

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

GSSD: Galaxy Similarity from Sorted Differences, and its Relationship to Age, Metallicity, and Star Formation Rate

Permalink

<https://escholarship.org/uc/item/3hk9v5jc>

Author

English, John Edward

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

GSSD: Galaxy Similarity from Sorted Differences, and its Relationship to Age, Metallicity,
and Star Formation Rate

Dissertation

submitted in partial satisfaction of the requirements
for the degree of

Doctor of Philosophy
in Computer Science

by

John Edward English

Dissertation Committee:
Associate Professor Wayne B. Hayes, Chair
Professor Charless Fowlkess
Professor Aaron Barth

2025

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGMENTS	vi
VITA	vii
ABSTRACT OF THE Dissertation	ix
1 Introduction	1
2 Background	3
2.1 SpArcFiRe	3
2.2 Sloan Digital Sky Survey	4
2.3 Galaxy Zoo	5
2.4 Galaxy Morphology	5
2.4.1 Hubble Sequence	5
2.5 EFIGI	7
2.6 Decision Tree Regression	8
2.7 Tested parameters	8
2.7.1 Age	9
2.7.2 Star Formation Rate	9
2.7.3 Metallicity	9
3 Determining Similarity	11
3.1 Developing the similarity order	12
3.1.1 Morphology Selection	15
3.1.2 Classification Background	15
3.1.3 Feature Weights	17
3.2 Normalization	18
3.3 Minimum local maximum difference	19

4	Quality Analysis of Similarity	21
4.1	Building an Ideal Set of Galaxies	22
4.1.1	Step 1. Filter by Spirality	23
4.1.2	Step 2. Filter by Size of the Image	24
4.1.3	Step 3. Least Different set for each Hubble Type	25
4.1.4	Curated Good Reference Set	25
4.2	Where the algorithm is limited	27
4.3	Expected Results	30
4.4	How Feature Removal Affects Rankings: Illustrated by Hubble Type	32
4.4.1	Hubble Type T 0	32
4.4.2	Hubble Type T 1	36
4.4.3	Hubble Type T 2	39
4.4.4	Hubble Type T 3	43
4.4.5	Hubble Type T 4	46
4.4.6	Hubble Type T 5	50
4.4.7	Hubble Type T 6	53
4.4.8	Hubble Type T 7	57
5	Novel relationships to spiral structure	62
5.1	Global Relationship between Pairwise Distance and Non-Spiral Parameters	62
5.2	Single Galaxy Variance from Global Relationships	64
5.3	Neighborhood to Neighborhood Relationships	64
5.3.1	NGram Analysis	66
5.4	Analysis of Selected SpArcFire Features	74
6	Future Work	77
6.1	Adapting the Algorithm to Other Surveys	77
6.1.1	Well Established Surveys	77
6.1.2	New Surveys	78
6.2	Comparing Real and Simulated Galaxies	78
6.3	SpArcFiRe Galaxy Morphological Classification	79
6.4	Cross Survey Search Engine	79
	Bibliography	80

LIST OF FIGURES

	Page
2.1 A visualization of Hubble’s initial galaxy morphology classification.	6
2.2 Caption	7
3.1 Cross-class Morphological Ordering Example	14
3.2 Intersection of GZ and EFIGI	16
4.1 Galaxies with a Spirality of 1	23
4.2 Galaxy Area Distirbution	24
4.3 Galaxies filtered by Spirality and Area	25
4.4 Crated Example Galaxies	26
4.5 Example of a well fit target with bad neighbors	28
4.6 Example of a badly fit target with bad neighbors	29
5.1 Initial t-SNE	63
5.2 t-SNE Progression Table With Means	65
5.3 t-SNE Progression Table Without Means	65
5.4 SFR vs totalArcLength Direct Comparison	76

LIST OF TABLES

	Page
3.1 SpArcFiRe RFR Weights Table	20
4.1 Randomly Sampled Orderings	30
4.2 T 0 Feature Removal table	33
4.3 T 1 Feature Removal table	36
4.4 T 2 Feature Removal table	39
4.5 T 3 Feature Removal table	43
4.6 T 4 Feature Removal table	46
4.7 T 5 Feature Removal table	50
4.8 T 6 Feature Removal table	54
4.9 T 7 Feature Removal table	58
5.1 Sample Galaxy Relationship Table	64
5.2 GSSD Data Saturation Relationships Table	68
5.3 Hubble Sequence Relationship Comparison table	69
5.5 Tabel of 2gram Features	70
5.6 Table of 9grams	71
5.7 3x3 table of 1grams	71
5.8 3x3 Table of 2grams	72
5.9 Table of 6grams	73
5.10 Age and Metallicity Direct Feature Comparison	75

ACKNOWLEDGMENTS

First I would like to thank my Advisor Wayne Hayes. Without him giving me the opportunity to do research as an undergraduate back in 2015, I likely never would have pursued a career in academic research and would still be miserable in the software industry. His philosophy on research and the faith he put in me to work on large problems matched my outlook and goals. This thesis started as a 1 week task for me to manually look over some images from Illustris and find good looking spiral galaxies. When I decided to try and use cosine similarity to find galaxies because I didn't have time that week to look at the data manually, it put me on an adventure that let me to this thesis. Thank you very much for always giving me the resources and time I needed to work on hard problems, and the pep talks I needed to cheer me up when my ideas failed.

Second I would like to thank my committee Aaron Barth and Charless Fowlkes. Most of my interactions with you two have been disorganized, from my behavior, and taken almost as much time up as 2 final defenses. The advise you gave me during my advancement to candidacy led me to the strongest publishable results that I found while working on my thesis. Thank you again for all the time and advice you gave me. I also would like to thank Michael Dillencourt who was on my committee for my advancement but not my final Defense. When one of my original committee members dropped out last minute, I found him in the hallway of Donald Bren Hall and he agreed to drop in like the 101st airborne and save the day.

I'm so grateful and proud to have received all of the help my family gave me. My parents let me move back in with them during the covid lock-downs. As people with athletic hobbies, they were very supportive of my overwhelmingly nerdy interests. They have always done their best to check in on me every week with a video call. Thank you so much for everything. My aunt Linda and uncle Kent also hosted me at their house for dinner every so often. Those meals were high impact stress relief.

I want to thank all of my friends from over the course of both my master's degree and my PhD journey. I can confidently say, without your patience in study groups I would have been unable to maintain the GPA I needed to be accepted into the PhD program. Also the fun times going to dinner, movies, and other adventures helped clear out the miasma in my mind left by focusing on one problem for 6 years.

Finally I would like to thank UCI for accepting me as a student. It would be impossible for me to advance, or defend without being accepted in the first place. As well thank you for following the labor contract and allowing me to be a TA for the majority of my time here. It kindled a passion I never thought I find in teaching.

VITA

John Edward English

EDUCATION

Doctor of Philosophy in Computer Science

2025

University of California, Irvine

Irvine, California

Master's of Science in Computer Science

2019

University of California, Irvine

Irvine, California

Bachelor of Science in Computational Sciences

2016

University of California, Irvine

Irvine, California

RESEARCH EXPERIENCE

Graduate Research Student

2018–2025

University of California, Irvine

Irvine, California

TEACHING EXPERIENCE

Teaching Assistant

2018–2025

University of California, Irvine

Irvine, California

REFEREED JOURNAL PUBLICATIONS

Peng, Tianrui (Rae) and English, John Edward and Silva, Pedro and Davis, Darren R and Hayes, Wayne B SpArcFiRe morphological selection effects due to reduced visibility of tightly winding arms in distant spiral galaxies 2018

Monthly Notices of the Royal Astronomical Society

Fetherolf, Tara and Pepper, Joshua and Simpson, Emilie and Kane, Stephen R and Močnik, Teo and English, John Edward and Antoci, Victoria and Huber, Daniel and Jenkins, Jon M and Stassun, Keivan and others: Variability Catalog of Stars Observed during the TESS Prime Mission 2023

The Astrophysics Journal, Supplemental Series

ABSTRACT OF THE DISSERTATION

GSSD: Galaxy Similarity from Sorted Differences, and its Relationship to Age, Metallicity,
and Star Formation Rate

by

John Edward English

Doctor of Philosophy in Computer Science

University of California, Irvine, 2025

Associate Professor Wayne B. Hayes, Chair

SpArcFiRe is an algorithm written by a previous PhD student, Darren Davis, that finds, delineates, and quantifies structure in the arms of spiral galaxies. For each arm it measures the length, width, area, and pitch angle of the arm, among many other statistics. The goal of this thesis is to use SpArcFiRe’s output measures to create a “similarity” measure between pairs of galaxies in order to find if galaxies of similar “shape” or “structure” have similar physical properties. while this can be useful for pairs of real galaxies, we also note that when one galaxy is real and the other is simulated, we can also use such a measure to test the result of galaxy simulations against real galaxies. Our method is called GSSD (Galaxy Similarity from Sorted Differences). GSSD defines the similarity between two galaxies as the minimum local maximum among pairwise differences in their normalized features. Given that SpArcFiRe outputs hundreds of features for each galaxy, a substantial portion of this thesis comprises how we determined the most discriminating features and chose a minimal set of useful ones. We then used GSSD to compute pairwise differences between 29,000 of the most “spiral-like” galaxies (as chosen by the Galaxy Zoo project) in the Sloan Digital Sky Survey. For each such galaxy, GSSD allows us to create a list of “nearest-neighbors” in feature space, essentially allowing us to find other galaxies that look very similar to a target galaxy. Finally, we show that for the specific existing measurements of star formation rate,

mean stellar age, and metallicity, the difference in these physical parameters is correlated with GSSD’s measure of their “shape” difference. Following in the footsteps of the galaxy morphological classification schemes of the Hubble Sequence and the De Vaucouleurs System, we believe that GSSD will aid in objectively quantifying the relationship between the shapes of galaxies and other observable parameters, as well as allow rigorous quantified comparison of the shapes and properties of real galaxies vs. simulated ones.

Chapter 1

Introduction

Understanding the sky and space beyond has been one of the oldest forms of study universal to the human race. From the Mayans to the Norse to the Greeks, civilizations across the world have looked up at the stars in awe. The authors of Norse Mythology believed the various planets to be the realms of the warring factions of the gods that controlled all of the world's natural phenomena. The Greeks similarly named the planets after their gods, or their gods after the planets depending on how you interpret their mythology. Meanwhile civilizations with star-based astronomy, such as the Mayans, studied them in such great detail that they devised one of the most accurate calendars of the ancient world-one that rivals the modern Greco-Roman calendar.

Constellations are some of the most basic patterns observed and every culture has different naming schemes for these collections of stars. However, with the advent of telescopes mankind was finally able to see beyond just stars. In the 1900s, we finally discovered cosmological structures that dwarfed stars.

Only after hundreds of years of technological advancement with telescopes were the concepts of galaxies and general groups developed by Edwin Hubble in 1926 [9], under the concept

of galaxy morphology. Categories from his naming scheme-such as elliptical, disk, bar, and spiral galaxies-are still used today. Numerous attempts have been made to expand his categories have been attempted since then, but most authors agree that advancing this field depends on a lot of opinion, with various schemes being designed for different types of predictions.

However this thesis rather than trying to advance astronomy's zeitgeist by grouping galaxies into universal groups, will be trying to contribute the power to quickly, and meaningfully order galaxies based on their visual similarity. To this end I created a tool that allows its user to choose one galaxy, then order the rest from most to least visually similar. This similarity is based on geometric features extracted using SpArcFiRe [4] and is applicable to any galaxy image compatible with said system. The system to perform said task is completely interpretable and shows statistical significance with other properties of galaxies. We focus on their age, star formation rate, and metallicity.

Chapter 2

Background

This chapter describes the tools and concepts that support the construction of our method for finding galaxy similarity called Galaxy Similarity from Sorted Difference, or GSSD, and demonstrate its efficacy. There is a distinct difference between galaxy-to-galaxy comparison and galaxy morphology classification. To contrast the concepts, I will explain what they are first in order to efficiently reference them later. In order to prove the efficacy of GSSD, I will need to explain both the difference between how humans observe galaxies and what telescopes see, and what physical information we can glean from galaxy images. This relationship between physical information and the results of my algorithm will be critical for showing that it is able to quantify observation of differences between galaxies.

2.1 SpArcFiRe

SpArcFiRe is an acronym for Spiral Arc Finder and Reporter [4]. In 2014 Darren R. Davis developed SpArcFiRe for automating the process of detecting and measuring the spiral arms of galaxies. It is able to describe in detail a number of “arm segments” or patches

of geometry that represent spiral arms. A spiral galaxy can be as disordered as a cotton ball or as clean as the arms of a clock. When we describe galaxies that look like cotton candy we call them flocculant, as described in the 1982 paper by [7], meaning having the appearance of wool. When we see galaxies that look like a clock we call them grand design galaxies [12]. Fortunately, even in spaghetti shaped galaxies SpArcFiRe is able to find arm segments and represent them statistically. However, for the purposes of this thesis we will use meta features that describe the aggregate behavior of the arms and not individual arm information. Features such as the total length of the arms, or brightness and contrast of arms are used, but individual arm features are not, as individual arm features were not effective when comparing galaxies with different numbers of arms.

2.2 Sloan Digital Sky Survey

All of the images of galaxies used for this work were retrieved from the Sloan Digital Sky Survey or SDSS for short [11]. We used images from data release 7 [1] and 8 [2] which were taken from their New Mexico facility. The survey consists of millions of galaxies. However, only around 900,000 galaxy objects that can show some kind of structure when parsed by SpArcFiRe. However, of these objects the majority of this thesis was done on a set of around 29,000 identified by Galaxy Zoo one team and used by Darren Davis [4]. SpArcFiRe was designed to detect spiral structure in any size galaxy and was tuned to work well on SDSS. Later in this work when analyzing the results I will analyze how the quality of fit that SpArcFiRe provides affects the results.

2.3 Galaxy Zoo

Galaxy Zoo One, referred to as GZ1, is a citizen science project that released hand-labeled galaxy shape information for almost 900000 SDSS images [13]. For the purposes of this paper, we will use GZ’s values as a ground truth measure for the likelihood that galaxies have spiral structure. From GZ we have two features that combine to form our spiral likelihood metric. These are PCW or the fraction of people who voted a galaxy having a clockwise spiral, and PACW or the fraction of people that voted a galaxy had a counter-clockwise spiral. When we say spirality in this paper, we are referring to the sum of PCW and PACW, meaning the fraction of humans that saw spiral structure at all, regardless of winding direction.

2.4 Galaxy Morphology

Galaxy morphology studies galaxy shape, and aim at classification: to group galaxies into reasonably distinct categories. Galaxy morphology then is just the general shape of a galaxy.

In 1926, Edwin Hubble created the field by introducing the Hubble sequence [9]. He split galaxies into three broad types: elliptical galaxies, which have smooth star distributions with little in the way of structure, spiral galaxies which are galaxies that have distinct arms, giving them a spiral shape not entirely different from a whirlpool, and finally the third type, lenticular galaxies, which have a distinct bright bulge in the center with uniform structureless light spreading out from the center. A diagram describing this can be seen in Figure 2.1

2.4.1 Hubble Sequence

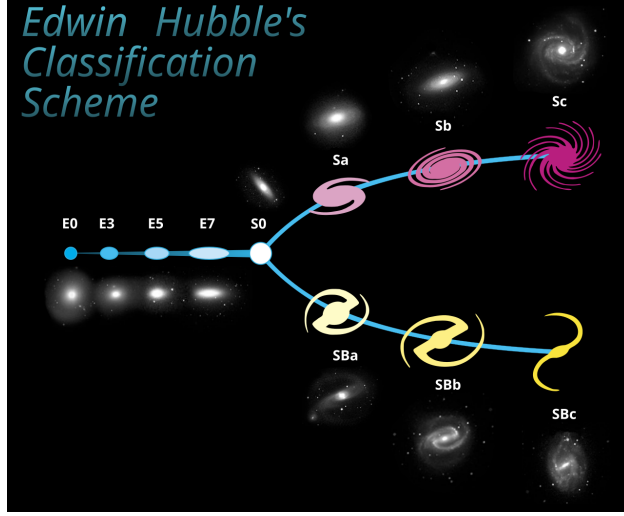


Figure 2.1: A visualization of Hubble’s initial galaxy morphology classification.

In 1959 Gérard de Vaucouleurs added a few new categories of galaxy morphology with the addition of bars and rings [6]. Galaxies with bars are galaxies whose central bulge has the appearance of a linear bar at the center. Ring galaxies have a concentration of stars around a specific radius creating a bright ring inside of the galaxy. Elliptical, spiral, lenticular, barred, and ring galaxies have remained the primary morphological categories with schemes after the De Vaucouleurs scheme rarely adding new keywords to categories. Most schemes classified them based on either hybrid categories, such as a spiral galaxy with a bar at the center, or adding new dimensions of categorization such as whether the edges are closer to Earth, or the face is closer to Earth, called edge-on and face-on galaxies respectively. All of this information is relevant to GSSD because for any galaxy with structure that can be detected by SpArcFiRe, we can find similar galaxies. GSSD identifies similar galaxies independent of galaxy morphology classification. For example, we may find a barred spiral galaxy whose closest matches include both barred and spiral galaxies that exhibit similar structure, even if they don’t fit neatly into traditional morphology classifications. This can be seen in [Figure 2.2](#)

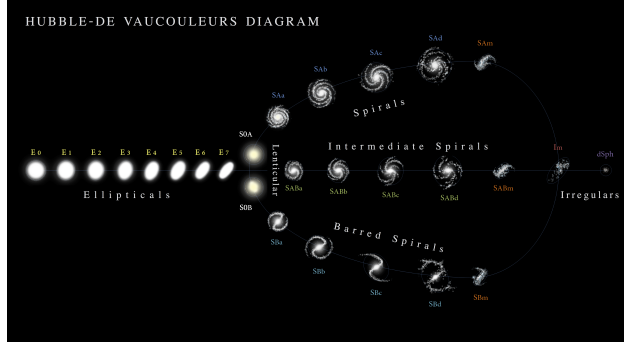


Figure 2.2: Caption

The other contribution that de Vaucouleurs provided was a numbering scheme that represented the galaxy’s morphological class. Known as Hubble stage, or Hubble type T , T values run from -6 to 10. These numbers are important because their general trend of T values is the higher the number, the further the disk extends spatially past the bulge. Generally positive numbers represent spiral galaxies, while negative numbers are elliptical galaxies. In chapter 4 we will extensively depend on T values to validate the similarity ordering presented in this thesis.

2.5 EFIGI

EFIGI is the survey I referenced for the Hubble Sequence T values of galaxies used in this thesis. EFIGI hand labeled 4458 SDSS galaxies by their Hubble Sequence T score. They labeled galaxies from SDSS 4, however I was able to forward translate those same galaxies to SDSS 7. This allowed me to build a set of galaxies matched by their T score and their age, SFR and metallicity.

