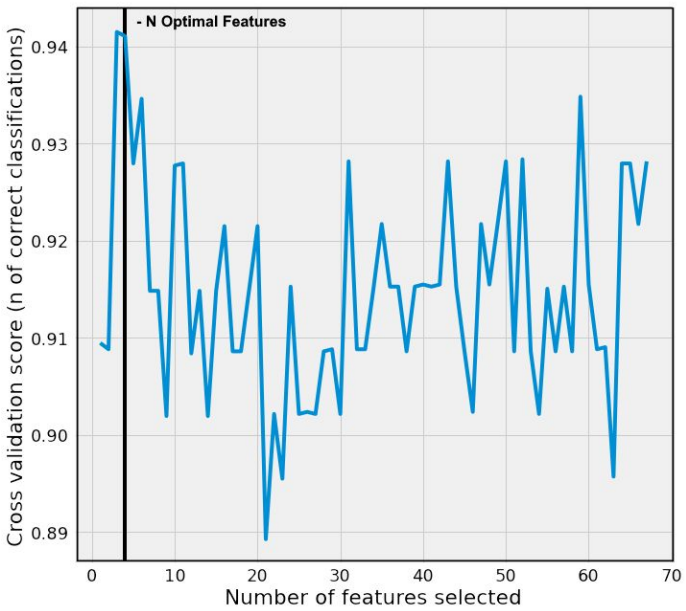



| Class | Precision | Recall | f1-score | | Precision | Recall | f1-score | | Support |
|---|---|---|---|---|---|---|---|---|---|
| QPeri | 0.82 | 0.90 | 0.86 | | 0.89 | 0.90 | 0.90 | | 52 |
| QApo+ | 0.56 | 0.50 | 0.53 | | 0.57 | 0.40 | 0.47 | | 10 |
| QNever | 0.75 | 0.50 | 0.60 | | 0.79 | 0.92 | 0.85 | | 12 |

Figure 22: Confusion Matrix and evaluation metrics of the decision tree classifier dry run on the left, and after dimensionality reduction on the right.

After dimensionality reduction, we reached a set of 5 features using RFECV, and the model was still unsuccessful in predicting QApo+. As a result, we dropped the class, continuing with a two-class expanded dataset: QPeri and QNever.

We refitted the model and went through feature reduction, with RFECV arriving at 4 optimal features.



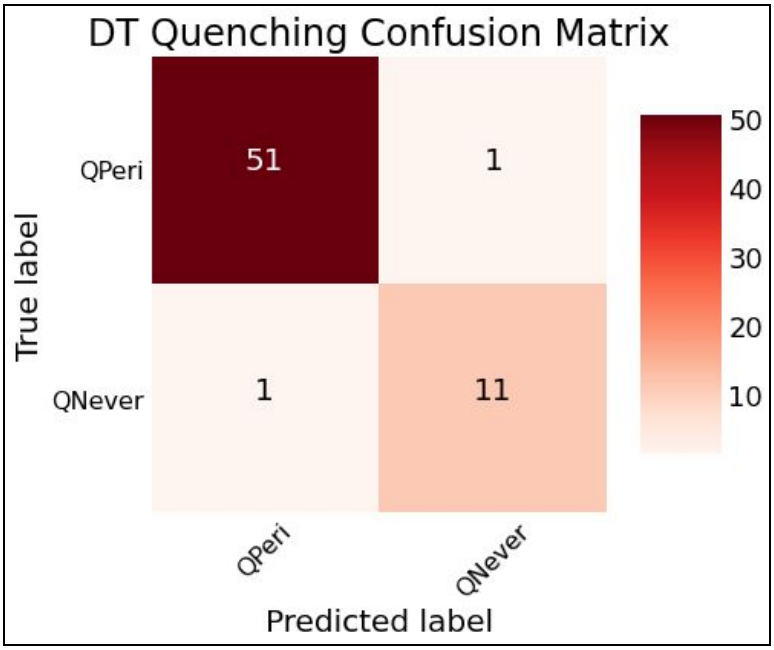

| Optimal number of features :4 |
|---|

Selected Features:

| Eccentricity(Rperi_1,Rapo_1) |
|---|
| Msat/Mvir |
| Rsat |
| EllipsArea_stars(Rsat) |
| \|Correlation\| = 0.488 |

Figure 23: RFE cross-validation results, with final chosen features. RFECV finds that 4 features (reduced from 67) are optimal for classification. The selected features are listed in the table above.

With these chosen features, we re-split the data, and ran the DT classifier.




| Class | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| QPeri | 0.98 | 0.98 | 0.98 | 52 |
| QNever | 0.92 | 0.92 | 0.92 | 12 |

Figure 24: Confusion Matrix and Metric Evaluation Report of final Quenching Model. We succeeded at creating a DT model with a high f1-score ($\geq 0.90$), with only two misclassifications in the testing data.

Our RF + RFECV → DT pipeline performed stellarly on the 2-class classification problem, reaching a weighted f1-score of 0.94, using only 4 features (min 3, max 10). We were surprised to find *EllipsArea_stars(Rsat)* in the selected optimal features, as seen in chapter 4.1.1 there was significant clustering of all Qlabels as spherical shapes, with few outliers.

This further signifies the usefulness of a randomized bootstrapped algorithm such as RF during feature selection. Mainly, it avoids 'feature selection overfitting', and picks up on alternate features that may fit better within the final predictive model.