2.6 Decision Tree Regression

Machine learning is often contentious in the hard sciences because sufficiently complex models-such as convolutional neural networks (CNNs)-can produce results that are difficult to interpret in terms of scientific ground truth [18]. To avoid this problem, I used a Random Forest Regressor which is a method that, while often less accurate than deep CNNs, has easy to interpret weights for the features used in its learning model. Every feature used is assigned a number representing a weight. The greater value the weight is assigned, the more significant it is to the prediction of the learning model. A Random Forest Regressor constructs an ensemble of decision trees, each trained on a randomly resampled subset of the SpArcFiRe features. This bootstrapping approach introduces variation among trees, improving robustness and reducing overfitting. Each tree splits the feature space by branching on values of individual features to make predictions. For example, if SpArcFiRe detects no arms, then the tree structure naturally prioritizes other features, reducing reliance on spiral-related predictions. In Chapter 3 I will discuss the results, weights, and specific methodology used to build the model. It is important to state here, my similarity metric does not use machine learning in generating the metric. I only used machine learning to find which features I should use to determine similarity. Also, I used the weights as guidelines. Using only the highest-weighted features performed worse than manually selecting a well-balanced subset.

2.7 Tested parameters

Most of the galaxy morphology schemes are relatively subjective and iterative improvements to the point that new work in the field is mostly being done with deep learning models. To answer the natural question, “why is your claimed similarity metric actually representative of similar galaxies” we showed that my groups of similar galaxies have a strong correlations

with a galaxy’s age, star formation rate, and metallicity. I will provide a brief explanation of each of these three parameters now.

2.7.1 Age

Age is relatively straightforward. Age, as used by SDSS, refers to the mean stellar age of galaxy [10], [17], or put more simply the average age of observable stars in a galaxy. For our purposes we used a static set of galaxies whose ages were all labeled by the same sources. According to Kauffman [10] and Tremonti [17], stellar population ages are estimated through spectral fitting and photometric analysis. We won’t be predicting age here, but Age-whenver we capitalize age we are referring to age from [10] or [17]- nevertheless is mostly distinct from what SpArcFiRe “sees” which is spiral arm shape.

2.7.2 Star Formation Rate

Star Formation Rate (SFR) is a metric that quantifies how actively a galaxy is forming stars. Younger galaxies tend to have high SFRs since they are still full of gas and dust. As stars form the observed amount of gas and dust decreases such that older galaxies, especially elliptical galaxies, tend to have low star formation rates. In our paper we will use the SFR values for SDSS given by Tremonti in their 2004 paper [17].

2.7.3 Metallicity

In 2004 Christy A. Tremonti referred to [17] metallicity as “... the gas reprocessed by stars and any exchange of gas between the galaxy and its environment.” Hydrogen and helium were produced during the Big Bang, while heavier elements-collectively called “metals” in

astronomy-are synthesized in stars and dispersed into the interstellar medium through supernovae and stellar winds. To calculate the exact value of metallicity, you need to measure the amount of mass that is not hydrogen or helium then divide that by the total mass of the star. For example Sun's mass is made up of about 2% elements that are not helium or hydrogen. Therefore the sun's metallicity is about 0.02. Again this is a solely star based metric and is relatively distinct from general shape.

Chapter 3

Determining Similarity

Galaxy similarity is abstract, subjective, and complex with regards to observed features of galaxies. We define a pairwise similarity measure that estimates the shape similarity between two galaxies as measured by differences in a specific set of SpArcFiRe output parameters. Given the measure across all galaxies and a target galaxy, we can rank galaxies in terms of their similarity to that target. We developed an automated method for ranking galaxies by similarity, showing that as computed differences increase, visual dissimilarity also tends to increase. This chapter will cover how I developed a similarity measure. We first “standardize” each feature by subtracting the mean across galaxies and then dividing by the standard deviation. The result is that for each feature, the distribution of that feature across galaxies is approximately Gaussian, with mean 0 and standard deviation 1. Then, given a pair galaxies g_1 and g_2 that have feature f values g_1^f and g_2^f , we define their difference as:

$$D(g_1, g_2) = \max_f \left| g_1^f - g_2^f \right| \tag{3.1}$$

Finally, we say that g' is galaxy g 's most similar neighbor if g' has the smallest difference from g across all galaxies.

First I will cover the general method by which our similarity ordering works. Then I will cover how I arrived at the current set of features. I will cover how I leveraged machine learning on a set of galaxies hand classified by morphology. Then I will explain how after using these automated methods I by hand improved the feature choices to fix some obvious problems and ended up at my current feature set.

3.1 Developing the similarity order

GSSD was first developed by first figuring out how galaxies are most visually different. First we needed to find what features represent a galaxy's visual appearance. The features used by our method are curated from weights used to predict a galaxy's morphology. Using these features, we then calculated the maximum difference among relevant features between two galaxies using equation 3.1. Finally after we had our set of maximum galaxy differences, we sorted by the least different galaxy. Or if we wanted to say this another way, we find a local maximum feature difference, then we sort by the minimum local maximum difference. However we do not just take the maximum difference from raw feature values, but instead first normalize them. This ensures features such as number of arms, which generally falls within a single digit range, will have an equivalently weighted value as pitch angle whose value may contain up to three digits. The normalization scales were based on the range of values from our chosen galaxies from across all of the Sloan Digital Sky Survey's data release 7, outside of our 29000 set of selected galaxies.

To see an example of how this similarity ordering works in practice please see Figure 3.1. The image is ordered from top to bottom with the top image being the target galaxy, and

every lower galaxy being the next nearest neighbor. The left column shows the original images, the center column shows SpArcFiRe-generated deprojections (face-on views), and the right column shows the cluster masks—representations of the arm segments SpArcFiRe “sees”. There will be a few orderings in this paper that don’t appear to be good fits until you look at the cluster mask, as the cluster mask is where a large number of the SpArcFiRe features are calculated.

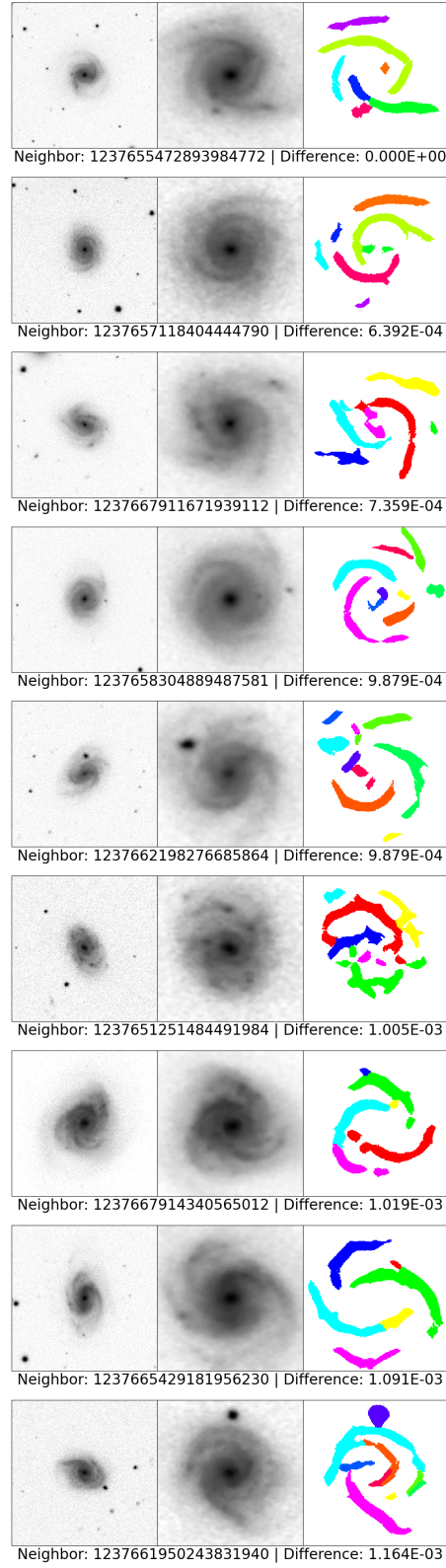


Figure 3.1: An example galaxy ordering that shows how similar looking galaxies branch out based on visual similarity between multiple classes of galaxies

3.1.1 Morphology Selection

Data Selection

Galaxy morphological classification, as referenced earlier, is a well established field in astronomy [9] [6]. As such, the EFIGI catalogue [3] was a straightforward place to start as a dataset for feature selection. It is a hand labeled set of galaxies from the Sloan Digital Sky Survey (SDSS). This is the same set that the largest dataset for SpArcFiRe has been processed so it was a natural choice as a training set for developing features. For the purposes of this paper when we reference a galaxy’s morphological class we will be drawing it from the EFIGI dataset. EFIGI scores a galaxy’s general morphology by its Hubble type listed in their data release as T. A distribution of our labeled set of galaxies is included in figure 3.2

3.1.2 Classification Background

In order to build an interpretable and unbiased set of features as a baseline I used random forest regression using the SpArcFiRe output for the set of galaxies listed in EFIGI as the training features, and predicted their T values. Only 809 galaxies between were found in both our GZ labeled spiral galaxies and our T labeled EFIGI galaxies. The distribution of those galaxies is displayed on figure 3.2. The regressor was implemented in Python using the scikit-learn library[14]. The specific method was `sklearn.ensemble.RandomForestRegressor` and default settings were used. In an apples to apples comparison, Alexandre Gauthier et al. tested random forest regression using 109 galaxies and raw pixel information. They only had 109 galaxies and did their best to tune their model. Rather than predict T score from EFIGI they predicted galaxy zoo classes. Their accuracy under 10-fold cross validation was 0.67[8]. Our accuracy under 10-fold cross validation was 0.69, however this improvement comes with caveats. We used almost 8 times as much data and we had the benefit of using

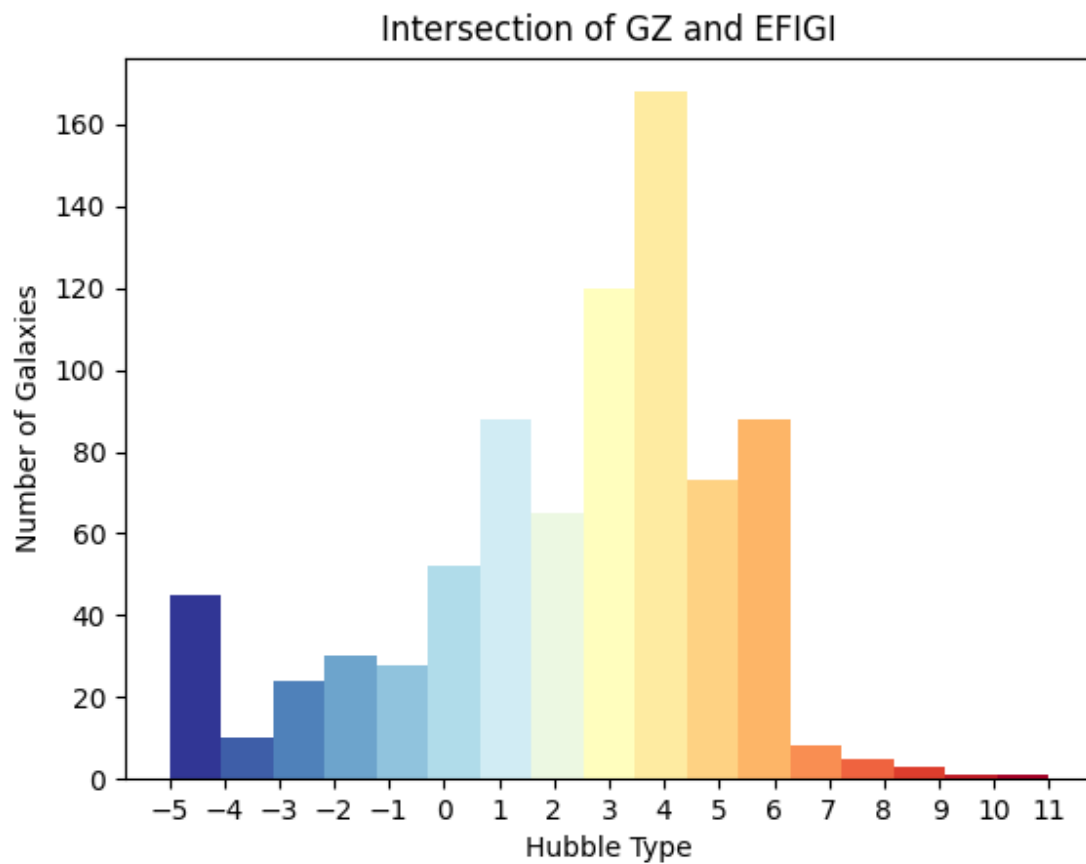


Figure 3.2: Hubble score T for all 809 galaxies that overlap from our SpArcFiRe trained GZ set and EFIGI

SpArcFiRe. Needless to say, this test was just to prove that the weights had meaningful predictive power.

3.1.3 Feature Weights

We selected features based on their importance scores from the trained random forest model. We did this by selecting the most important features that were both, not artifacts of the SpArcFiRe arc generation process, and not similar to previously chosen features. These features are all SpArcFiRe features and will have a dedicated section of analysis on how they affect the results later in chapter 4. Listed in order, the highest-weighted SpArcFiRe features generated from the random forest can be found in table 3.1. In the next subsection I will give definitions for the feature used, as well as an explanation why we changed the features from the strict tree ordering plus filtering process described above.

Tree Feature Descriptions

totalArcLength: Stands for Total Arc Length. This feature is the total length of all arcs found by SpArcFiRe for a given galaxy.

pa_alenWtd_avg__abs: Stands for pitch angle arc length wtd average absolute value. This feature is the average pitch angle across all arm segments weighted by length of the arc.

diskAxisRatio: Stands for disk axis ratio. The disk axis ration gives the ratio of the minor axis length to the major axis length.

brtUnifScore: Stands for brightness uniformity score. It represents the score give to measure how uniform brightness is on average along arcs withing the galaxy.

bar_cand_score_0rifld: Stands for bar candidate score orientation field: It is SpArcFiRe's internal score that the galaxy has a bar based on information from the orientation field.

alenAt25pct: Stands for arc length at 25 percent. This measures the amount of arc frag-

mentation. It does so by laying all of the arc segments end to end sorted in decreasing order by arc length, reaching the total arc length of the galaxy. It then gives the length of the arc present at 25 percent of the total arc length.

diskMinAxsLen: Stands for disk minor axis length. It gives the length of the minor axis in the input image.

contourBrtratio: Stands for contour brightness ratio. It is the ratio between the inner and outer countours generated by SpArcFiRe. Countours are ellipses generated to find the boundaries of the galaxy and assist in the ellipse fit portion of SpArcFiRe.

Curated Features

The tree-weighted features produced a reasonable baseline. However, there were still some common problems. Our orderings were often disrupted by flocculent galaxies. To address this **rankAt75pct** was added. **rankAt75pct** stands for rank at 75 percent. It is the rank of the arm located at 75 percent of the total arc length described earlier in the definition of **ale-nAt25pct**. Under the recommendation of professor Hayes **diskMinAxsLen** was replaced with **diskMajAxsLen**, due to the major axis of the galaxy giving more relevant shape information. **diskMajAxsLen** Stands for disk major axis length, and gives the length of the major axis of the input galaxy image.

3.2 Normalization

To ensure each feature contributed equally, values were normalized from their original ranges to a 0–1 scale. Normalization parameters were derived from nearly 600,000 galaxies in SpArcFiRe’s 2016 data release. We applied 2-norm normalization to each feature column.

$$x_{\text{norm}} = \frac{x}{\|x\|_2}$$

3.3 Minimum local maximum difference

As stated before, we define similarity here as two galaxies are similar when all other galaxies are more different than the two galaxies being compared. To do this we use three simple steps given the data we have calculated so far. First step is to choose a galaxy a "target" galaxy. The next step is for each galaxy to calculate the maximum difference between any of the normalized features of the two galaxies $D(g, g_i) = \max_f |g^f - g_i^f|$. Once we compute all maximum feature differences for a target galaxy, we sort the list from smallest to largest. The most similar galaxy appears first; the least similar, last.

Feature	Weight
totalArcLength	6.13×10^{-1}
pa_alenWtd_avg__abs	4.86×10^{-2}
diskAxisRatio	3.92×10^{-2}
bar_cand_score_orifld	3.73×10^{-2}
brtUnifScore	3.43×10^{-2}
alenAt25pct	2.01×10^{-2}
diskMinAxsLen	1.71×10^{-2}
contourBrtratio	1.31×10^{-2}
pa_avg__abs	1.27×10^{-2}
bar_half_length_standardized_img	1.26×10^{-2}
bar_score_img	1.14×10^{-2}
wtdLik	1.11×10^{-2}
bar_half_length_input_img	1.05×10^{-2}
alenAt50pct	1.02×10^{-2}
diskMajAxsLen	9.06×10^{-3}
avgArcLength	9.02×10^{-3}
muDistProp	7.74×10^{-3}
muDist	7.64×10^{-3}
paErr_alenWtd_stdev_domChiralityOnly	7.45×10^{-3}
likOfCtr	6.79×10^{-3}
numArcsGE040	6.29×10^{-3}
numElpsRefits	4.37×10^{-3}
numArcsGE020	3.21×10^{-3}
numDcoArcsGE010	3.03×10^{-3}
numDcoArcsGE020	2.86×10^{-3}
numDcoArcsGE000	2.79×10^{-3}
numArcsGE010	2.67×10^{-3}
numArcsGE000	2.56×10^{-3}
numDcoArcs_largest_length_gap	1.84×10^{-3}
totalNumArcs	1.80×10^{-3}
numDcoArcsGE040	1.80×10^{-3}
numArcs_largest_length_gap	1.72×10^{-3}
rankAt75pct	1.70×10^{-3}
numArcs_arclength_function_flattening	1.66×10^{-3}
numDcoArcs_arclength_function_flattening	1.64×10^{-3}
numDcoArcsGE060	1.53×10^{-3}
numArcsGE085	1.43×10^{-3}
numDcoArcsGE100	1.25×10^{-3}
numArcsGE060	1.23×10^{-3}
numArcsGE080	1.04×10^{-3}
numDcoArcsGE075	1.03×10^{-3}
rankAt50pct	1.01×10^{-3}
numDcoArcsGE120	1.00×10^{-3}
numDcoArcsGE085	9.76×10^{-4}
numArcsGE090	9.41×10^{-4}
numDcoArcsGE090	9.32×10^{-4}
numArcsGE075	8.63×10^{-4}
numArcsGE100	8.27×10^{-4}
numArcsGE120	7.01×10^{-4}
numDcoArcsGE160	6.61×10^{-4}
numArcsGE140	6.19×10^{-4}
numArcsGE095	5.97×10^{-4}
numDcoArcsGE080	5.83×10^{-4}
rankAt25pct	5.67×10^{-4}
numArcsGE160	5.31×10^{-4}

Table 3.1: SpArcFiRe Feature Importance sorted by weight with 3 Significant Digits. Some redundant features not included.

Chapter 4

Quality Analysis of Similarity

Similar galaxies in this thesis are galaxies that are less different than other galaxies. This means that for every galaxy targeted by our algorithm, it should be straightforward to rank all other galaxies from most similar to most different. In the previous section we covered the method by which we developed the features used as well as how they were used to find a maximum difference score between two galaxies, a neighbor and their target. Finally the sorting by feature difference led to the similarity order. However, the accuracy of our similarity score has a few bottlenecks as we will show later. In order to test the robustness of GSSD(Galaxy Similarity from Sorted Difference) I will show where the algorithm works, where it fails, and how it works on average. Then I will break down feature by feature all of the problems it overcomes to show visually our results are reliable even when applied to adverse datasets.