### 4.2.3 - Predictive Model

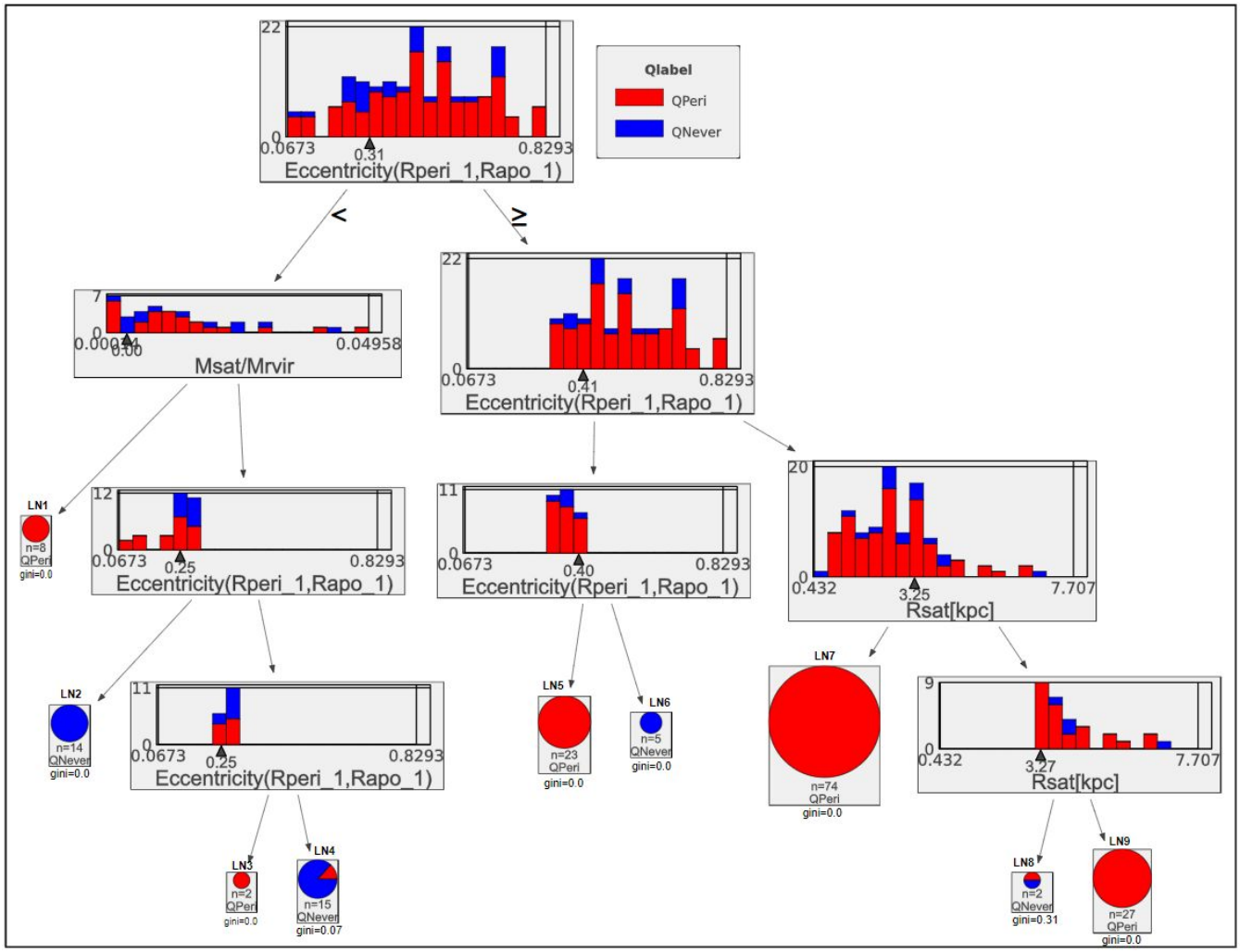



Figure 25: Decision Tree visualized using DTreeViz. We additionally see the sample size and gini index for each leaf node (LN). The split is based on the value seen in each node, with $X_i <$ Split going to the left, and $X_i \geq$ split going to the right.

Analyzing the model, we notice several major attributes:

LN7 contains the largest homogenous sample sizes of 74, accounting for 54% of QPeri identification in the training data. Other significant LN's include LN9 and LN6, accounting for another 19.7% and 16.8% respectively. LN2 and LN4 together account for 81.8% of QNever in the training data.

We see that the 2 class final model is able to accurately divide the training data using only 3 of the 4 provided features; EllipsArea_stars, as predicted based on 4.1.1 P/Q graphs, is not a

strong enough feature to present itself in the DT. Additionally, eccentricity is the overall strongest feature for this classification problem, as predicted in *Nussbaum, 2018.*

Finally, we point out that using these chosen features, the DT is able to achieve homogeneity (gini impurity = 0.0) in 7/9 LNs, without overfitting.

### 4.2.4 - Main Features Distribution Analysis

Graphing the feature distribution among the two classes, we can see a trend towards a split between the classes. However, it can be seen that 1, or even 2 features alone would not be enough for an accurate model.

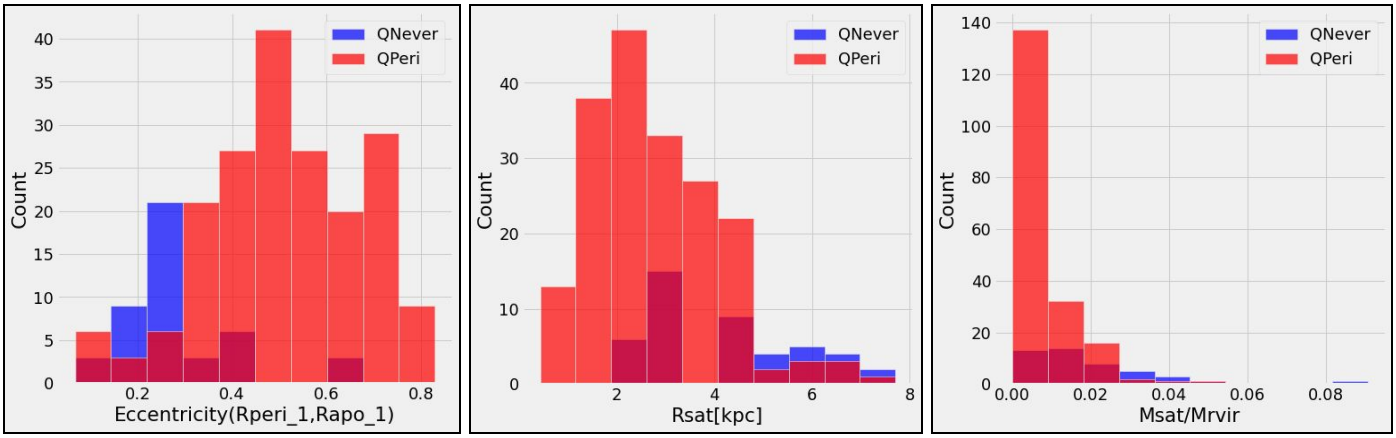



Figure 26: Eccentricity, Rsat, Msat/Mvir distribution among QPeri and QNever.

The final predictive model used 3 features, and the distribution of the two classes can be visualized with a 3D scatterplot:

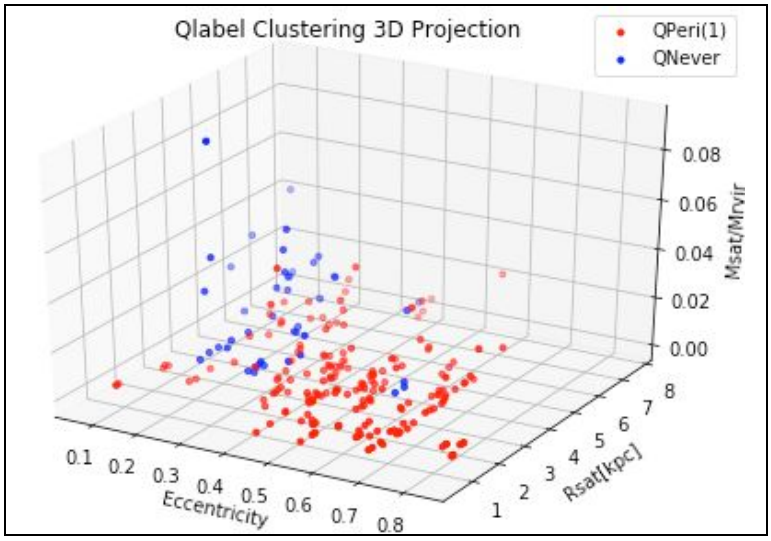



Figure 27: 3D projected scatter-plot of the SG dataset provided to the ML.

We see that while a trend towards a split between the two classes exists, the distribution of the data more closely resembles different clusters sharing a Qlabel. These results point to the idea that there exists more than one physical cause for non-quenching SGs.

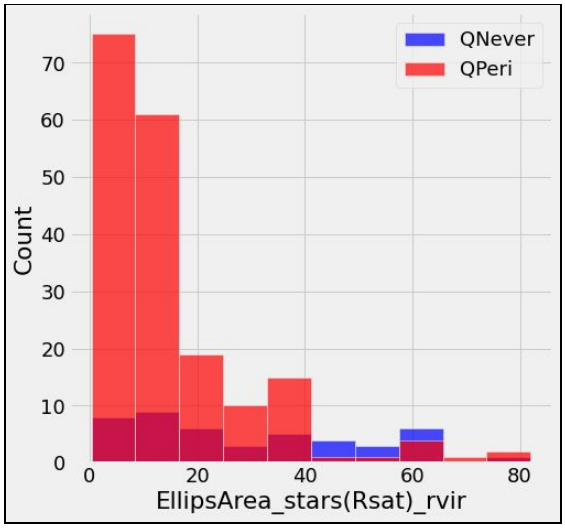



Figure 28: EllipsArea_stars(Rsat) distribution among QPeri and QNever.

Additionally, observing the EllipsArea_stars(Rsat) distribution plot: We see a significant portion of QPeri SGs with low values, while QNever has an even spread throughout the range of values.

## 4.3 - Swansong Model

Insofar, testing on the Swansong labels have been inconclusive. Further feature engineering and testing remains. While *Nussbaum, (2018)* found that these cases can be predicted using dynamical forces at the peri-center, our RFC Model fails using the provided features in the dataset.

# 5 - Discussion

In this chapter, we discuss the results of the ShapeTensor at entrance to halo and ML, in addition to providing an analysis of the chosen features, and their physical implications.

## 5.1 - SG Shape Dependency

As shown in chapter 4.1, it is clear that SG evolution from its very beginnings are fundamentally different to that of CG evolution. Importantly, the lack of a significant prolate/oblate shape in the stellar and DM components throughout the data sample. SGs are smaller cosmological bodies, and in addition undergo fewer mergers throughout their lives. Further research remains to be done on this subject.