4.1 Building an Ideal Set of Galaxies

One of the major limiting factors of our algorithm’s performance is that it’s only as good as the accuracy of our features. What I mean by that is that SpArcFiRe has some areas where it is known to have limited results[4]. It also doesn’t fare well on flocculent or morphologically ambiguous galaxies. As stated by Davis and Hayes (2014) [4], SpArcFiRe reads arm segments then merges them into arms. So, necessarily, flocculent arms are especially challenging to fit correctly. Also, arms can be obscured by dust, visually splitting them into multiple segments. In order to find a representative set of galaxies that minimizes the natural limitations of SpArcFiRe and thus produce the clearest representation of how features affect the similarity results, we filtered galaxies by three specific criteria: spirality, image size, and Hubble type. We discuss these three filters in the next sections.

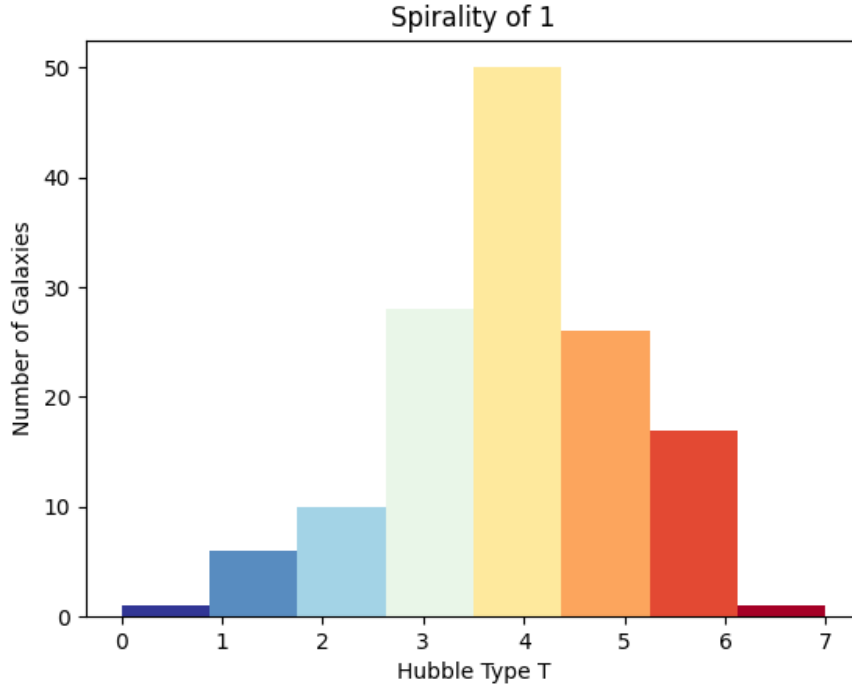


Figure 4.1: Hubble score T for all 139 galaxies that overlap from our SpArcFiRe trained GZ set and EFIGI whose spirality is equal to 1, meaning the citizen scientists of galaxy zoo were unanimous the galaxy was spiral.

4.1.1 Step 1. Filter by Spirality

First, chose galaxies with GZ spirality score of 1. This means that 100% of the citizen scientists voting in GZ1 agree that arms were visible, regardless of winding direction. Of the galaxies in our 809 galaxy overlap between EFIGI and GZ, only 139 had a spirality of 1.(Figure 4.1). This also has the side effect of limiting our choices of Hubble types from 0 to 7.

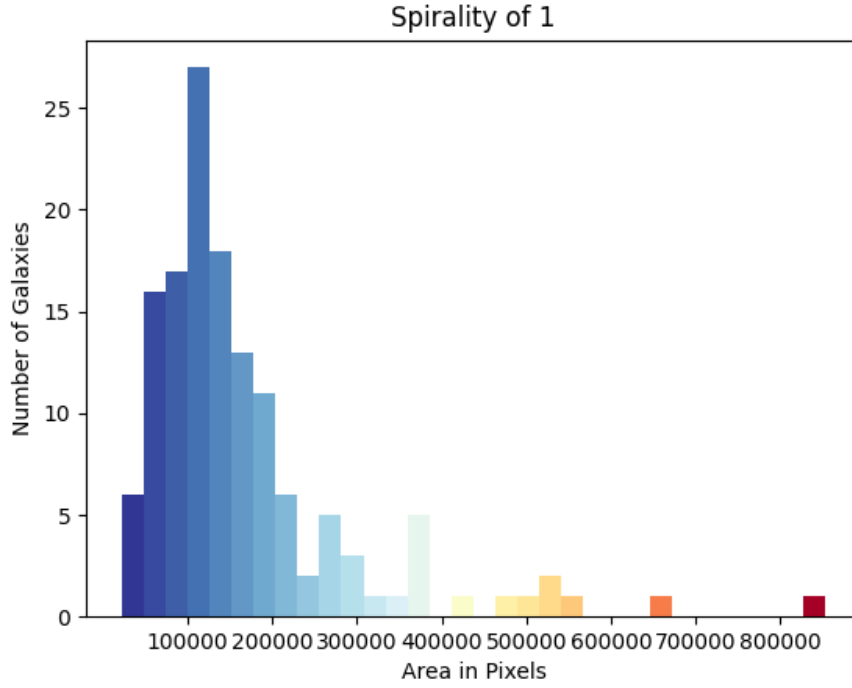


Figure 4.2: Area in pixels for all galaxies with spirality of 1 from the GZ1 EFIGI union sets.

4.1.2 Step 2. Filter by Size of the Image

Next I needed to find galaxies that had manageable resolution. To do this I filtered our set of galaxies to only include those above a specific size threshold. At 100,000 square pixels we retain all eight Hubble T types ($T = 0$ to $T = 7$), as shown in Figure 4.2. When I apply all of these filters we are now reduced to 99 galaxies in figure 4.2.

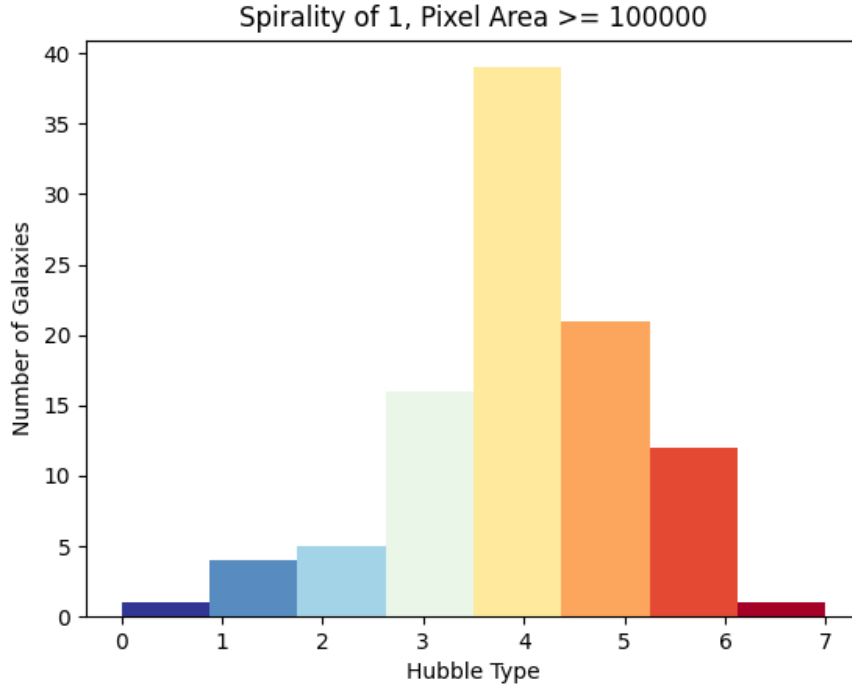


Figure 4.3: Hubble T score for 99 galaxies filtered by spirality 1 and size ≥ 100000 square pixels from the GZ EFIGI Intersection Sets.

4.1.3 Step 3. Least Different set for each Hubble Type

Finally, for each of our remaining Hubble Types from range 0 to 7 I chose the galaxy with the smallest GSSD for their 10th neighbor. As shown in 4.3 At T=1 and T=7 we had only had 1 galaxy that fit our selection criteria, therefore I picked our only option rather than our most similar ordering for those two T scores.

4.1.4 Curated Good Reference Set

Figure 4.4 is a the result of our filtering to representative results based on filtering criteria rather than cherry picking arbitrary good orders. Please note, these 99 galaxies are just targets we filtered down to. We still choose galaxies from the 29000 galaxies in order to select neighbors.

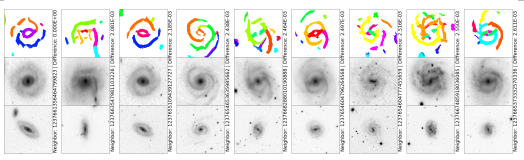
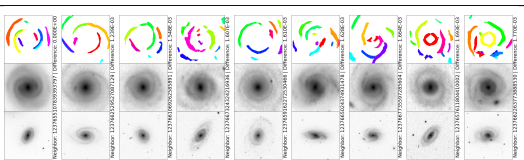
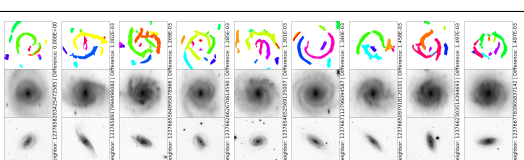
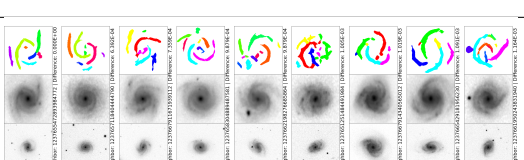
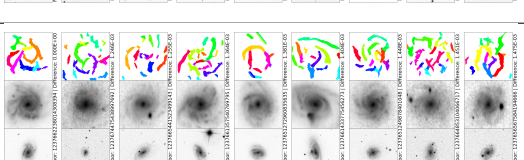
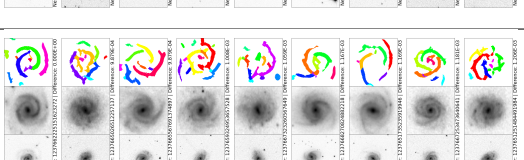
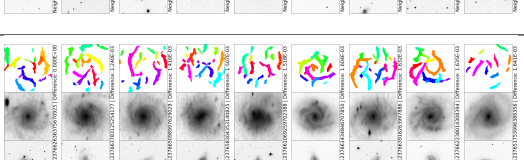
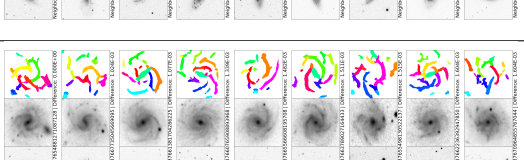
	T = 0
	T = 1
	T = 2
	T = 3
	T = 4
	T = 5
	T = 6
	T = 7

Figure 4.4: Curated examples of galaxy similarity for Hubble T-types 0 through 7. Each row shows the selected image most similar to the target for the respective T-type.

4.2 Where the algorithm is limited

The previous section covered how we chose a set of galaxies that we hope will perform well, and this section will cover where it struggles. There are categorically two ways GSSD consistently fails. First, if the input images themselves are flawed or have artifacts, SpArcFiRe also does poorly. Second is when SpArcFiRe fits poorly to input images. As an example, in Figure 4.5 we show what happens when we expand our selection criteria outside of our curated set and images with artifacts join the GSSD search space.

Figure 4.6 shows examples of artifacts that cause poor SpArcFiRe fits. Davis found in his 2014 work [4] found that SpArcFire had trouble finding arms with very low surface brightness. Figure 4.6 depicts such a galaxy, and SpArcFiRe detects the endpoints of a short bar, and only one side of the ring, resulting in the pixel mask looking like a "smiley face". As a result of the target galaxy's bad fit, the list of "similar" galaxies is unfortunately rendered meaningless. It matched what it saw as a frowning face as the galaxy's nearest neighbor. SpArcFiRe as we mentioned earlier, doesn't fit every image of a galaxy perfectly. In the previous sections we attempted to filter galaxies to choose the best fit possible in order to avoid this situation where even if there is nothing wrong with the input image, sometimes SpArcFiRe just has a bad fit.

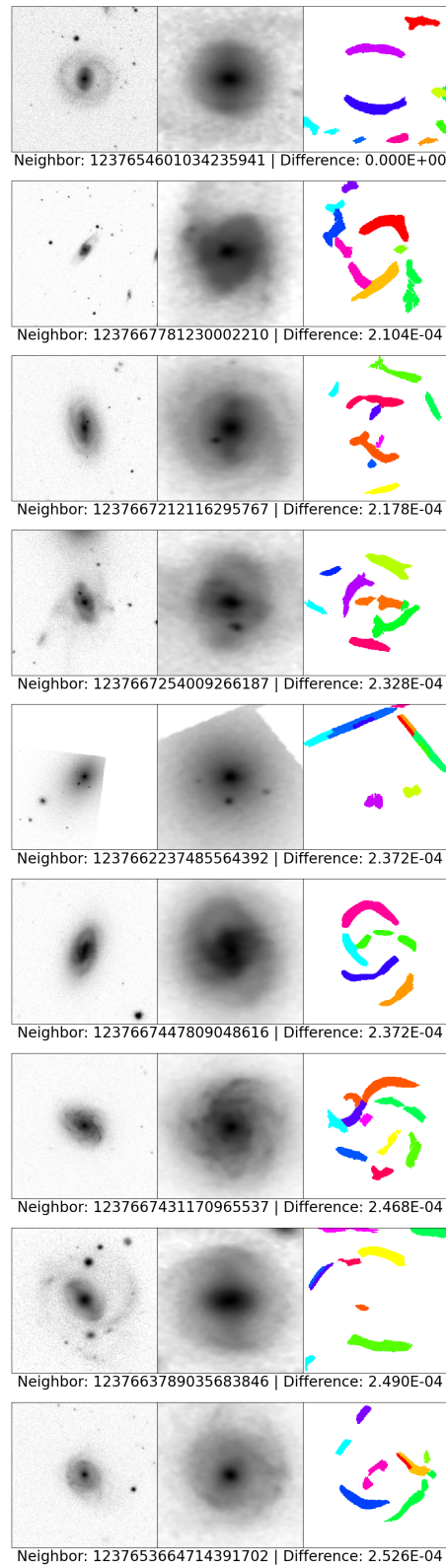


Figure 4.5: Example of a well fit target with bad neighbors

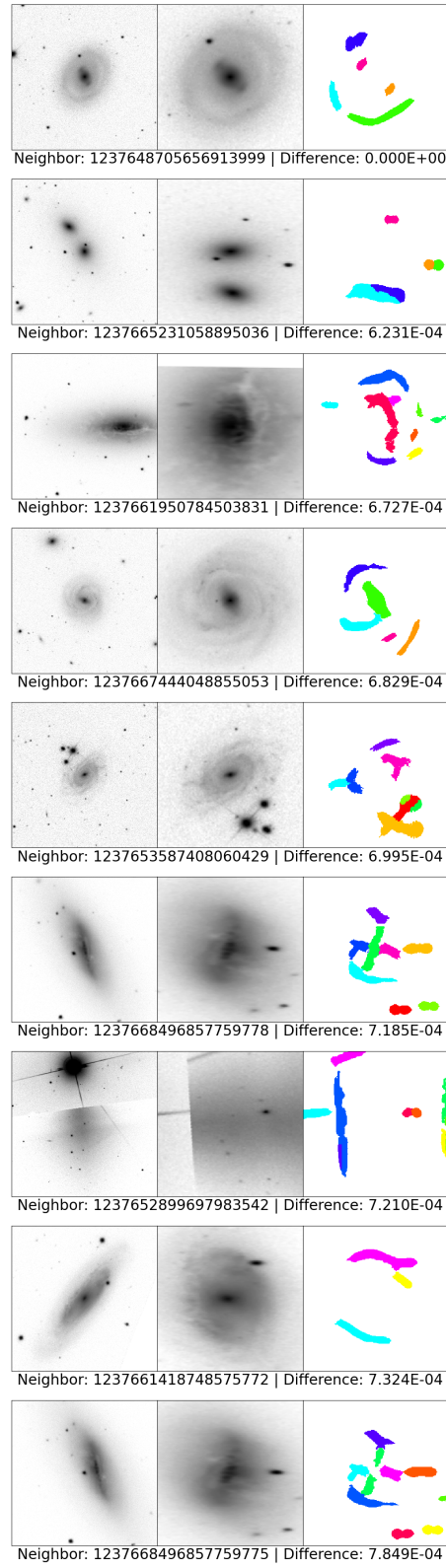


Figure 4.6: Example of a badly fit target with bad neighbors

4.3 Expected Results

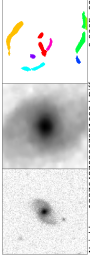
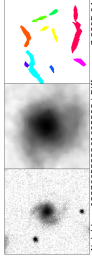
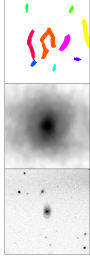
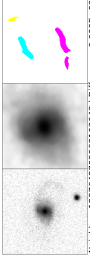
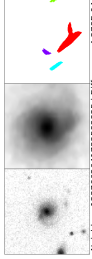
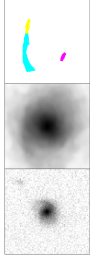
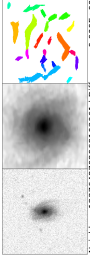
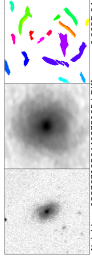
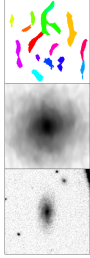
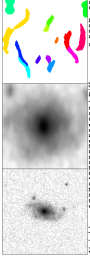
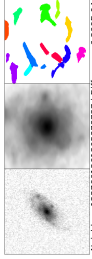
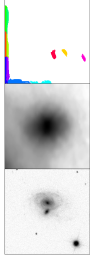
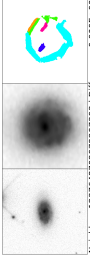
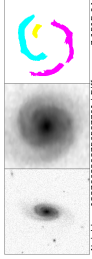
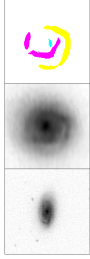
So far we have selected galaxies that explicitly exist within the intersection of our GZ and EFIGI galaxies. In order to showcase how well the results are when trained on average galaxies that SpArcFiRe is known to function correctly on, I will show a few samples of galaxies not chosen from where EFIGI and GZ overlap, but instead from the "Nice" set of approximately twenty nine thousand galaxies referenced by Darren Davis in his SpArcFiRe paper [4]. This was done to show that galaxies were SpArcFiRe should have good but average performance still mostly produce sets of galaxies that look similar.

Table 4.1: Randomly chosen galaxies from the 29,000 used in [4]

Image										Label
										0
										
Neighbor: 12376487029698820 Difference: 0.000E+00	Neighbor: 1237665129607528464 Difference: 5.590E-04	Neighbor: 12376615080505050 Difference: 6.493E-04	Neighbor: 12376932656580841 Difference: 6.951E-04	Neighbor: 123766711059331171 Difference: 7.005E-04	Neighbor: 1237662262178807906 Difference: 9.879E-04	Neighbor: 123766773633797051 Difference: 9.879E-04	Neighbor: 123765324470657781 Difference: 9.879E-04	Neighbor: 123766112454056628 Difference: 9.879E-04	Neighbor: 123766112454056628 Difference: 9.879E-04	
										
										
Neighbor: 1237648702967054406 Difference: 0.000E+00	Neighbor: 123766872445507682 Difference: 4.750E-04	Neighbor: 123766735581884592 Difference: 6.000E-04	Neighbor: 123766735564050811 Difference: 7.191E-04	Neighbor: 1237655469834809451 Difference: 7.570E-04	Neighbor: 123766219934870907 Difference: 8.129E-04	Neighbor: 123765330909704386 Difference: 8.199E-04	Neighbor: 1237667429031473401 Difference: 8.733E-04	Neighbor: 1237667537475534941 Difference: 8.835E-04	Neighbor: 1237667537475534941 Difference: 8.835E-04	

Continued on next page

Table 4.1 – continued from previous page

Image			Label
 Neighbor: 1237648702967251093 Difference: 0.000E+00	 Neighbor: 12376655261095598 Difference: 7.389E-04	 Neighbor: 123765605805981623 Difference: 7.405E-04	2
 Neighbor: 1237648702968103043 Difference: 0.000E+00	 Neighbor: 1237662441732583825 Difference: 5.278E-04	 Neighbor: 1237664880490512429 Difference: 7.789E-04	3
 Neighbor: 1237648702968103061 Difference: 0.000E+00	 Neighbor: 123761956440875195 Difference: 6.802E-04	 Neighbor: 1237668495776241746 Difference: 7.430E-04	4
 Neighbor: 1237648702971511000 Difference: 0.000E+00	 Neighbor: 1237661361305019737 Difference: 6.399E-04	 Neighbor: 123765221488414483 Difference: 6.101E-04	5
 Neighbor: 123764870297297393 Difference: 0.000E+00	 Neighbor: 1237662309756587701 Difference: 7.151E-04	 Neighbor: 123766743173931286 Difference: 9.382E-04	6

Continued on next page

Table 4.1 – continued from previous page

Image										Label
										7
Neighbor: 123767764786020385 Difference: 0.000E+00			Neighbor: 123766614976510752 Difference: 2.938E-03			Neighbor: 12376581305072806 Difference: 2.898E-03			Neighbor: 123765348350746672 Difference: 3.125E-03	
										
Neighbor: 1237667764786020385 Difference: 0.000E+00			Neighbor: 123766614976510752 Difference: 2.938E-03			Neighbor: 12376581305072806 Difference: 2.898E-03			Neighbor: 123765348350746672 Difference: 3.125E-03	
										
Neighbor: 1237667764786020385 Difference: 0.000E+00			Neighbor: 123766614976510752 Difference: 2.938E-03			Neighbor: 12376581305072806 Difference: 2.898E-03			Neighbor: 123765348350746672 Difference: 3.125E-03	
										
Neighbor: 1237667764786020385 Difference: 0.000E+00			Neighbor: 123766614976510752 Difference: 2.938E-03			Neighbor: 12376581305072806 Difference: 2.898E-03			Neighbor: 123765348350746672 Difference: 3.125E-03	
										
Neighbor: 1237667764786020385 Difference: 0.000E+00			Neighbor: 123766614976510752 Difference: 2.938E-03			Neighbor: 12376581305072806 Difference: 2.898E-03			Neighbor: 123765348350746672 Difference: 3.125E-03	

4.4 How Feature Removal Affects Rankings: Illustrated by Hubble Type

In this section I show feature by feature the results of removing it from my similarity algorithm. This sections goal will be to provide visual evidence that every feature contributes to the goal of proving the visual difference of galaxies. Some features do not make a difference for some morphological orderings, however every feature will be shown to make a large difference for at least one type of spiral galaxy. For each galaxy from the control set representing a different Hubble type, I will first show that galaxy with all of the features included in the difference calculation. Then I will show 8 images, one for each feature removed in order to showcase the sensitivity of each Hubble type to each feature. Please refer to our table at the end of [3](#) page [11](#) to see how each feature was weighted based on relevance to galaxy morphology.

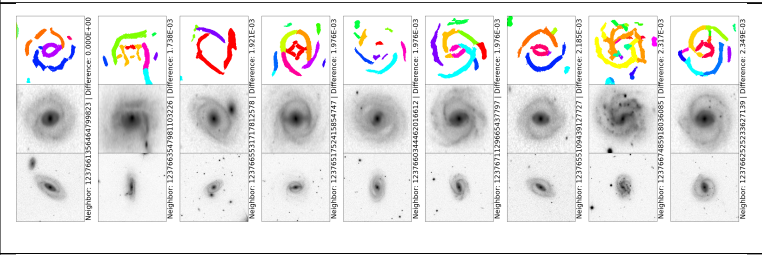
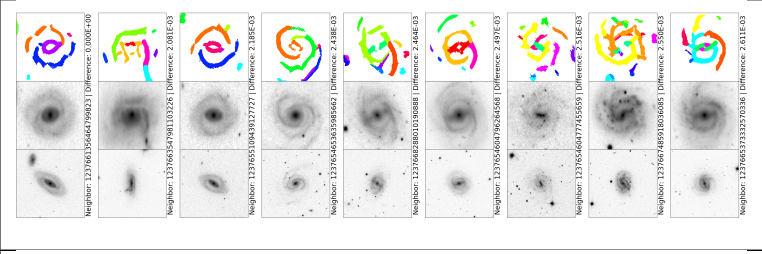
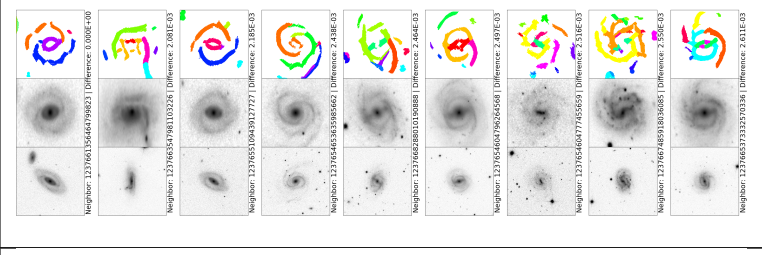
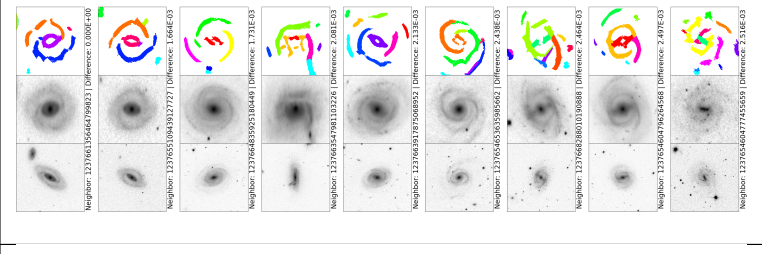
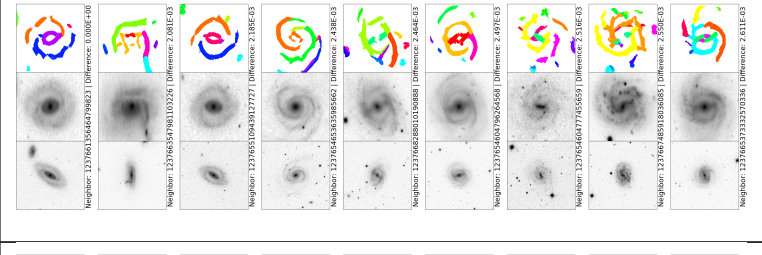
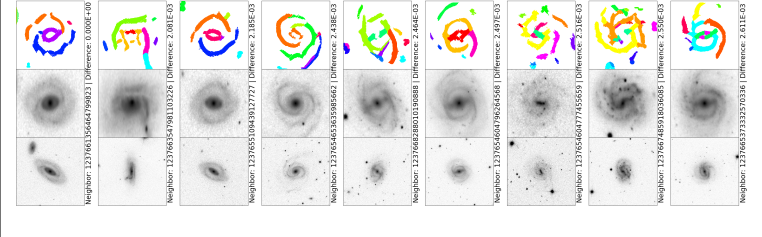
4.4.1 Hubble Type T 0

Table 4.2: Visual comparisons of ordering results for Hubble type T=0 with individual SpArcFiRe features removed.

Image	T-Type / Removed
	T = 0 (All features)
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore
	contourBrtrRatio

Continued on next page

Table 4.2 – continued from previous page

Image	T-Type / Removed
	alenAt25pct
	rankAt75pct
	totalArcLength
	pa_avg__abs
	pa_alenWtd_avg__abs
	bar_cand_score_orifld

Commenting on Figure 4.2:

Control Group: This group is representative of eye shaped bulges with tightly wound arms that form an outer ring. While some of the most similar galaxies to the target in Figure 4.10 are missing the hallmark eye and outer ring, at least half of them have it. Due to GSSD's sensitivity to artifacts it appears that since both galaxies are noisy in the parts of geometry that define the cluster, they were ordered improperly. Neighbor 2 has a fuzzy bulge as well as dust obscuring the arms. The way I interpret this is that as galaxies look less and less similar, the algorithm starts choosing between similar arms and bulges as the neighbors travel further and further away from the target.

Removed diskAxisRatio: When diskAxisRatio is removed the ordering appears to be over dependent on arm length and arm count. Galaxies 2 and 3 are grand design. Neighbors don't return to eye bulge rings until the fourth and 10th galaxy.

Removed brtUnifScore: Removing brtUnifScore appears to cause the clustering to focus on the gradient of light within the arms. Only one of the original eye shaped bulges remains within the set.

Removed brtUnifScore: Removing brtUnifScore appears to cause the clustering to focus on the gradient of light within the arms. Only one of the original eye shaped bulges remains within the set.

Removed contourBrtratio: The order is unchanged.

Removed alenAt25pct: This order over fits the bulge. Instead of matching 2 tightly wound almost ringlike arms, it just matches rings and bars earlier in the order.

Removed rankAt75pct: The order is unchanged

Removed totalArcLength: The order is unchanged

Removed pa_avg__abs: The ordering stops distinguishing the clear brightness contrast around the central bulge.

Removed pa_alenWtd_avg__abs: The order is unchanged

Removed bar_cand_score_orifld: The order is unchanged

4.4.2 Hubble Type T 1

Table 4.3: Visual comparisons of ordering results for Hubble type T=1 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 1 (All features)
	diskAxisRatio
	diskMajAxsLen

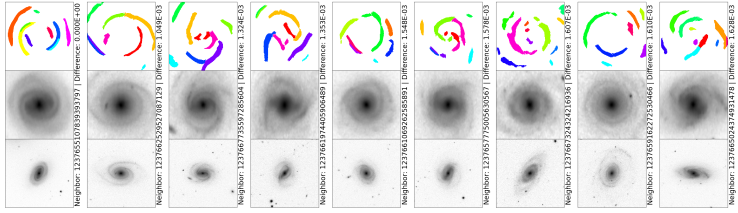
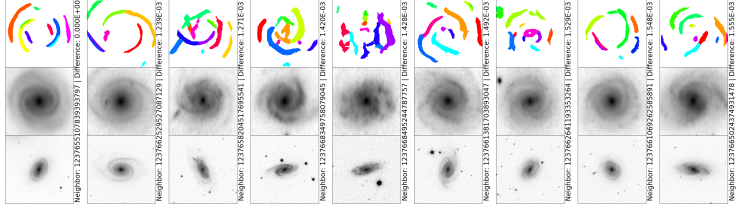
Continued on next page

Table 4.3 – continued from previous page

Image	T-Type / Feature
	brtUnifScore
	contourBrtratio
	alenAt25pct
	rankAt75pct
	totalArcLength
	pa_avg__abs

Continued on next page

Table 4.3 – continued from previous page

Image	T-Type / Feature
	pa_alenWtd_avg__abs
	bar_cand_score_orifld

Commentary on Table 4.3

Control Group: The general shape fit in this set of galaxies is a large dark bulge with two tightly wound, smooth, wide arms. There is little empty space in between the arms as they spin outwards. As neighbors travel further away, more arms are added, or instead of tightly wound arms, rings with as similar brightness distribution are substituted.

diskAxisRatio: The emphasis on smooth arms suffers. Bulges are no longer surrounded by smooth spirals, but by lumpy arms or blotchy light patterns.

diskMajAxsLen: Arms begin to have more light between their curves, and bulges often resemble bars.

brtUnifScore: Ring galaxies (e.g., at position 9) move up to earlier ranks. Arms are more emphasized, and the central bulge contrast disappears.

contourBrtrRatio: The ordering is unchanged.

alenAt25pct: Smoothness constraints are relaxed. Multiple small arm segments of similar shape are now favored.

rankAt75pct: The number of arms is less strictly enforced. Matched galaxies are more flocculent than in the original set.

totalArcLength: Overfits the central bulge. Arms are largely ignored if the general light gradient resembles the original.

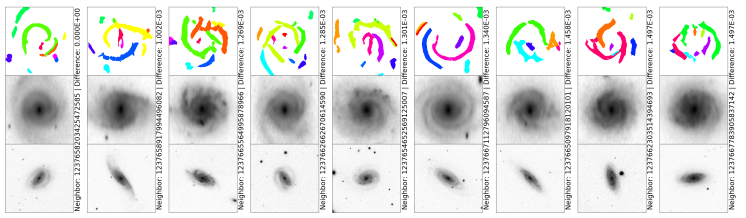
pa_avg__abs: The final galaxy in the ordering shows 3 fuzzy arms instead of 2 defined ones.

pa_alenWtd_avg__abs: Bar galaxies appear in the ranking, but most still retain 2 smooth arms, except near the end.

bar_cand_score_orifld: Arm count and definition are reduced in importance. Fuzzy-armed galaxies appear, but bars are still mostly avoided and bulge shape is retained.

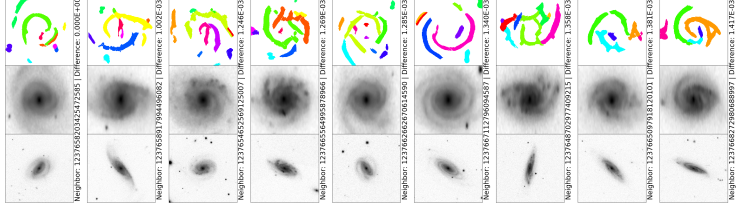
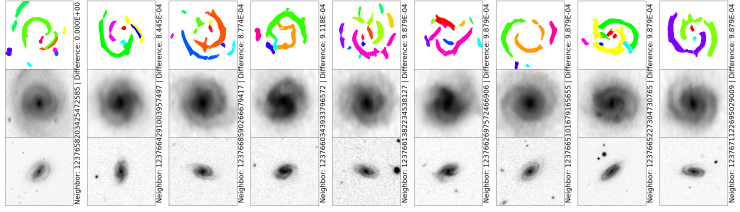
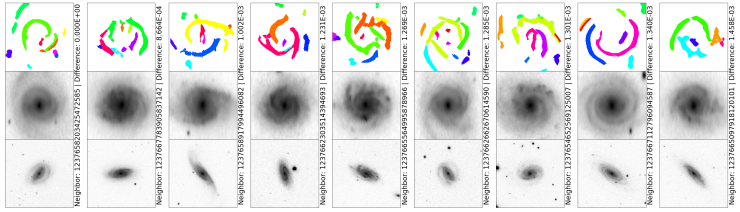
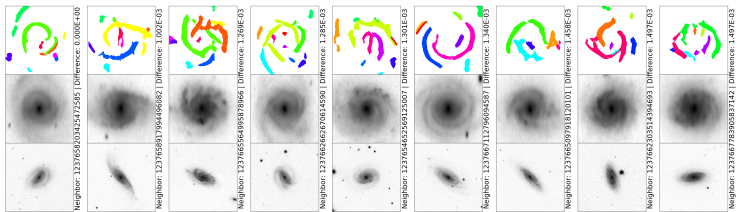
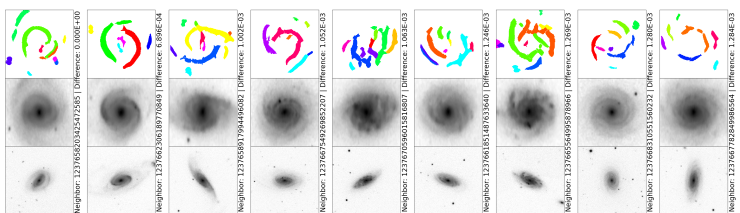
4.4.3 Hubble Type T 2

Table 4.4: Visual comparisons of ordering results for Hubble type T=2 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 2 (All features)

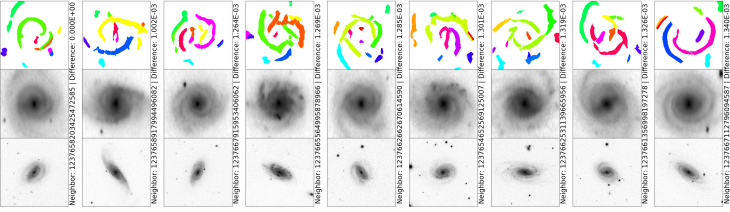
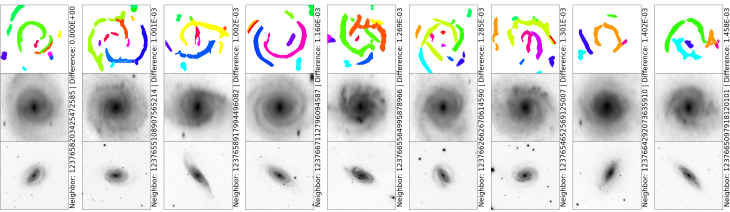
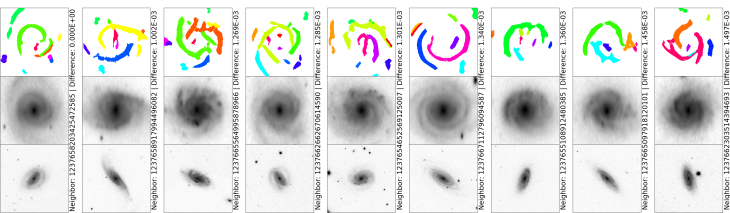
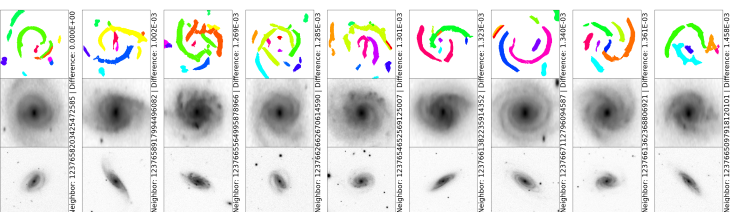
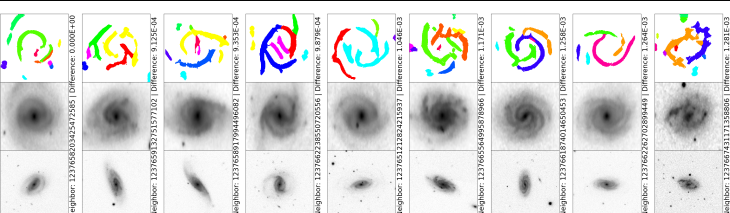
Continued on next page

Table 4.4 – continued from previous page

Image	T-Type / Feature
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore
	contourBrtrRatio
	alenAt25pct

Continued on next page

Table 4.4 – continued from previous page

Image	T-Type / Feature
	rankAt75pct
	totalArcLength
	pa_avg_abs
	pa_alenWtd_avg_abs
	bar_cand_score_orifld

Commentary on Table 4.4

Control Group The general shape of this cluster appears to be tightly wound arms around an inner bulge, either a large bright bulge or a bulge with inner arms. From this inner circle, two long arms extend outward. As the galaxies increase in difference, the central bright region becomes more chaotic; however, the trend of two long arms is maintained.

diskAxisRatio: The emphasis on the bright central region is reduced, and the emphasis on the two long arms increases.

diskMajAxsLen: The inner region is projected as the entire galaxy. These results match well if you ignore the dim outer arms.

brtUnifScore: This cluster appears to emphasize the wispy outer structure of the original galaxy.

contourBrtRatio: No changes to the ordering.

alenAt25pct: Flocculent galaxies that ignore either the wispy outer arms or the well-defined bulge appear in the ordering.

rankAt75pct: Contrast between the inner and outer regions is ignored in some galaxies. A barred galaxy with an inner ring (rather than arms) also appears.

totalArcLength: The ordering favors a dark central bulge and two long arms rather than maintaining distinct inner and outer regions.

pa_avg__abs: A galaxy with a similar brightness gradient but no distinct arms enters the top 10, pushing others down.

pa_alenWtd_avg__abs: The ordering appears to overemphasize the wispy outer regions in the final entries.

bar_cand_score_orifld: Brightness concentration in the inner region is ignored. The results reflect selection based more on the inner region's shape.