However, the findings of QNever stellar component trending towards oblate shape, while QPeri,QApo+ towards prolate, is very interesting. As shown in in *(Wang, Wang, Mo, Lim, et al., 2018; Wang, Wang, Mo, van den Bosch, et al., 2018)*, diffuse gas components are more likely to quench, and likely to be quenched at halo. We find this in keeping with our results, where the median of QHalo gas shape lands very near P~0, and similarly QPeri gas shape has a high percentage of P~0 samples. More dense gaseous SGs have a more defined shape, and are predicted to survive for more peri-centers, while diffuse SGs are expected to disrupt tidally. Our findings support this, as QNever and QApo+ populations trend further towards sphericality or oblate disks.

QNever stellar and gas components have oblate shape at Rsat and R0.9coldgas, with a dense spherical core at 0.5kpc.

## 5.2 - SG Physical Implication Analysis

Using the methods detailed in the paper, we find that the 3 features $Eccentricity_{(Peri_1, Apo_1)}$, $M_{sat}/M_{Rvir}$, and $R_{sat}$ are the greatest predictors of QPeri vs QNever. In addition, we find that our RF+RFECV feature selection stage choosing 4 optimal features (from 3 to 10 possible number o features), including $Ellips_{area}(Stars)_{Rsat}$. Despite that, the final DT classification model does not find statistical importance in that feature.

As seen in the appendix, eccentricity and Msat/Mrvir are non-correlative to other features, a signal that they are indeed physically important. Rsat however, is highly correlative to several other features, including $Ellips_{area}(Stars)_{Rsat}$, $Ellips_{depth}(Stars)_{Rsat}$, $Mdm$, and more. As such, we are tentative to pinpoint SG stellar radius specifically being more or less influential in SG evolution in the dynamical environment of the CG halo.

The following will include an analysis of the 3 primary features, as they affect the SG dynamical forces:

From the forces ratio $f_{ram}/f_{tidal}$, we learn that ram pressure will be the dominant force throughout the SG orbit, while tidal pressure either opposes or supports the stripping depending on the eccentricity of the orbit. The more eccentric an orbit (e>=0.3), the higher the fraction of quenched galaxies. This is due to a closer approach to CG during peri-center, resulting in stronger tidal stripping and heating

$f_{ram}/f_{self-gravity}$ is affected by eccentricity, as circular orbits cause lower ram-pressure. The forces ratio is further dependent on the ratio of SG vs CG masses, whereas SG mass and radius decreases, ram-pressure increases. This supports our findings that Msat/Mrvir and Rsat are primary features for SG quenching.

$f_{tidal}/f_{self-gravity}$ too shows a dependency on SG vs CG mass ratio. As the difference between the masses and the respective potential wells grows, tidal forces grow to overpower the self-gravity. Low eccentricity orbits additionally cause less rapid and violent peri-center events; the potential well stays in the same depth with trimming in the edges of the satellite as the time taken to get to peri-center is lengthened. High eccentricity orbits result in flatter potential well due to the tidal heating. $f_{tidal}/f_{self-gravity}$ further explains the importance of shapes, as oblate galaxies with a dense bulge have higher self-gravity in the dense core, and resist the tidal heating and stripping effects. On the other hand, prolate and circular shapes are more diffuse, resulting in higher tidal heating and stripping.

## 5.3 - Qlabel Division

We find that there may be a gap in the current dataset, as even after feature selection a predictive model can't be found for classifying Swansong events at peri-centers.

There remains further testing to be done on the unused class - QApo+. Several questions remain, namely: Is the cause of the failure feature related, or related to our definition of the class?

One option may be that it is due to certain physical forces or effects acting on the SG during its time in the CG halo, which would be related to gas mass and shape. This would mean more feature engineering and studying each snapshot within the data between halo and peri-center(1).

Another possibility is that the class must be divided further into SGs quenching at apo-center(1), and quenching at peri-center(2). We previously grouped the two on assumption that the same mechanism was responsible for their avoiding quenching at the 1st peri-center, but this may be incorrect.

Tests must be conducted on 2-class ML models with each of the Qlabels, and possible re-tuning of the hyper-parameters set originally may be required.

## 5.4 - Study Limitations

It is important to note that the methods used here are an advisory tool: they do not and cannot replace the physical research, as shown in 5.2. Without the physical context, we would not have a way to check our results.

ML model fitting inherently benefits from larger datasets, something that in the cosmological simulation field is unattainable when high-resolution is a factor. With only 34 central galaxies, and 118 SGs, we were limited by our data. We did expand our dataset using 3 snapshots from each SG. This was allowed by the large time difference between snapshots (on the scale of hundred millions of years) gave way to significant changes in features among the 3 snapshots. This means each snapshot we added acts as a new, slightly different SG sample, rather than an identical sample. We found this method to be superior compared to over-sampling.

We did a small test during the work with the original non-expanded dataset, and we achieved similar evaluations with the selected features. The expanded model however was for more robust and reliable. We believe we succeeded in expanding our dataset by sampling from additional snapshots around Rvir, avoiding overfitting.

In addition, the use of bootstrapping and feature randomness in an ensemble learning algorithm such as RF helps avoid overfitting and ensures model generalization ability during the feature selection stage.

There still remains the limitations of our input data: VELA is one of the most advanced cosmological simulations to date, but still isn't of high enough resolution for certain small effects, such as Kelvin-Helmolts effect. In addition, the simulation evolves only until z=1 (half the age of the universe).

# 6 - Conclusions

Using the ShapeTensor algorithm, we have calculated the ellipsoidal shape for individual SG components at all snapshots. Additionally, we confirmed our results with VIVID, and built on the VELA Simulation Suite catalogs, adding new features to the SG dataset.

We find that the dark matter component for SGs is spherical throughout all quenching paths. Stellar component shape is primarily spherical for already quenched Qhalo galaxies, while long-living non-quenching SGs are slightly more oblate in shape. Peri-center slow and fast quenching is spherical with a tendency towards prolate. Gas component is found to have high variance throughout all quenching paths. This is in accordance with previous research done, finding that dense SGs survive longer around the CG, while diffuse gas and star components are quenched at halo, or quench at first peri-center. No shape-mass dependence is found, neither does baryon dominance have any relation to the results.

We find the ML methods used in this paper to be a powerful tool in predicting SG quenching. Additionally, it provides us with new ways of analyzing data and physical phenomena in the VELA simulation. Many unexpected correlations and non-correlations were found during the feature reduction stage of this work. Further research may be done on any facet of this data.

We successfully created a prediction model for identification of SG quenching at the first peri-center, using a Decision Tree Classifier. Our DT model had a weighted f1-score of 0.97 over a testing set of 64 sample SGs. This was done using the following features: Eccentricity, Msat/MRvir, and Rsat. Additionally, we proved our results using physical explanations of dynamical forces on SGs in the environment of their central galaxy.

A link to our full code is located in Appendix C.

# 7 - Further Research

We are still in an ongoing search to find the physical explanation for fast vs slow quenching and peri-center starburst events. The success of the methods for prediction on QPeri vs QNever suggests that the issue is in the provided features in the dataset, rather than the methods.

In addition to the methods detailed, we will look into alternate dimensionality reduction options for feature selection and may look into alternative ML algorithms. Further, we would like to test the abilities of regression algorithms on VELA, for features such as sSFR, dynamical forces, and more.

We emphasize that we now have new information that helps us to prepare a complete analytical model, which is our goal in order to understand the world in depth. Another point is to prepare a Bayesian or other continuous model and with its help reach a semi-analytical model of SG evolution.

# Bibliography


Allday, J. (2002). *Quarks, leptons and the big bang.* Bristol: Institute of Physics Publ.

Carroll, S. M. (2007). *Dark matter, dark energy. the dark side of the universe.* Chantilly, VA: Teaching Co.

Dekel, A., Lapiner, S., Freundlich, J., Ginzburg, O., Jiang, F., Kretschmer, M., . . . Primack, J. (n.d.). *Compaction to a blue nugget: a critical phase in galaxy evolution.* (In preparation)

Tacchella, S., Dekel, A., Carollo, C. M., Ceverino, D., DeGraf, C., Lapiner, S., . . . Joel, R. P. (2016). *The confinement of star-forming galaxies into a main sequence through episodes of gas compaction, depletion and replenishment.* Monthly Notices of the Royal Astronomical Society, 457(3), 2790–2813. doi: 10.1093/mnras/stw131

Press, W. H., & Schechter, P. (1974). *Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation.* Astrophysical Journal, 187, 425-438. doi: 10.1086/152650

Wuyts, S., Schreiber, N. M. F., van der Wel, A., Magnelli, B., Guo, Y., Genzel, R., . . . Tacconi, L. (2011). *GALAXY STRUCTURE AND MODE OF STAR FORMATION IN THE SFR-MASS PLANE FROMz^a 2.5 TOz^a 0.1.* Astrophysical Journal, 742(2), 96. doi: 10.1088/ 0004-637x/742/2/96

jie Peng, Y., Lilly, S. J., Kovac, K., Bolzonella, M., Pozzetti, L., Renzini, A., . . . Scaramella, R. ˇ (2010). *Mass and environment as drivers of galaxy evolution in sdss and zcosmos and the origin of the schechter function.* Astrophysical Journal, 721(1), 193.

van den Bosch, F. C., Aquino, D., Yang, X., Mo, H. J., Pasquali, A., McIntosh, D. H., . . . Kang, X. (2008). *The importance of satellite quenching for the build-up of the red sequence of present-day galaxies.* Monthly Notices of the Royal Astronomical Society, 387(1), 79–91. doi: 10.1111/j.1365-2966.2008.13230.x

Woo, J., Carollo, C. M., Faber, S. M., Dekel, A., & Tacchella, S. (2017). *Satellite quenching, galaxy inner density and the halo environment.* Monthly Notices of the Royal Astronomical Society, 464(1), 1077-1094. doi: 10.1093/mnras/stw2403

Kravtsov, A. V., Klypin, A. A., & Khokhlov, A. M. (1997). *Adaptive Refinement Tree: A New High-Resolution N-Body Code for Cosmological Simulations.* Astrophysical Journal, Supplement, 111, 73-94. doi: 10.1086/313015

Ceverino, D., & Klypin, A. (2009). *The Role of Stellar Feedback in the Formation of Galaxies.* Astrophysical Journal, 695, 292-309. doi: 10.1088/0004-637X/695/1/292

Mitchell, M. T. (1997). *Machine Learning.* New York : McGraw-Hill

Alpaydin, E. (2010). *Introduction to Machine Learning.* The MIT Press

Rokach L., Maimon O. (2005). *Decision Trees.* In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. doi: 10.1007/0-387-25465-X_9

Quinlan, J. R. (1986). *Induction of decision trees.* Mach Learn **1,** 81–106. doi: 10.1007/BF00116251

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees. Statistics/probability series*. Wadsworth & Brooks/Cole Advanced Books & Software.