4.4.4 Hubble Type T 3

Table 4.5: Visual comparisons of ordering results for Hubble type T=3 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 3 (All features)
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore

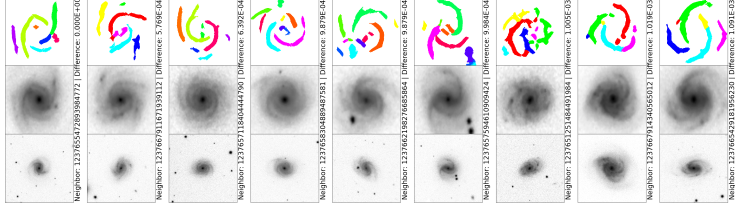
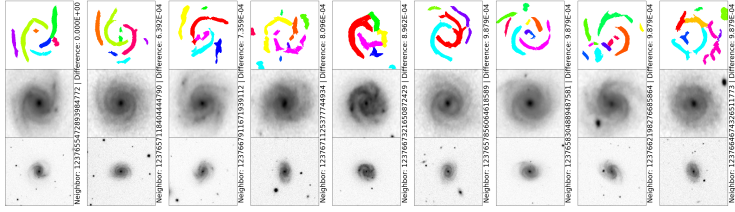
Continued on next page

Table 4.5 – continued from previous page

Image	T-Type / Feature
	contourBrtrRatio
	alenAt25pct
	rankAt75pct
	totalArcLength
	pa_avg__abs

Continued on next page

Table 4.5 – continued from previous page

Image	T-Type / Feature
	pa_alenWtd_avg__abs
	bar_cand_score_orifld

Commentary on Table 4.5

Control Group: The general shape of these galaxies appears to fit a three-arm structure with a wide central bulge. Two of the arms are roughly symmetric but oppositely curved, while the third arm branches off one of the main arms.

diskAxisRatio: The three-arm fit degrades to matching galaxies with just two arms, losing the characteristic branching structure.

diskMajAxsLen: This ordering overemphasizes bulge shape, at the expense of correctly matching arm geometry.

brtUnifScore: The algorithm appears to mistake the third branching arm for general flocculent structure, reducing match quality.

contourBrtrRatio: No changes to the ordering.

alenAt25pct: The inner bulge's eye shape is no longer preserved.

rankAt75pct: The ordering is slightly shuffled but generally consistent.

totalArcLength: The final galaxy in the list lacks a third, branching arm, weakening the structural consistency.

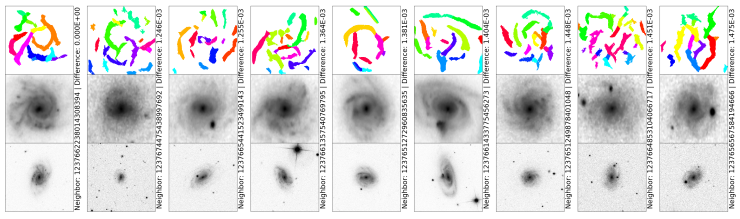
pa_avg__abs: No significant change to the order.

pa_alenWtd_avg__abs: One new galaxy appears in position 5; it lacks a bulge that matches the target, weakening the match.

bar_cand_score_orifld: The brightness gradient becomes a stronger matching criterion than arm shape. In particular, neighbor 9 has dissimilar arms but appears in the list due to similar brightness trends.

4.4.5 Hubble Type T 4

Table 4.6: Visual comparisons of ordering results for Hubble type T=4 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 4 (All features)

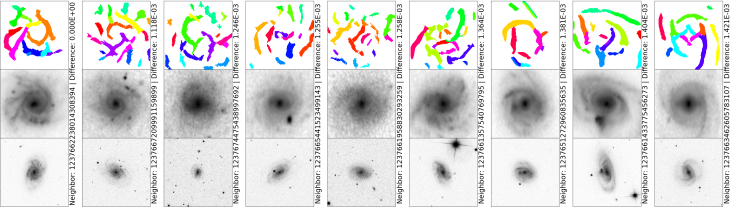
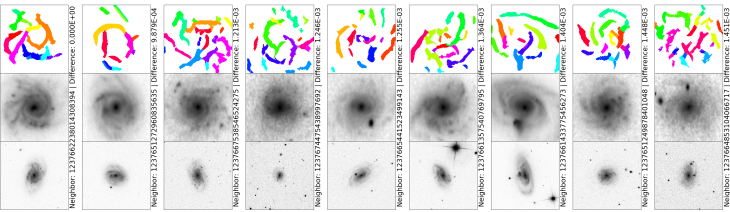
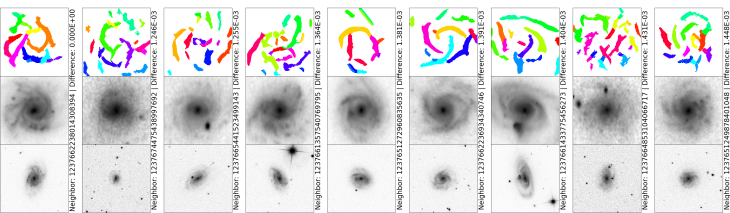
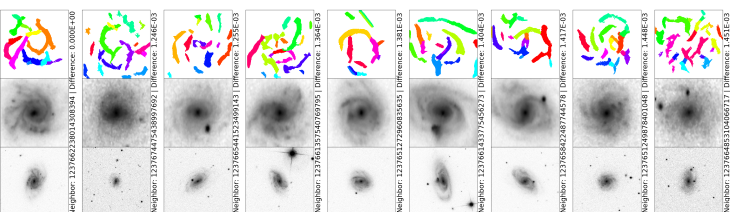
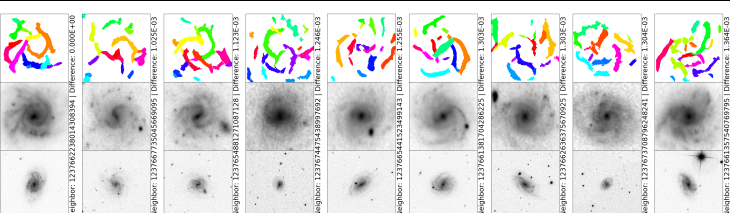
Continued on next page

Table 4.6 – continued from previous page

Image	T-Type / Feature
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore
	contourBrtrRatio
	alenAt25pct

Continued on next page

Table 4.6 – continued from previous page

Image	T-Type / Feature
	rankAt75pct
	totalArcLength
	pa_avg_abs
	pa_alenWtd_avg_abs
	bar_cand_score_orifld

Commentary on Figure 4.6

Control Group: The general shape of this cluster of galaxies is a medium-brightness central bulge with a single distinct arm spiraling out from the center. Surrounding this bulge is flocculent structure. A secondary geometric cue — possibly spurious — involves bright foreground stars or unresolved clusters in the target image. The ordering consistently selects galaxies with similar bulges and foreground stars, then differentiates based on either arm structure or brightness gradient.

diskAxisRatio: The ordering shifts toward fuzzy galaxies with two arms, moving away from the single-arm structure.

diskMajAxsLen: The feature appears critical for identifying stars or short bright arms — its removal leads to less emphasis on these bright point-like features.

brtUnifScore: A few barred galaxies with three arms are introduced into the ordering, indicating confusion with arm symmetry.

contourBrtratio: No significant changes observed.

alenAt25pct: Greater emphasis is placed on matching arm structure rather than the full visual balance of bulge and surroundings.

rankAt75pct: Two-armed galaxies with a similar bulge structure are pulled into the order, diverging from the flocculent + single-arm target.

totalArcLength: The order is broadly reshuffled without obvious thematic consistency.

pa_avg__abs: A 3-armed galaxy with fewer prominent stars enters at position 5, and the final 4 galaxies are reordered.

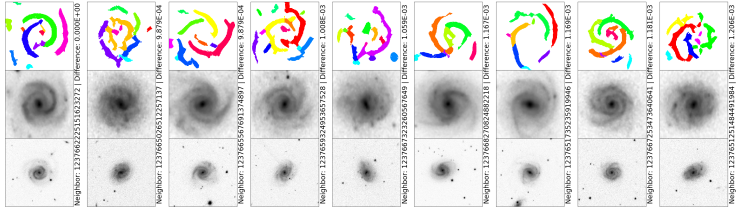
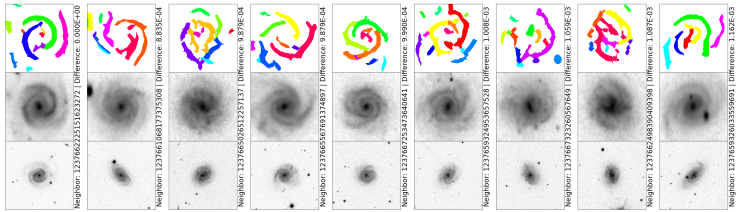
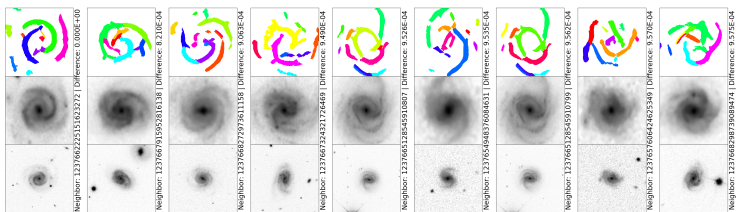
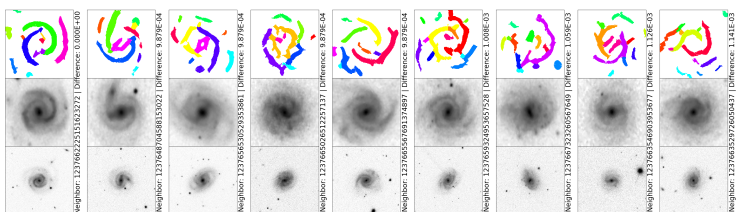
pa_alenWtd_avg__abs: A two-armed galaxy with a prominent star enters at position 7; the

rest of the order remains mostly intact.

bar_cand_score_orifld: Less weight is given to the central bulge structure. The resulting galaxies tend toward thin bulges with tight arms or wider, moderate-brightness bulges.

4.4.6 Hubble Type T 5

Table 4.7: Visual comparisons of ordering results for Hubble type T=5 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 5 (All features)
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore

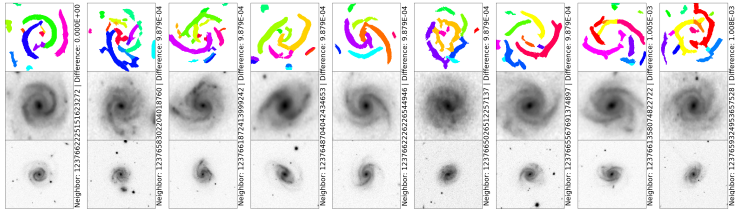
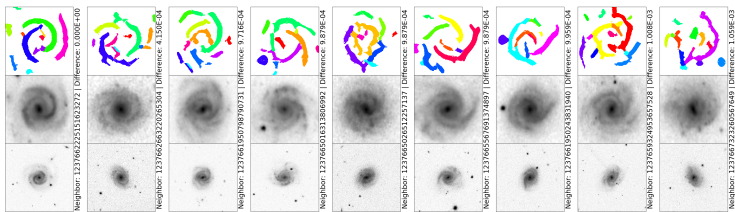
Continued on next page

Table 4.7 – continued from previous page

Image	T-Type / Feature
	contourBrtrRatio
	alenAt25pct
	rankAt75pct
	totalArcLength
	pa_avg__abs

Continued on next page

Table 4.7 – continued from previous page

Image	T-Type / Feature
	pa_alenWtd_avg__abs
	bar_cand_score_orifld

Commentary on Figure 4.7

Control Group: The target galaxy is a tightly wound, two-armed spiral with a bright bulge that is clearly separated from the spiral arms. Notably, Galaxy 4 lacks obvious visible arms, but SpArcFiRe identifies two elongated arm segments in its mask. This suggests the feature extraction is detecting subtle structure not easily visible in raw or deprojected images.

diskAxisRatio: Galaxies are reordered with greater emphasis on bulge shape over spiral structure.

diskMajAxisLen: Galaxies with flocculent arms are pulled into the ranking, reducing similarity to the structured two-arm target.

brtUnifScore: Matches now include barred galaxies and others where the dominant arms connect directly to the bulge, unlike the control where they are clearly separated.

`contourBrtratio`: The ordering is nearly unchanged, with the exception of one new galaxy in position 5 lacking distinct arms.

`alenAt25pct`: Bar-like galaxies with tight, nearly circular arms around the bulge are included; the arm fits remain visually good.

`rankAt75pct`: A flocculent galaxy with a similar bulge enters at position 8. The rest of the order is stable.

`totalArcLength`: The new ordering appears to focus on bulge size and arm brightness distribution. Flocculent and near-ring structures enter the list.

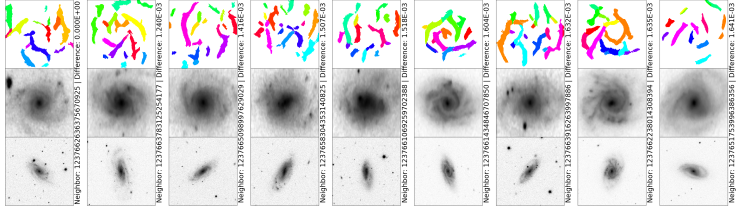
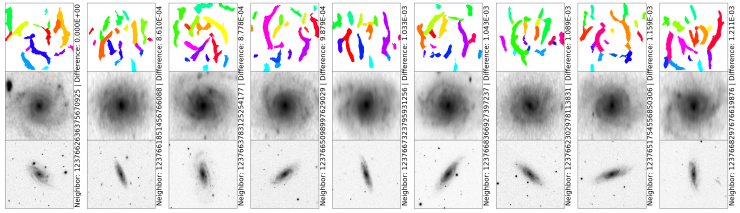
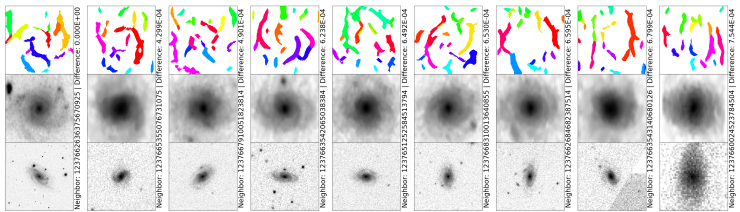
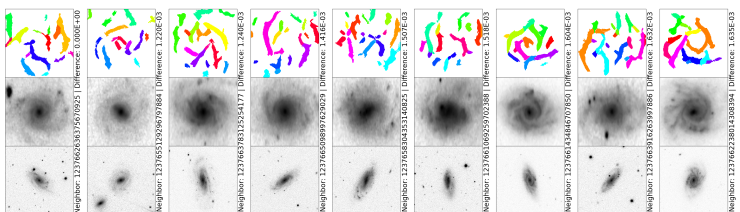
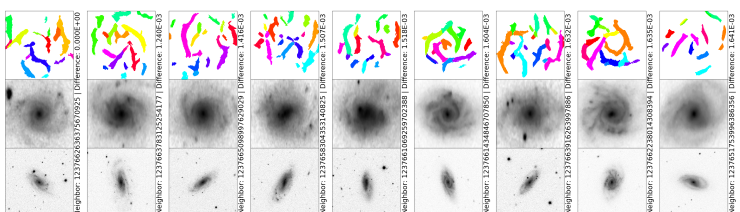
`pa_avg__abs`: The ordering is shuffled. Neighbor 7 is a new galaxy where the fit appears to emphasize arm shape more than brightness or bulge structure.

`pa_alenWtd_avg__abs`: Matching emphasizes arm length and brightness more than pitch angle, allowing in galaxies with similar arm profiles but less geometric alignment.

`bar_cand_score_orifld`: Bulge shape appears to be ignored. Several new galaxies lack the clean separation between arms and bulge seen in the control.

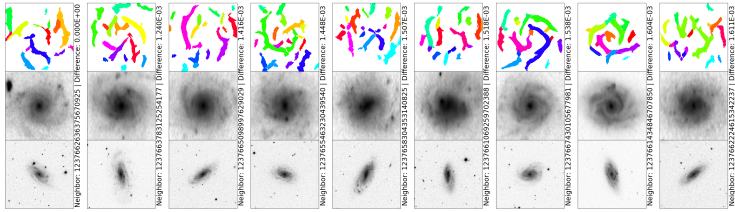
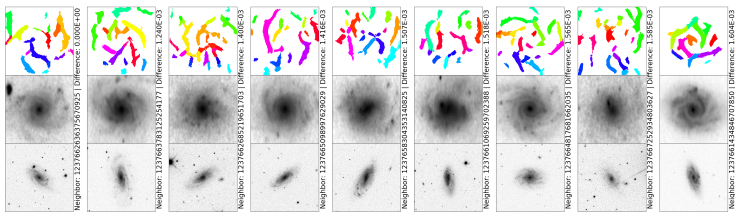
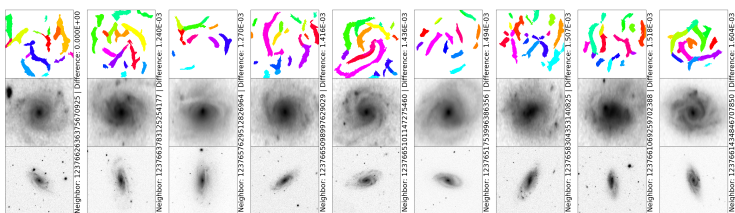
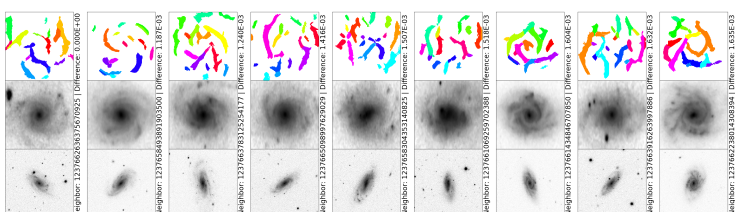
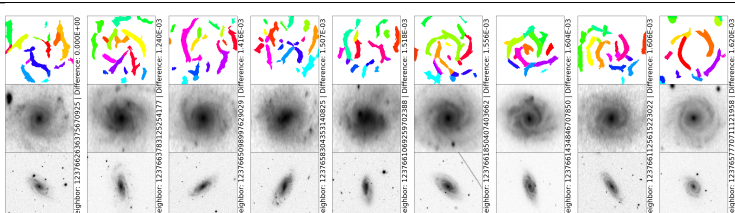
4.4.7 Hubble Type T 6

Table 4.8: Visual comparisons of ordering results for Hubble type T=6 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 6 (All features)
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore
	contourBrtrRatio

Continued on next page

Table 4.8 – continued from previous page

Image	T-Type / Feature
	aleneAt25pct
	rankAt75pct
	totalArcLength
	pa_avg_abs
	pa_alenWtd_avg_abs

Continued on next page

Table 4.8 – continued from previous page

Image	T-Type / Feature
	bar_cand_score_orifld

Commentary on Figure 4.8

Control Group: SpArcFiRe interpreted the target galaxy as a series of broken ring segments rather than two smooth, wispy arms spiraling around a wide, moderate-brightness bulge. Despite this imperfect fit, the selected neighbors mostly consist of galaxies with two faint, extended arms wrapping around similarly bright bulges. This demonstrates that the SpArcFiRe feature space can still yield useful similarities even when its direct output is suboptimal.

diskAxisRatio: The bulge shapes dominate the ordering, while arm structure is less emphasized. The list is slightly shuffled compared to the control.

diskMajAxsLen: Matching shifts toward galaxies with broader, fainter arms, suggesting this feature helps avoid overly diffuse matches.

brtUnifScore: A new galaxy with a dense bulge and dim arms enters first, shifting the rest downward — suggesting brightness uniformity helps constrain this balance.

contourBrtratio: No change in the order, indicating this feature did not significantly

impact neighbor selection.

alenAt25pct: The number of arms seems less constrained; galaxies with more segmented or fragmented arm structures appear in the ranking.

rankAt75pct: The list is significantly shuffled, and new galaxies with flocculent arm patterns are introduced.

totalArcLength: Brightness distribution — especially in the central bulge — becomes the dominant match criterion. Several galaxies with few arms enter the list.

pa_avg__abs: A new first neighbor has arm fragments rather than smooth spirals, highlighting this parameter's influence on perceived continuity.

pa_alenWtd_avg__abs: Two galaxies with more numerous arms but similar brightness and distribution profiles are included.

bar_cand_score_orifld: Bulge size and shape are ignored; instead, matches focus solely on arm geometry and light distribution.

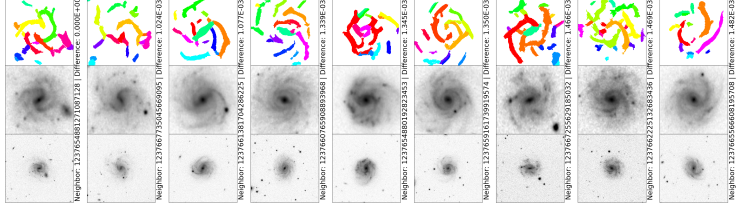
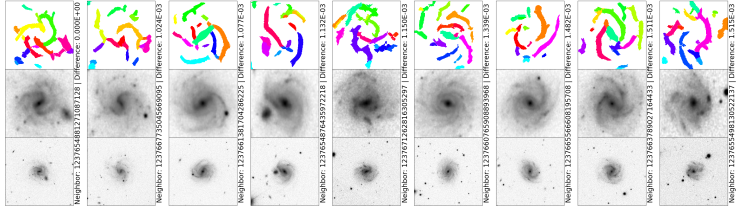
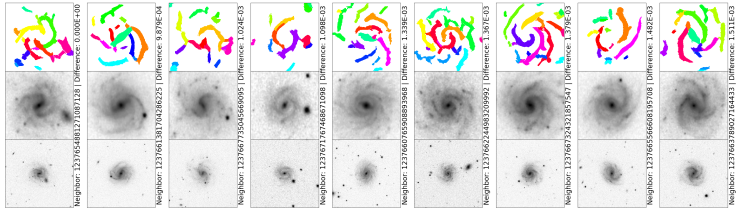
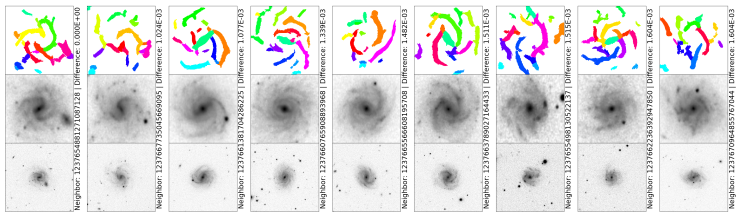
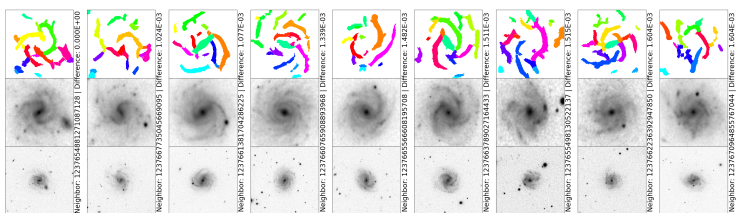
4.4.8 Hubble Type T 7

Table 4.9: Visual comparisons of ordering results for Hubble type T=7 with individual SpArcFiRe features removed.

Image	T-Type / Feature
	T = 7 (All features)
	diskAxisRatio
	diskMajAxsLen
	brtUnifScore
	contourBrtrRatio

Continued on next page

Table 4.9 – continued from previous page

Image	T-Type / Feature
	aleneAt25pct
	rankAt75pct
	totalArcLength
	pa_avg_abs
	pa_alenWtd_avg_abs

Continued on next page

Table 4.9 – continued from previous page

Image	T-Type / Feature
	bar_cand_score_orifld

Commentary on Figure 4.9

Control Group: The target galaxy is a tightly wound grand design spiral with two distinct arms that span from the central bar-like bulge to the galaxy’s outer edge. The neighbors appear to share this defining structure: bar galaxies with two dominant, extended arms. One image contains a second galaxy, a case that highlights how different feature sets influence which images pass similarity thresholds—even allowing secondaries in crowded fields.

diskAxisRatio: A neighbor containing two galaxies is introduced. Ignoring the secondary, the fit is still visually plausible. Arm separation is less consistent, with some late entries showing bar-arm continuity.

diskMajAxsLen: Matching shifts to include galaxies with noisier outer structures and fewer cleanly defined arms.

brtUnifScore: Inner structure is prioritized over arm continuity. Several galaxies lack the distinct inner-ring geometry seen in the control group.

contourBrtRatio: No significant change to the ordering.

alenAt25pct: The bar shape of the bulge is de-emphasized, and later neighbors feature rounder central bulges than the target.

rankAt75pct: Another two-galaxy image enters. The smaller secondary is likely misidentified by SpArcFiRe as a spiral arm segment.

totalArcLength: Matches favor arm shape, restoring some focus on two-arm geometry—though less emphasis on bar-bulge separation.

pa_avg__abs: No change from the full-featured ordering.

pa_alenWtd_avg__abs: No change from the full-featured ordering.

bar_cand_score_orifld: The bar structure is not prioritized. The resulting galaxies tend to show a strong bulge with irregular or flocculent arm segments.

Chapter 5

Novel relationships to spiral structure

In this chapter we demonstrate novel relationships between the spiral arm parameters measured by SpArcFiRe, and other observable quantities that are not directly related to arm structure. To demonstrate the novel relationships of our similarity algorithm, we analyze its relationship to three galaxy parameters largely independent of spiral structure: age, star formation rate, and metallicity and show that they are statistically correlated with visual appearance [17] [10].

5.1 Global Relationship between Pairwise Distance and Non-Spiral Parameters

Our first naive attempt to show this relationship was simply to compare the difference measure of target galaxies and the difference in star formation rate, age, and metallicity. We initially saw almost no signal, likely due to the significant noise inherent in precisely measuring visual appearance of highly complex objects like spiral galaxies. In order to try and maximize this approach I calculated distance between each of the 29000 “nice” GZ1

galaxies as used by Davis [4] and each of their and their nearest 3000 neighbors. I then sorted all target neighborhood pairs to find global best pairs. After this I compared these collections of distances to the respective star formation rate, age, and metallicity differences of the pairs to produce table. While this slightly improved observed correlations, they were still of marginal statistical significance. We had inconsistent signal with small groups of data, but as we increased our datasets to the upper limit, we converged on lackluster results.

The distribution of galaxy shape is not uniform. It is well known that large scale cosmic phenomenon can cause clustered galaxy morphology, as stated by Davis *et al* (1976).[5]. When we originally tried to find a correlation between the distance metric used by this paper and the physical parameters of galaxies we simply took the pairs of galaxies with the smallest difference and looked at how that difference related to star formation rate, age, and metallicity. These results were confounding. In an attempt to try and make sense of how our results were distributed across our high dimensional feature space we turned to a t-SNE plot, which essentially project high dimensional feature spaces down to 2 dimensional plots while maintaining, as much as possible, the pairwise distance between them. When we visualized using a t-SNE plot we saw very complex structure as depicted in Figure 5.1.

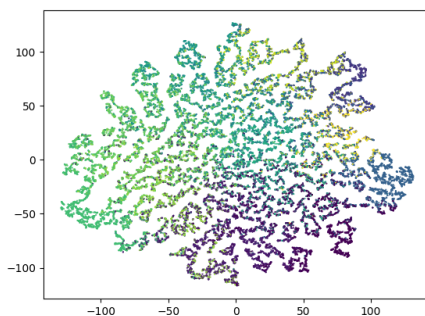


Figure 5.1: While there is clearly some structure in the distribution of galaxies in the sprial-structre feature space, it is not clear how this structure is related to other galaxy properties.

5.2 Single Galaxy Variance from Global Relationships

A breakthrough came about when I inspected individual galaxies. Some galaxy orderings had almost no relationship between their difference, rank, and physical attributes while others had a clear relationship. An example of how some galaxies dramatically out-performed the globally selected pairs of differences can be seen in table 5.1. This led us to the insight that cluster to cluster relationships were potentially more significant than galaxy to galaxy relationships and that insight produce powerful results.

Name	DiffPearson	DiffPvalue	RankPerason	RankPvalue
SFR	-0.034277354	8.09E-09	-0.015199267	0.01057462
Age	0.158657058	7.55E-159	0.123053407	7.31E-96
Metallicity	-0.014030364	0.018283522	-0.021750967	0.000253558

Table 5.1: The numbers in this table are just from the similarity order of galaxy 1237648703503794332 and its neighbors. The Pearson correlation between the feature difference of two galaxies and their physical parameter differences. Both shown by feature difference and similarity rank.

5.3 Neighborhood to Neighborhood Relationships

Once we were aware of the fact that some similarity orderings were more related to physical parameters than others, we redid our t-SNE plot one cluster at a time as shown following figures 5.2 and 5.3. Unlike pairwise comparisons that failed to reveal clear structure, these visualizations demonstrate that similarity-based groupings emerge naturally without enforcing clustering by physical parameters. The similarity orders naturally divide themselves if we pick target galaxies at random and just plot their 100 nearest neighbors.

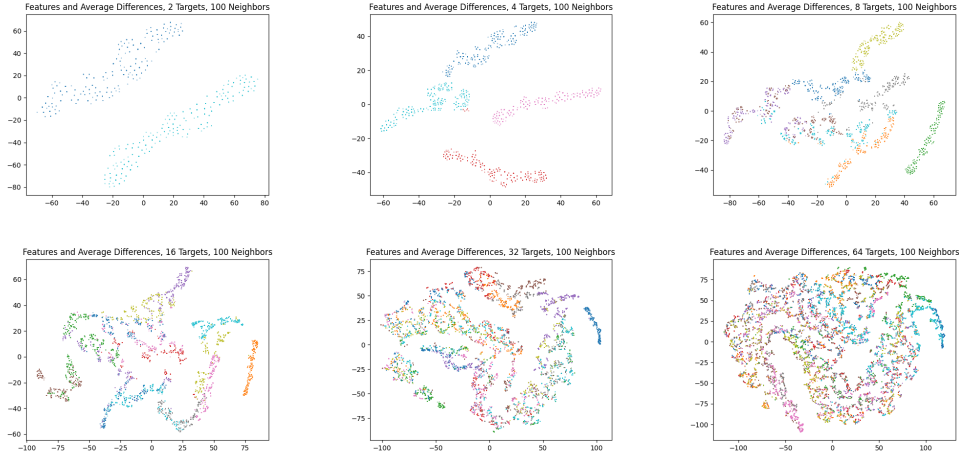


Figure 5.2: Showcase of the progression of Cluster based t-SNE plots that begin as distinct neighborhoods then converges to a similar outcome as figure 5.1

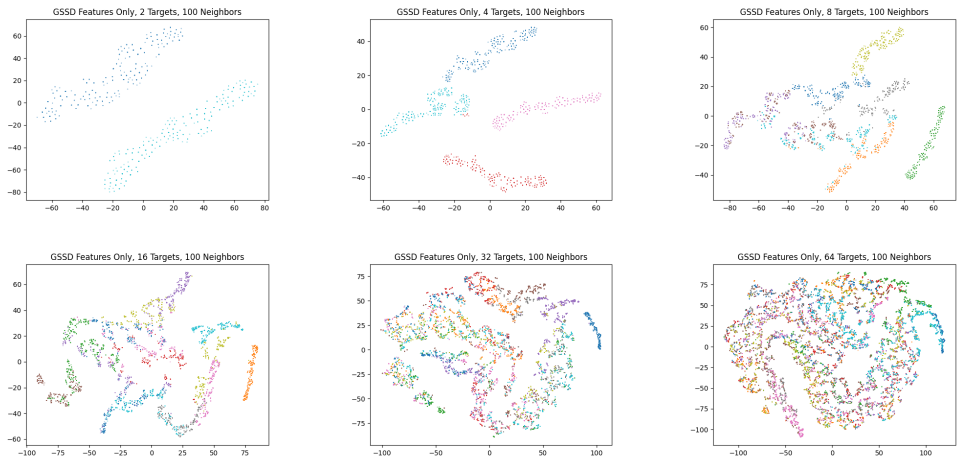


Figure 5.3: Showcase of progression of neighborhood based t-SNE without using the mean physical parameters in the vector space. It begins mixing earlier, however distinct structure is still clear until around 64 neighborhoods

Armed with the knowledge that at last in 10 dimensional vector space, we had clear separation of clusters it was time to again check if there was a relationship between our difference metric and physical parameters. To represent this relationship we took the target of two different clusters of galaxies and related that to the difference in mean physical parameter for the two clusters. So we compared the $\text{maxdiff}(\text{targetA}, \text{targetB})$ with $\text{abs}(\text{mean}(\text{clusterA}) - \text{mean}(\text{clusterB}))$. This led to almost a 10x stronger correlations for each target and cluster average value shown in 5.2.

Lastly in order to give context to the previous table, I will be creating an equivalent table to represent morphological classes. In the set of galaxies that fit within the union of our GZ and EFIGI datasets only 5 Hubble types had more than 50 galaxies. The results of choosing one random target per Hubble type and the resulting feature vector difference correlated with age, sfr, and metallicity can be found in table 5.3. We note that Table 5.3 does not include GSSD, we are simply showing the correlation between Hubble Type and the non-spiral measures.

5.3.1 NGram Analysis

We now know that SpArcFiRe’s features build a very strong base for determining visual difference. In order to gain a more detailed understanding of how SpArcFire’s features relate to physical parameters we can examine subsets of features rather than all 10 at once. We already have some preliminary results in this shown in table 5.5. From table 5.5 we can immediately see that the total arc length of clusters is related to their average physical parameters. This has been extensively explored by [16]. The next group of similar weighted relationships is pitch angle weighted average with metallicity.

GSSD differences are calculated in a similar manner to a decision tree where the value with

the largest difference is how the difference is calculated. As such, some values dominate the results of our ngrams. As seen in table 5.5 when totalArcLength is chosen as one of the two features in the 2gram the results are mostly similar. As such by the time we include all features in 9grams the relationships are dominated by totalArcLength.

In order to provide a clearer picture of how general spiral structure relates to SFR, AGE, and Metallicity we will provide four sets of plots. One set will be of all features included as 9grams, one set will be curated to remove totalArcLength as well some SpArcFiRe features such as diskAxisRatio and diskMajAxesLen as they don't provide much spiral structure information, the third set will be a couple of 2grams to show that the co-relations found in the 2gram table match the relationships we found between groups of galaxies, and finally individual features or 1grams in order to show how the relationships change at the 2, 6, and 9gram levels.