Breiman, L. (2001). *Random Forests.* Machine Learning **45,** 5–32. Doi: 10.1023/A:1010933404324

Powers, D. M. W. (2011), *Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation*. Journal of Machine Learning Technologies **2** (1), 37--63. doi: 10.9735/2229-3981

Tweed, D., Devriendt, J., Blaizot, J., Colombi, S., & Slyz, A. (2009). *Building merger trees from cosmological n-body simulations.* Astronomy and Astrophysics, 506(2), 647–660. doi: 10.1051/0004-6361/200911787

More, S., Kravtsov, A. V., Dalal, N., & Gottlöber, S. (2011). *The overdensity and masses of the friends-of-friends halos and universality of halo mass function.* Astrophysical Journal Supplement Series, 195(1), 4.

Simpson, C. M., Grand, R. J. J., Gómez, F. A., Marinacci, F., Pakmor, R., Springel, V., . . . Frenk, C. S. (2018). *Quenching and ram pressure stripping of simulated milky way satellite galaxies.* Monthly Notices of the Royal Astronomical Society, 478(1), 548–567. doi: 10.1093/mnras/sty774

Gunn, J. E., & J. Richard, I. G. (1972). *On the infall of matter into clusters of galaxies and some effects on their evolution.* Astrophysical Journal, 176, 1. doi: 10.1086/151605

Dekel, A., Devor, J., & Hetzroni, G. (2003). *Galactic halo cusp-core: tidal compression in mergers.* Monthly Notices of the Royal Astronomical Society, 341(1), 326-342. doi: 10.1046/ j.1365-8711.2003.06432.x

Matteo Tomassetti, Avishai Dekel, Nir Mandelker, Daniel Ceverino, Sharon Lapiner, Sandra Faber, Omer Kneller, Joel Primack, Tanmayi Sai, (2016). *Evolution of galaxy shapes from prolate to oblate through compaction events*. Monthly Notices of the Royal Astronomical Society, 458(4), 4477–4497. doi: 10.1093/mnras/stw606

Ferguson, C. C. (1979) *Intersections of Ellipsoid and Planes of Arbitrary Orientation and Position*. Mathematical Geology, 11(3), 329-336. doi: 10.1007/BF01034997

Pedregosa et al., (2011). *Scikit-learn: Machine Learning in Python.* JMLR 12, 2825-2830.

Harshil M. Kamdar, Matthew J. Turk, Robert J. Brunner, (2015). *Machine learning and cosmological simulations – I. Semi-analytical models.* Monthly Notices of the Royal Astronomical Society, 455(1), 642–658. doi: 10.1093/mnras/stv2310

Wang, E., Wang, H., Mo, H., Lim, S. H., van den Bosch, F. C., Kong, X., . . . Chen, S. (2018). *The dearth of difference between central and satellite galaxies. i. perspectives on star formation quenching and AGN activities.* Astrophysical Journal, 860(2), 102. doi: 10.3847/1538-4357/aac4a5

Wang, E., Wang, H., Mo, H., van den Bosch, F. C., Lim, S. H., Wang, L., . . . Chen, S. (2018). *The dearth of differences between central and satellite galaxies. II. comparison of observations with l-GALAXIES and EAGLE in star formation quenching.* Astrophysical Journal, 864(1), 51. doi: 10.3847/1538-4357/aad554

# Appendix A (Galaxy Example)

In figure 29, we present an evolution of SG galaxy parameters over time. The blue dotted line in the first 2 rows represents Rorbit, the distance from SG to CG. The yellow line represents sSFR. We can see that as the SG approaches its CG, it quenches (sSFR ~ 0). When we reach the minimum point on the Rorbit - the pericenter - the sSFR spikes to pre-quenching values for one snapshot, before quenching again. This would be classified as a QHalo - Swansong SG.

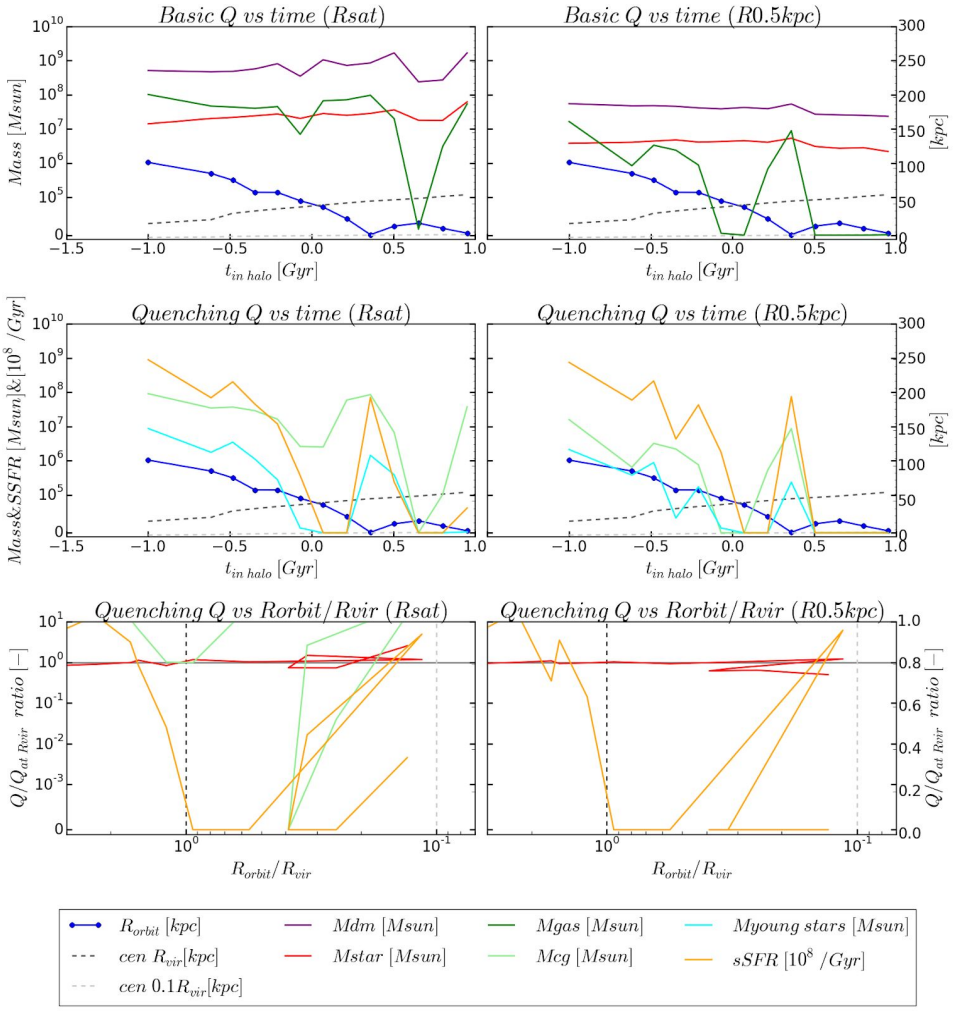



Figure 29: Satellite galaxy 04-005 forces evolution: Each plot depicts a SG property against the time or against the distance, and splits to two columns; left column focuses on the SG enclosed in a sphere of $R_{sat}$ radius and right column focus on the $0.5kpc$ radius sphere. By rows, the first row describes basic galactic parameters over time in halo: $R_{orbit}$ (blue), $R_{vir}, 0.1R_{vir}$ (dashed lines), dark matter mass (purple), stellar mass (red), gas (green). The second row describes additional star formation related parameters: in addition to the $R_{orbit}$ and $R_{vir}$ over time in halo, which are the sSFR (orange), cold gas mass (light green) and young stars (cyan). The last row describes important quenching and stripping parameters depending on the distance to the central galaxy normalized by their amount while entering $R_{vir}$: sSFR, stellar mass and cold gas (with the same coloring as before). This figure presents a fast quenching case with swansong event of an SG. We can see the SG enter the halo at the point the blue line $R_{orbit}$ intersects with the grey dashed line $R_{vir}$, at which point sSFR on both columns is greatly reduced (despite the cold gas component still remaining within the SG). Later around the first peri-center which is the minimum of the blue line, we can see by the light green and orange lines that the SG experiences cold gas compaction very sharply (less than 100 Myr) and go through starburst event, before equally sharply quenching and losing cold gas component.

# Appendix B (Feature List)

Features dropped during pearson correlation:

| Prime Feature | Correlative Feature | Correlation *r* |
|---|---|---|
| Mcg(Rcoldgass_0.9)[Msun] | Mcg(Rsat)[Msun] | 0.835 |
| | Mgas(Rsat)[Msun] | 0.840 |
| | Mdm(Rcoldgass_0.9)[Msun] | 0.825 |
| | Mgas(Rcoldgass_0.9)[Msun] | 0.999 |
| | Mtot(Rcoldgass_0.9)[Msun] | 0.822 |
| | Mncg(Rcoldgass_0.9)[Msun] | 0.917 |
| | Mncg(Rsat)[Msun] | 0.814 |
| Mdm(R0.5kpc)[Msun] | Mstar(R0.5kpc)[Msun] | 0.830 |
| | Mtot(R0.5kpc)[Msun] | 0.894 |
| | density(R0.5kpc)[Msun/kpc^3] | 0.830 |
| | density(R1kpc)[Msun/kpc^3] | 0.856 |
| | sigma(R0.5kpc)[Msun/kpc^2] | 0.830 |
| | sigma(Rsat)[Msun/kpc^2] | 0.814 |
| Mgas(R0.5kpc)[Msun] | Mstar_young(R0.5kpc)[Msun] | 0.825 |
| | Mcg(R0.5kpc)[Msun] | 1.000 |
| | sfr(Rcoldgass_0.9)[Msun/yr] | 0.820 |
| center_Mstar(0.1rvir)[Msun] | M_rvir[Msun] | 0.945 |
| | Mdm(rvir)[Msun] | 0.941 |
| | Mdm(0.15rvir)[Msun] | 0.971 |
| | Mcold_gas(0.15rvir)[Msun] | 0.811 |
| | Mtot(rvir)[Msun] | 0.945 |
| | CGH_Mtot(Rorbit)[Msun] | 0.907 |
| | Mdarkmatter(0.2rvir) | 0.968 |
| density(Rsat)[Msun/kpc^3] | Mstar(R0.5kpc)[Msun] | 0.857 |
| | Mtot(R0.5kpc)[Msun] | 0.838 |
| | density(R0.5kpc)[Msun/kpc^3] | 0.857 |
| | density(R1kpc)[Msun/kpc^3] | 0.805 |
| | sigma(R0.5kpc)[Msun/kpc^2] | 0.857 |
| | sigma(Rsat)[Msun/kpc^2] | 0.947 |
| Myoung_stars(0.15rvir)[Msun] | SFR(disc)[Msun/yr] | 0.950 |
| | SFR(0.15rvir)[Msun/yr] | 0.986 |
| | SFR(R_disc)[Msun/yr] | 0.979 |
| Rorbit-Rvir[kpc] | R_orbit/R_vir[kpc] | 0.925 |