Due to each feature having a dramatically different range and distribution, we normalized the tables in order to provide a constant reference between feature sets. For each type of parameter between GSSD, age, metallicity, and sfr, we used the formula parameter set p, and value in group n: $\sqrt[6]{(n - \mu_p)/\sigma_p + 1.5}$

Group Physical Paramater Relationships								
			SFR		Age		Metallicity	
Targets	Neighbors	Pairs	Corr	Pvalue	Corr	Pvalue	Corr	PValue
2	10	3	1	0	1	0	1	0
4	10	10	0.671	3.35E-02	0.907	2.88E-04	0.803	5.20E-03
6	10	21	0.586	5.21E-03	0.388	8.19E-02	0.635	1.97E-03
8	10	36	0.554	4.57E-04	0.252	1.38E-01	0.605	9.19E-05
16	10	136	0.412	6.17E-07	0.119	1.68E-01	0.383	4.18E-06
32	10	528	0.242	1.80E-08	0.090	3.92E-02	0.284	3.10E-11
64	10	2016	0.168	2.92E-14	0.062	5.40E-03	0.240	6.74E-28
2	20	3	1	0	1	0	1	0
4	20	10	0.581	7.81E-02	0.860	1.43E-03	0.875	9.10E-04
6	20	21	0.531	1.33E-02	0.508	1.88E-02	0.784	2.58E-05
8	20	36	0.513	1.37E-03	0.352	3.51E-02	0.720	7.46E-07
16	20	136	0.442	6.91E-08	0.202	1.84E-02	0.515	1.44E-10
32	20	528	0.332	5.05E-15	0.164	1.52E-04	0.327	1.15E-14
64	20	2016	0.204	2.02E-20	0.134	1.49E-09	0.276	1.90E-36
2	30	3	1	0	1	0	1	0
4	30	10	0.577	8.09E-02	0.864	1.28E-03	0.803	5.15E-03
6	30	21	0.561	8.17E-03	0.443	4.41E-02	0.767	4.99E-05
8	30	36	0.533	8.10E-04	0.340	4.24E-02	0.650	1.77E-05
16	30	136	0.437	1.04E-07	0.214	1.25E-02	0.488	1.73E-09
32	30	528	0.375	4.13E-19	0.168	1.08E-04	0.311	2.82E-13
64	30	2016	0.220	1.86E-23	0.149	1.53E-11	0.264	2.25E-33
2	40	3	1	0	1	0	1	0
4	40	10	0.582	7.76E-02	0.882	7.22E-04	0.771	8.98E-03
6	40	21	0.571	6.82E-03	0.442	4.50E-02	0.744	1.11E-04
8	40	36	0.565	3.27E-04	0.366	2.81E-02	0.645	2.20E-05
16	40	136	0.450	3.98E-08	0.261	2.11E-03	0.510	2.35E-10
32	40	528	0.387	2.74E-20	0.194	7.27E-06	0.330	6.47E-15
64	40	2016	0.223	4.13E-24	0.172	6.32E-15	0.284	9.59E-39
2	50	3	1	0	1	0	1	0
4	50	10	0.578	8.02E-02	0.935	7.04E-05	0.775	8.45E-03
6	50	21	0.566	7.52E-03	0.436	4.82E-02	0.746	1.05E-04
8	50	36	0.578	2.23E-04	0.392	1.81E-02	0.656	1.39E-05
16	50	136	0.444	5.94E-08	0.302	3.51E-04	0.533	2.30E-11
32	50	528	0.392	7.45E-21	0.209	1.28E-06	0.347	2.25E-16
64	50	2016	0.223	4.06E-24	0.180	3.32E-16	0.291	1.14E-40

Table 5.2: These numbers are generated by taking the pearson correlation between the *mean* GSSD among neighbors of the two targets, and the difference in non-spiral parameters—in other words, distance in GSSD correlates with distance in non-spiral parameters.. Naming convention examples: SFRCo is the Pearson corelation for difference in sfr and SFRPV is the pvalue for the pearson correlation in the previous column

Hubble Sequence Physical Paramter Relationships								
			SFR		Age		Metallicity	
targets	neighbors	Pairs	Corr	PValue	Corr	PValue	Corr	PValue
6	10	21	0.128	0.581	0.678	7.26E-04	0.736	1.42E-04
6	20	21	0.126	0.587	0.766	5.18E-05	0.672	8.49E-04
6	30	21	0.160	0.489	0.739	1.29E-04	0.698	4.31E-04
6	40	21	0.182	0.429	0.682	6.68E-04	0.690	5.35E-04
6	50	21	0.226	0.326	0.678	7.34E-04	0.690	5.41E-04

Table 5.3: These numbers are the GSSD feature set distance between a randomly chosen galaxy of each Hubble type from 1 to 6 and the average SFR, Age, and Metallicity of 50 randomly chosen galaxies of the respective Hubble type.

Features	SFRPearson	SFRpvalue	AGEpearson	AGEpvalue	METALPearson	METALpvalue
alenAt25pct_totalArcLength	2.22e-01	5.44e-24	1.80e-01	3.27e-16	3.25e-01	6.32e-51
pa_alenWtd_avg_abs_totalArcLength	2.24e-01	2.84e-24	1.45e-01	6.05e-11	3.21e-01	1.40e-49
pa_avg_abs_totalArcLength	2.24e-01	2.87e-24	1.45e-01	6.06e-11	3.21e-01	1.46e-49
bar_cand_score_orifld_totalArcLength	2.24e-01	2.93e-24	1.45e-01	6.06e-11	3.21e-01	1.57e-49
diskAxisRatio_totalArcLength	2.24e-01	2.96e-24	1.45e-01	6.06e-11	3.21e-01	1.59e-49
brtUnifScore_totalArcLength	2.24e-01	2.96e-24	1.45e-01	6.06e-11	3.21e-01	1.59e-49
contourBrtratio_totalArcLength	2.24e-01	2.96e-24	1.45e-01	6.06e-11	3.21e-01	1.59e-49
bar_cand_score_orifld_pa_alenWtd_avg_abs	2.69e-02	2.27e-01	1.24e-01	2.62e-08	1.86e-01	4.04e-17
pa_alenWtd_avg_abs_pa_avg_abs	7.85e-02	4.22e-04	1.25e-01	1.74e-08	1.79e-01	5.04e-16
contourBrtratio_pa_alenWtd_avg_abs	2.88e-02	1.97e-01	9.59e-02	1.62e-05	1.77e-01	1.37e-15
brtUnifScore_pa_alenWtd_avg_abs	2.88e-02	1.97e-01	9.59e-02	1.62e-05	1.77e-01	1.37e-15
bar_cand_score_orifld_pa_avg_abs	7.94e-02	3.58e-04	1.49e-01	1.68e-11	1.74e-01	3.51e-15
brtUnifScore_pa_avg_abs	7.83e-02	4.35e-04	1.20e-01	6.73e-08	1.66e-01	6.26e-14
bar_cand_score_orifld_contourBrtratio	-2.48e-02	2.65e-01	3.24e-01	2.38e-50	1.04e-01	2.81e-06
bar_cand_score_orifld_brtUnifScore	-2.48e-02	2.65e-01	3.24e-01	2.47e-50	1.04e-01	2.82e-06
alenAt25pct_pa_alenWtd_avg_abs	-6.65e-03	7.65e-01	4.04e-01	4.85e-80	3.12e-02	1.61e-01
alenAt25pct_brtUnifScore	-6.52e-03	7.70e-01	4.02e-01	2.24e-79	2.93e-02	1.88e-01
alenAt25pct_bar_cand_score_orifld	-6.53e-03	7.70e-01	4.02e-01	2.28e-79	2.93e-02	1.89e-01
brtUnifScore_contourBrtratio	-4.23e-02	5.73e-02	9.07e-02	4.53e-05	7.08e-03	7.51e-01

Table 5.4

Table 5.5: Table of 2grams, their sample size, Pearson correlations to average physical values and the respective p-values. The averages were taken from neighborhoods of size 100. The suspicious similarity between 2grams that shared totalArcLength led us to examine our relationships with various Ngrams. The three physical parameters tested are Star Formation Rate, Age, and Metallicity

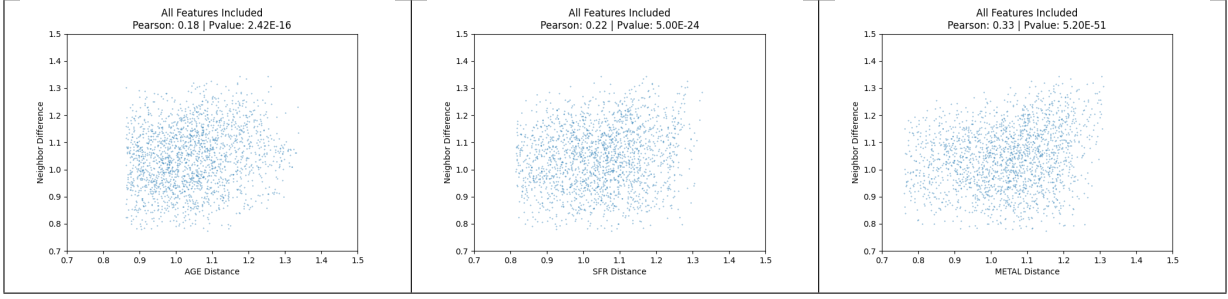


Table 5.6: Table of 9grams. We have consistent results among Age, SFR, and metallicity labeled metal on the table. A problem we noticed as that this very closely matches the totalArcLength row of the 1gram table 5.7. The x-axis is the GSSD neighbor difference between two galaxies as $GSSD(A, B)$ and the Y axis is the physical parameters of group a and group b as GA and GB in the formula: $\mu_{GA} - \mu_{GB}$.

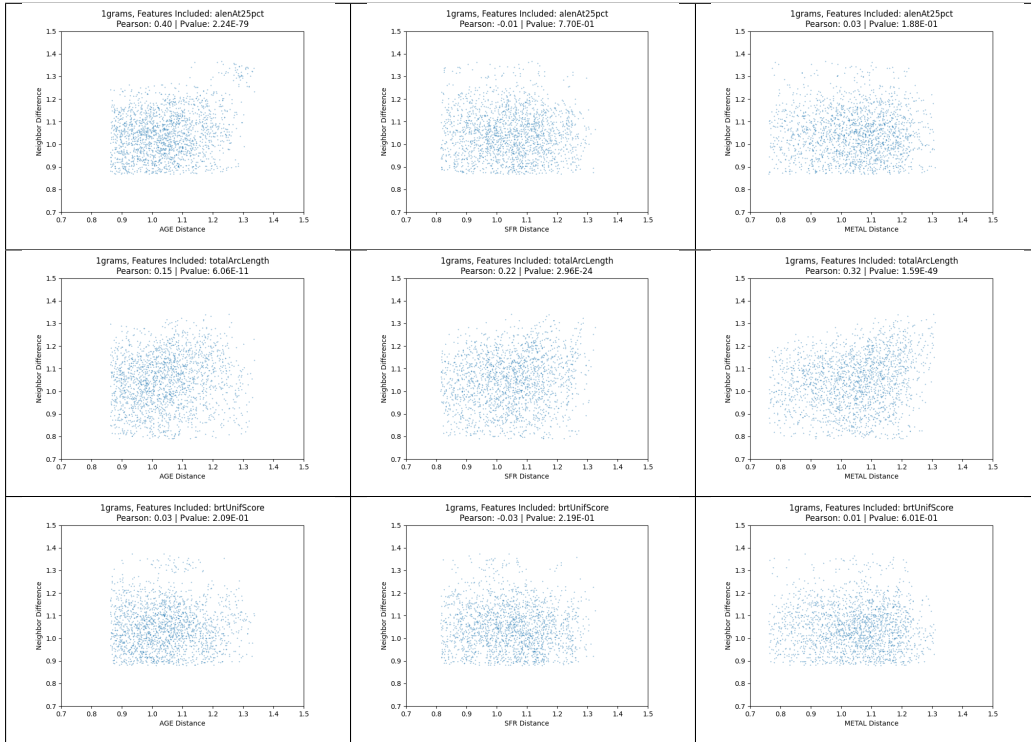


Table 5.7: The x-axis is the GSSD neighbor difference between two galaxies as $GSSD(A, B)$ and the Y axis is the physical parameters of group a and group b as GA and GB in the formula: $\mu_{GA} - \mu_{GB}$.

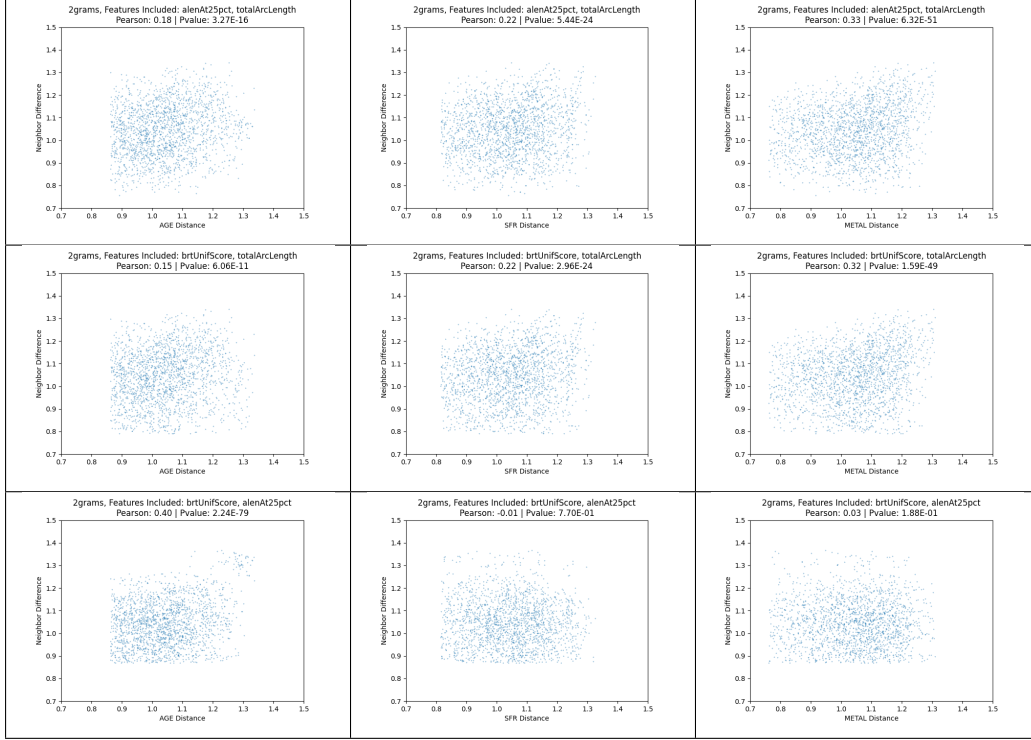


Table 5.8: Table of 2grams. When we remove totalArcLength, the results strongly match alenAt25pct. This indicates that whatever feature is most dominant is what the physical parameters are relating to. The x-axis is the GSSD neighbor difference between two galaxies as $GSSD(A, B)$ and the Y axis is the physical parameters of group a and group b as GA and GB in the formula: $\mu_{GA} - \mu_{GB}$.

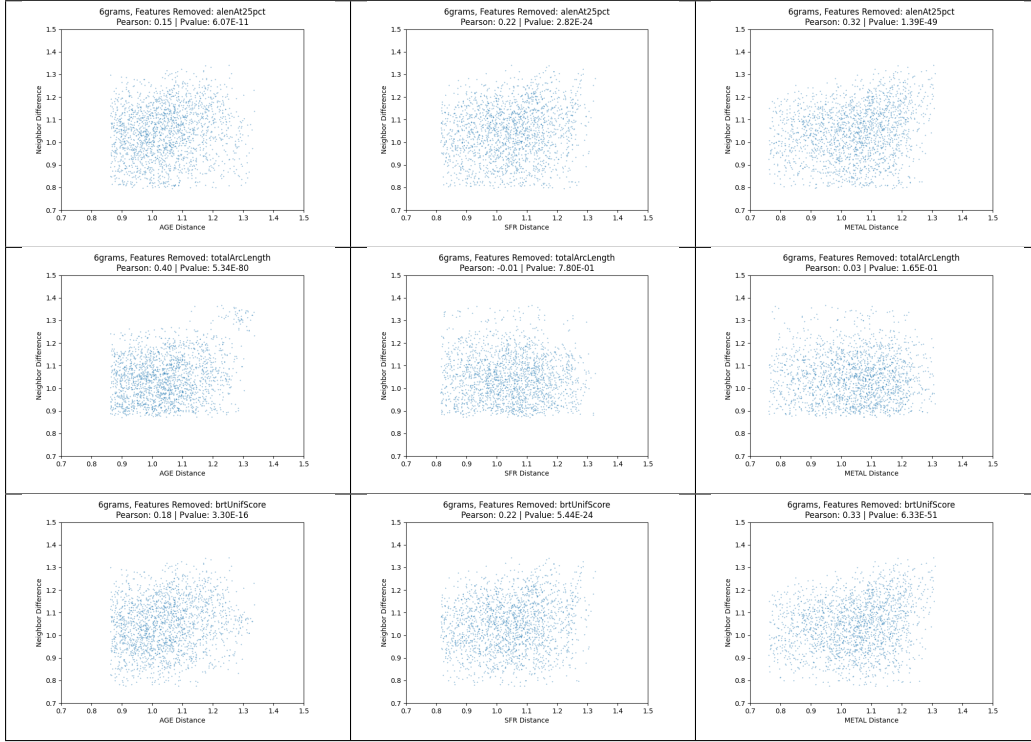


Table 5.9: Table of 6grams. The 6grams mostly strongly totalArcLength then, match alenAt25pc_age afterwards. The x-axis is the GSSD neighbor difference between two galaxies as $GSSD(A, B)$ and the Y axis is the physical parameters of group a and group b as GA and GB in the formula: $\mu_{GA} - \mu_{GB}$.

With information from tables 5.7, 5.5, 5.9, and 5.6, we gain a clear picture of mechanically how GSSD works with regards to different types of features. Whatever feature is most dominant in deciding the GSSD difference is what the data move towards. This is by no means a bad thing. The top two features are very reasonable choices for the highest value information in choosing how spiral galaxies are related. totalArcLength in our testing roughly translated to how much of the galaxy had spiral structure. alenAt25pct was a way for us to match flocculence.

5.4 Analysis of Selected SpArcFire Features

Understanding that 1grmas have a strong physical relationship in groups, we decided to check the two most dominant features, totalArcLength and alenAt25pc, for group analysis and see if they had individual relationships to age, sfr, or metallicity. Professor Hayes confirmed that the relationships between mean SpArcFiRe feature values and physical parameters were both strong and surprising. We tested this first with our most dominant feature and it's strongest 1gram relationship in figure 5.4.

We tested this with a few more feature-parameter relationships to verify that this relationship was not an aberration:

Given the patterns shown in Figure 5.10, we have demonstrated that the GSSD similarity orderings not only reflect morphological structure, but also correspond to meaningful physical differences among galaxies.

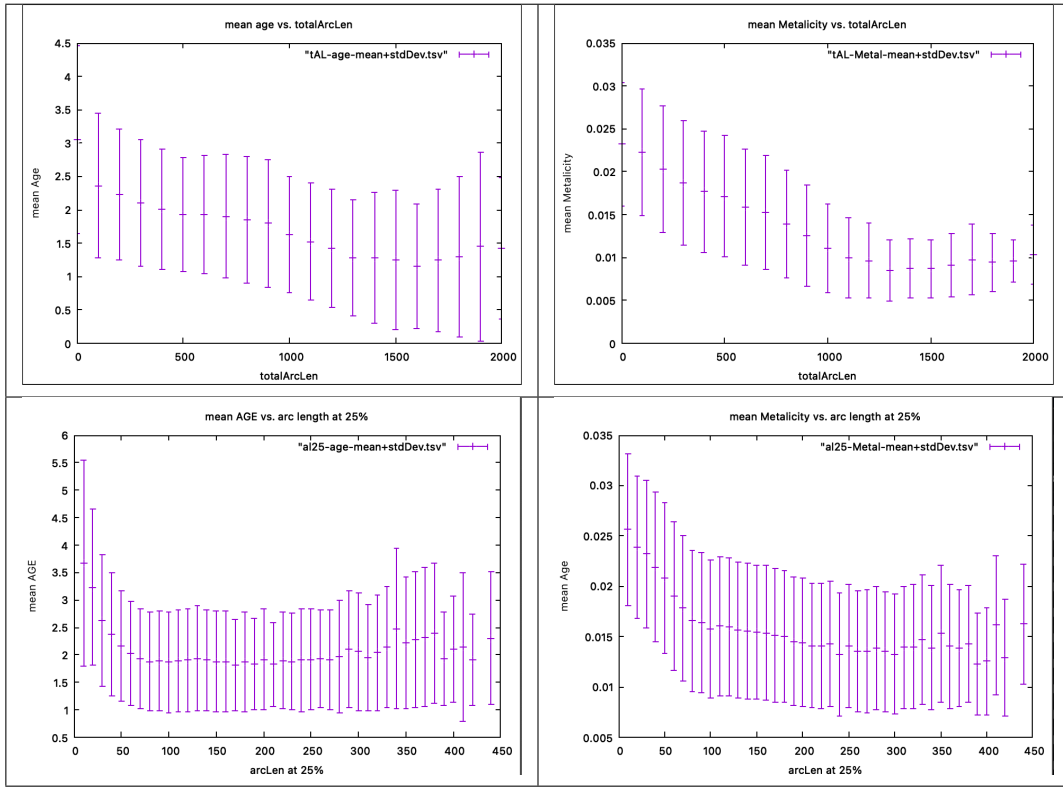


Table 5.10: This table shows the the relationship between Age and Metallicity for totalArcLength and alenAt25pct.