| | | |
|---|---|---|
| Rsat[kpc] | Mdm(Rsat)[Msun] | 0.827 |
| | Mncg(Rsat)[Msun] | 0.875 |
| | R_cg_eff(approx)[kpc] | 0.895 |
| | EllipsArea_stars(Rsat)_rvir | 0.878 |
| | EllipsDepth_stars(Rsat)_rvir | 0.933 |
| | EllipsArea_stars(Rsat)_rvir_Vorbit | 0.827 |
| center_Rvir[kpc] | M_rvir[Msun] | 0.846 |
| | Mdm(rvir)[Msun] | 0.853 |
| | Mdm(0.15rvir)[Msun] | 0.821 |
| | Mtot(rvir)[Msun] | 0.846 |
| | Mdarkmatter(0.2rvir) | 0.819 |
| stars_AM*Vorbit(0.9coldgas) | stars_AM*Vorbit(Rsat) | 0.852 |
| | stars_AM*dm_AM(Rsat) | 0.852 |
| stars_p(0.5kpc) | EllipsArea_stars(0.5kpc)_rvir | 0.906 |
| gas_p(0.5kpc) | EllipsArea_gas(0.5kpc)_rvir | 0.873 |
| | EllipsArea_gas(0.5kpc)_rvir_Vorbit | 0.823 |
| Rcoldgass_0.9[kpc] | Mdm(Rcoldgass_0.9)[Msun] | 0.832 |
| | Mncg(Rcoldgass_0.9)[Msun] | 0.848 |
| Mdarkmatter(0.1rvir) | M_rvir[Msun] | 0.977 |
| | Mdm(rvir)[Msun] | 0.976 |
| | Mdm(0.15rvir)[Msun] | 0.999 |
| | Mcold_gas(0.15rvir)[Msun] | 0.834 |
| | Mtot(rvir)[Msun] | 0.977 |
| | CGH_Mtot(Rorbit)[Msun] | 0.933 |
| | Mdarkmatter(0.2rvir) | 0.996 |
| V_rvir[km/s] | M_rvir[Msun] | 0.832 |
| | Mdm(rvir)[Msun] | 0.830 |
| | Mtot(rvir)[Msun] | 0.832 |
| | CGH_Mtot(Rorbit)[Msun] | 0.807 |
| | Mdarkmatter(0.2rvir) | 0.806 |
| cen_stars_p(Re) | cen_dm_p(Re) | 0.872 |
| EllipsArea_gas(0.9coldgas)_rvir | EllipsArea_gas(0.9coldgas)_rvir_Vorbit | 0.966 |
| EllipsDepth_gas(0.9coldgas)_rvir | EllipsDepth_gas(0.9coldgas)_rvir_Vorbit | 0.900 |
| Vorbit[km/s] | EllipsDepth_stars(0.5kpc)_rvir_Vorbit | 0.929 |
| EllipsDepth_stars(Rsat)_rvir_Vorbit | EllipsDepth_stars(Rsat)_rvir | 0.838 |
| | EllipsArea_stars(Rsat)_rvir_Vorbit | 0.807 |

Additional non-correlative features:

| | | | |
|---|---|---|---|
| CGH_gas_density (Rorbit+-Rsat)[Msun/kpc^3] | Eccentricity(Rperi_1,Rapo_1) | cen_stars_q(Re) | cen_gas_p(1kpc) |
| SFR(1kpc)[Msun/yr] | stars_q(0.5kpc) | cen_dm_q(Re) | cen_gas_q(1kpc) |
| Rorbit[kpc] | dm_p(0.5kpc) | cen_gas_p(Re) | cen_stars_p(10kpc) |
| c(Rsat/Reff) | dm_q(0.5kpc) | cen_gas_q(Re) | cen_stars_q(10kpc) |
| Rsat/Rvir[kpc] | gas_q(0.5kpc) | stars_p(1kpc) | cen_dm_p(10kpc) |
| theta(Vorbit*Rorbit) | stars_p(Rsat) | stars_q(1kpc) | cen_dm_q(10kpc) |
| Rorbit(peri_1)/R_vir(kpc) | stars_q(Rsat) | dm_p(1kpc) | cen_gas_p(10kpc) |
| Msat/Mrvir | dm_p(Rsat) | dm_q(1kpc) | cen_gas_q(10kpc) |
| gas_c_vec*Vorbit (0.9coldgas) | dm_q(Rsat) | gas_p(1kpc) | EllipsDepth_stars (0.5kpc)_rvir |
| gas_c_vec*gas_AM (0.9coldgas) | gas_p(0.9coldgas) | gas_q(1kpc) | EllipsDepth_gas (0.5kpc)_rvir |
| gas_c_vec*stars_AM (0.9coldgas) | gas_q(0.9coldgas) | cen_stars_q(1kpc) | EllipsArea_stars (0.5kpc)_rvir_Vorbit |
| stars_AM*Vorbit(0.5kpc) | ssfr(Rcoldgass_0.9)[1/yr] | cen_dm_q(1kpc) | EllipsDepth_gas (0.5kpc)_rvir_Vorbit |

# Appendix C (Code)

The code written for this paper is available at our github repository:

https://github.com/GalaxyHunters/Vela_satellites