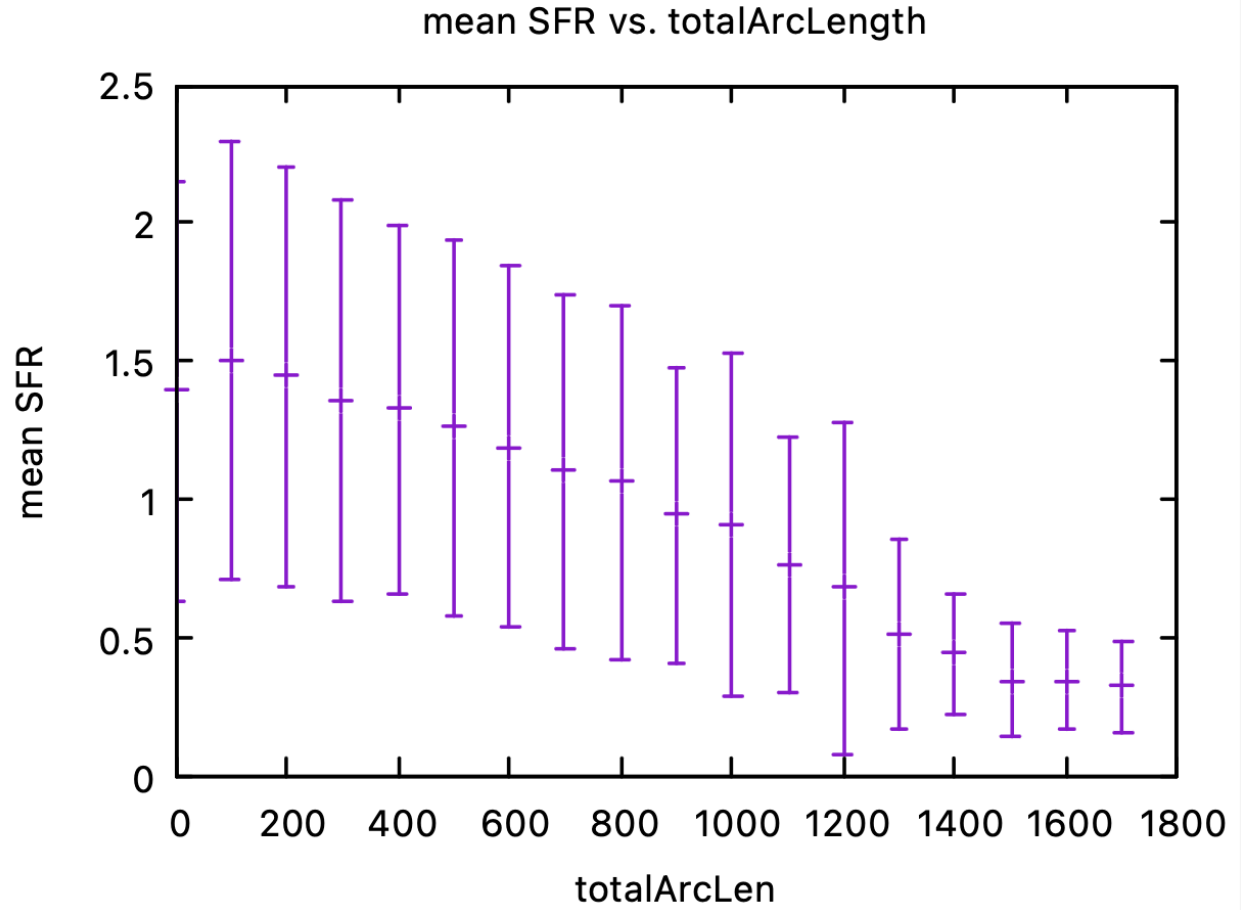


Figure 5.4: This is a plot of the mean SFR vs SpArcFiRe’s totalArcLength parameter across the 41,235 GZ1 galaxies satisfying the following properties: GZ1 spirality > 0.5 , nonzero totalArcLength, and nonzero SFR. The Pearson correlation of the raw data (before binning + taking means) is -0.2356 with a Pearson p-value of 10^{-521} (49σ). Taking the mean in bins of 100 totalArcLength gives the above plot (with error bars).

Chapter 6

Future Work

In this section I will be discussing some the opportunities that are opened up by having developed the Galaxy Similarity from Sorted Difference algorithm.

6.1 Adapting the Algorithm to Other Surveys

Since GSSD relies solely on SpArcFiRe output, it is in principle applicable to any survey where SpArcFiRe performs reliably, as discussed in Chapter 4.

6.1.1 Well Established Surveys

For example, in the EFIGI dataset they pull galaxies from multiple surveys in order to build up their list of about 4800 hand morphologically classified galaxies. Some ideas of how to prove cross survey feasibility: show that we don't favor grouping images within the same survey over other surveys assuming their SpArcFiRe feature distribution is the same among all surveys, show we maintain a similar level of physics relevance using only galaxies not used

in this body of work, and finally show the similar feature sensitivity for non SDSS images as SDSS images.

6.1.2 New Surveys

There are many new astronomical survey missions underway. One of the ones most relevant to pop culture in the United States is the James Webb Space Telescope or JWST. As of yet I have been unable to find any granular collections of spiral galaxies listed from the wide band image releases. The similarity ordering, since it was shown to work on most any image with an accurate SpArcFiRe fit, could be used to rapidly build collections of spiral galaxies from the swath of unlabeled objects within that survey. There are many more surveys that could also be enhanced but JWST was just chosen as a popular anecdote.

6.2 Comparing Real and Simulated Galaxies

It would also enable the original reason I wanted to find an automated, interpretable, unbiased way of finding similar looking galaxies, providing a framework for testing spirality prediction on simulated galaxies

One of the ways astronomers are trying to understand unobservable phenomena such as dark matter and dark energy is by simulating the Big Bang and running it until recent times. The Illustris and Illustris-TNG simulated galaxy surveys come to mind as one such effort. If we could find simulated galaxies that look very similar to real galaxies, astronomers could then compare and contrast the expected amount of dark matter in the simulated galaxy with the estimated amount in the real counterpart.

6.3 SpArcFiRe Galaxy Morphological Classification

Many papers that attempt to predict morphology do so for the broad categories from Hubble's 1926 paper[9]. Work such as Moonzarin Reza's 2021 paper [15] used random forest regression on SDSS images but only predicted for major classes of galaxies such as spiral, elliptical, or merging galaxies. However with our algorithm and a small dataset we were able to predict at 67% 10 fold cross validation accuracy for Hubble Types. It may be worth it in the future to examine how good SpArcFiRe features are at morphological classification with more effort. As said with some of the previous "future work" sections conceptually with how survey independent SpArcFiRe is, it would be straightforward to attempt to label JWST galaxies by their likelihood of being spiral just by replicating the results of other supervised learning efforts using SpArcFiRe features then applying that model to JWST images.

6.4 Cross Survey Search Engine

One last option to build another tool would be to use GSSD + SpArcFiRe's web interface to make a similar galaxy search engine. The least work version of this would be to simply input an image into the SpArcFiRe web interface, then receive back the names of a galaxy ordering that matches the input image. The more difficult version that would likely take more physical resources would be for us to run SpArcFiRe on more galaxy surveys and allow the user to find the same input galaxy among multiple surveys. An even harder version would be for us to authorize academic accounts to allow users to add galaxy images and SpArcFiRe outputs to a real-time database we create in order to build a living cross survey collection of easy to reference galaxies.

Bibliography

- [1] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. A. Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, J. C. Barentine, B. A. Bassett, A. C. Becker, T. C. Beers, E. F. Bell, V. Belokurov, A. A. Berlind, E. F. Berman, M. Bernardi, S. J. Bickerton, D. Bizyaev, J. P. Blakeslee, M. R. Blanton, J. J. Bochanski, W. N. Boroski, H. J. Brewington, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, L. N. Carey, S. Carliles, M. A. Carr, F. J. Castander, D. Cinabro, A. J. Connolly, I. Csabai, C. E. Cunha, P. C. Czarapata, J. R. A. Davenport, E. de Haas, B. Dilday, M. Doi, D. J. Eisenstein, M. L. Evans, N. W. Evans, X. Fan, S. D. Friedman, J. A. Frieman, M. Fukugita, B. T. Gänsicke, E. Gates, B. Gillespie, G. Gilmore, B. Gonzalez, C. F. Gonzalez, E. K. Grebel, J. E. Gunn, Z. Györy, P. B. Hall, P. Harding, F. H. Harris, M. Harvanek, S. L. Hawley, J. J. E. Hayes, T. M. Heckman, J. S. Hendry, G. S. Hennessy, R. B. Hindsley, J. Hoblitt, C. J. Hogan, D. W. Hogg, J. A. Holtzman, J. B. Hyde, S.-i. Ichikawa, T. Ichikawa, M. Im, v. Ivezić, S. Jester, L. Jiang, J. A. Johnson, A. M. Jorgensen, M. Jurić, S. M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, K. Konishi, R. G. Kron, J. Krzesinski, N. Kuropatkin, H. Lampeitl, S. Lebedeva, M. G. Lee, Y. S. Lee, R. F. Leger, S. Lépine, N. Li, M. Lima, H. Lin, D. C. Long, C. P. Loomis, J. Loveday, R. H. Lupton, E. Magnier, O. Malanushenko, V. Malanushenko, R. Mandelbaum, B. Margon, J. P. Marriner, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, H. L. Morrison, F. Mullally, J. A. Munn, T. Murphy, T. Nash, A. Nebot, E. H. Neilsen, H. J. Newberg, P. R. Newman, R. C. Nichol, T. Nicinski, M. Nieto-Santisteban, A. Nitta, S. Okamura, D. J. Oravetz, J. P. Ostriker, R. Owen, N. Padmanabhan, K. Pan, C. Park, G. Pauls, J. Peoples, W. J. Percival, J. R. Pier, A. C. Pope, D. Pourbaix, P. A. Price, N. Purger, T. Quinn, M. J. Raddick, P. R. Fiorentin, G. T. Richards, M. W. Richmond, A. G. Riess, H.-W. Rix, C. M. Rockosi, M. Sako, D. J. Schlegel, D. P. Schneider, R.-D. Scholz, M. R. Schreiber, A. D. Schwope, U. Seljak, B. Sesar, E. Sheldon, K. Shimasaku, V. C. Sibley, A. E. Simmons, T. Sivarani, J. A. Smith, M. C. Smith, V. Smolčić, S. A. Snedden, A. Stebbins, M. Steinmetz, C. Stoughton, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tanaka, M. Tegmark, L. F. A. Teodoro, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, Y. Wadadekar, S. Watters, D. H. Weinberg, A. A. West, S. D. M. White, B. C. Wilhite, A. C. Wonders, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, S. Zibetti, and D. B. Zucker. The seventh data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 182(2):543, may 2009.

- [2] H. Aihara, C. A. Prieto, D. An, S. F. Anderson, e. Aubourg, E. Balbinot, T. C. Beers, A. A. Berlind, S. J. Bickerton, D. Bizyaev, M. R. Blanton, J. J. Bochanski, A. S. Bolton, J. Bovy, W. N. Brandt, J. Brinkmann, P. J. Brown, J. R. Brownstein, N. G. Busca, H. Campbell, M. A. Carr, Y. Chen, C. Chiappini, J. Comparat, N. Connolly, M. Cortes, R. A. C. Croft, A. J. Cuesta, L. N. da Costa, J. R. A. Davenport, K. Dawson, S. Dhital, A. Ealet, G. L. Ebelke, E. M. Edmondson, D. J. Eisenstein, S. Escoffier, M. Esposito, M. L. Evans, X. Fan, B. F. Castellá, A. Font-Ribera, P. M. Frinchaboy, J. Ge, B. A. Gillespie, G. Gilmore, J. I. G. Hernández, J. R. Gott, A. Gould, E. K. Grebel, J. E. Gunn, J.-C. Hamilton, P. Harding, D. W. Harris, S. L. Hawley, F. R. Hearty, S. Ho, D. W. Hogg, J. A. Holtzman, K. Honscheid, N. Inada, I. I. Ivans, L. Jiang, J. A. Johnson, C. Jordan, W. P. Jordan, E. A. Kazin, D. Kirkby, M. A. Klaene, G. R. Knapp, J.-P. Kneib, C. S. Kochanek, L. Koesterke, J. A. Kollmeier, R. G. Kron, H. Lampeitl, D. Lang, J.-M. Le Goff, Y. S. Lee, Y.-T. Lin, D. C. Long, C. P. Loomis, S. Lucatello, B. Lundgren, R. H. Lupton, Z. Ma, N. MacDonald, S. Mahadevan, M. A. G. Maia, M. Makler, E. Malanushenko, V. Malanushenko, R. Mandelbaum, C. Maraston, D. Margala, K. L. Masters, C. K. McBride, P. M. McGehee, I. D. McGreer, B. Ménard, J. Miralda-Escudé, H. L. Morrison, F. Mullally, D. Muna, J. A. Munn, H. Murayama, A. D. Myers, T. Naugle, A. F. Neto, D. C. Nguyen, R. C. Nichol, R. W. O’Connell, R. L. C. Ogando, M. D. Olmstead, D. J. Oravetz, N. Padmanabhan, N. Palanque-Delabrouille, K. Pan, P. Pandey, I. Pâris, W. J. Percival, P. Petitjean, R. Pfaffenberger, J. Pforr, S. Phleps, C. Pichon, M. M. Pieri, F. Prada, A. M. Price-Whelan, M. Jordan Raddick, B. H. F. Ramos, C. Reylé, J. Rich, G. T. Richards, H.-W. Rix, A. C. Robin, H. J. Rocha-Pinto, C. M. Rockosi, N. A. Roe, E. Rollinde, A. J. Ross, N. P. Ross, B. M. Rossetto, A. G. Sánchez, C. Sayres, D. J. Schlegel, K. J. Schlesinger, S. J. Schmidt, D. P. Schneider, E. Sheldon, Y. Shu, J. Simmerer, A. E. Simmons, T. Sivarani, S. A. Snedden, J. S. Sobeck, M. Steinmetz, M. A. Strauss, A. S. Szalay, M. Tanaka, A. R. Thakar, D. Thomas, J. L. Tinker, B. M. Tofflemire, R. Tojeiro, C. A. Tremonti, J. Vandenberg, M. V. Magaña, L. Verde, N. P. Vogt, D. A. Wake, J. Wang, B. A. Weaver, D. H. Weinberg, M. White, S. D. M. White, B. Yanny, N. Yasuda, C. Yeche, and I. Zehavi. Erratum: “the eighth data release of the sloan digital sky survey: First data from sdss-iii” (2011, *apjs*, 193, 29). *The Astrophysical Journal Supplement Series*, 195(2):26, aug 2011.
- [3] A. Baillard, E. Bertin, V. de Lapparent, P. Fouqué, S. Arnouts, Y. Mellier, R. Pelló, J.-F. Leborgne, P. Prugniel, D. Makarov, L. Makarova, H. J. McCracken, A. Bijaoui, and L. Tasca. The efigi catalogue of 4458 nearby galaxies with detailed morphology. *Astronomy & Astrophysics*, 532:A74, July 2011.
- [4] D. R. Davis and W. B. Hayes. Sparcfire: scalable automated detection of spiral galaxy arm segments. *The Astrophysical Journal*, 790(2):87, 2014.
- [5] M. Davis and M. J. Geller. Galaxy correlations as a function of morphological type. *Astrophysical Journal*, Vol. 208, pp. 13-19 (1976)., 208:13–19, 1976.
- [6] G. De Vaucouleurs. Classification and morphology of external galaxies. In *Astrophysik iv: Sternsysteme/astrophysics iv: Stellar systems*, pages 275–310. Springer, 1959.

- [7] D. M. Elmegreen and B. G. Elmegreen. Flocculent and grand design spiral structure in field, binary and group galaxies. *Monthly Notices of the Royal Astronomical Society*, 201(4):1021–1034, 12 1982.
- [8] A. Gauthier, A. Jain, and E. Noordeh. Galaxy morphology classification. *Lecture Notes, Stanford University*, 16, 2016.
- [9] E. P. Hubble. Extragalactic nebulae. *The Astrophysical Journal*, 64:321–369, December 1926. Provided by the SAO/NASA Astrophysics Data System.
- [10] G. Kauffmann, T. M. Heckman, S. D. M. White, S. Charlot, C. Tremonti, J. Brinchmann, G. Bruzual, E. W. Peng, M. Seibert, M. Bernardi, M. Blanton, J. Brinkmann, F. Castander, I. Csábai, M. Fukugita, Z. Ivezic, J. A. Munn, R. C. Nichol, N. Padmanabhan, A. R. Thakar, D. H. Weinberg, and D. York. Stellar masses and star formation histories for 105 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 341(1):33–53, 05 2003.
- [11] S. M. Kent. Sloan digital sky survey. *Astrophysics and Space Science*, 217:27–30, 1994.
- [12] C. Lin and F. H. Shu. On the spiral structure of disk galaxies, ii. outline of a theory of density waves. *Proceedings of the National Academy of Sciences*, 55(2):229–234, 1966.
- [13] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178, Jan. 2011.
- [14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [15] M. Reza. Galaxy morphology classification using automated machine learning. *Astronomy and Computing*, 37:100492, 2021.
- [16] M. C. Sánchez-Gil, D. H. Jones, E. Pérez, J. Bland-Hawthorn, E. J. Alfaro, and J. O’Byrne. Age patterns in a sample of spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, 415(1):753–772, 07 2011.
- [17] C. A. Tremonti, T. M. Heckman, G. Kauffmann, J. Brinchmann, S. Charlot, S. D. M. White, M. Seibert, E. W. Peng, D. J. Schlegel, A. Uomoto, M. Fukugita, and J. Brinkmann. The origin of the mass-metallicity relation: Insights from 53,000 star-forming galaxies in the sloan digital sky survey. *The Astrophysical Journal*, 613(2):898, oct 2004.
- [18] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018.