UNIVERSITY OF CALIFORNIA,
IRVINE

Machine Learning for the Sciences

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Kevin M. Bache

Dissertation Committee:
Professor Pierre Baldi, Chair
Professor Christopher C.W. Hughes
Professor Xiaohui Xie

2017

DEDICATION

To my very many teachers

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vii**

LIST OF ALGORITHMS **viii**

ACKNOWLEDGMENTS **ix**

CURRICULUM VITAE **xi**

ABSTRACT OF THE DISSERTATION **xiii**

1 Introduction **1**
- 1.1 Dissertation Outline and Contributions . . . 2
  - 1.1.1 Text-Based Measures of Document Diversity . . . 2
  - 1.1.2 The Cosmic Prevalence of Quenching in Low Mass Satellites . . . 3
  - 1.1.3 Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images . . . 3

2 Text-Based Measures of Document Diversity **4**
- 2.1 Introduction . . . 4
- 2.2 Related Work . . . 7
  - 2.2.1 Interdisciplinarity in Scientometrics . . . 7
  - 2.2.2 Diversity as Outlier Detection . . . 9
  - 2.2.3 Diversity in Information Retrieval . . . 10
- 2.3 Defining Topic-Based Diversity . . . 10
  - 2.3.1 Topic Co-occurrence Similarity . . . 12
  - 2.3.2 Topic-Word Similarity . . . 13
  - 2.3.3 From Similarity to Distance . . . 13
- 2.4 Data Sets and Topic Models . . . 14
  - 2.4.1 Data Sets . . . 14
  - 2.4.2 Topic Modeling . . . 15
- 2.5 Pseudo-Document Experiments . . . 15
  - 2.5.1 Pseudo-Documents . . . 15
  - 2.5.2 Experiments . . . 16

2.6 Detecting Diverse Documents 21
2.7 Conclusions 24

3 The Cosmic Prevalence of Quenching in Low Mass Satellites 25
3.1 Introduction 25
3.2 Observations 27
3.3 Satellite Characterization 29
3.3.1 Satellite Finding 29
3.3.2 Star Formation Modeling 34
3.3.3 Mass Distribution Sensitivity Analysis 38
3.3.4 Algorithm Efficacy 39
3.4 Precision and Recall 40
3.5 Results and Discussion 43

4 Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images 47
4.1 Introduction 47
4.2 Methods 51
4.2.1 Data Collection 51
4.2.2 Preprocessing 51
4.2.3 Human Ratings 53
4.2.4 Loss Weighting 54
4.2.5 Training Procedure 55
4.2.6 Linear Models 57
4.2.7 Convolutional Neural Network Models 57
4.2.8 Hyperparameter Search for Convolutional Architectures 58
4.2.9 Pre-Trained Convolutional Architecture 59
4.2.10 Custom Convolutional Architecture 59
4.3 Results 61
4.3.1 Human Rating Results 61
4.3.2 Linear Model Results 61
4.3.3 Pre-Trained Convolutional Neural Network Results 62
4.3.4 Custom Convolutional Neural Network Results 62
4.4 Discussion 63
4.5 Conclusion 66

Bibliography 67

LIST OF FIGURES

Page

2.1 Histograms of topic-topic distances for $\delta(i,j) = 1 - s_c(i,j)$ and $\delta(i,j) = 1/s_c(i,j)$. 17
2.2 Pseudo-document ROC curves for PubMed data with 100 topics comparing Rao diversity to alternate methods. See also Table 2.1. 18
2.3 Two of the most diverse NSF grant proposals. 20
2.4 Two of the least diverse NSF grant proposals. 21
2.5 Two of the most diverse PubMed OA articles. 23
2.6 High diversity (top) and low diversity (bottom) ACL articles. 23

3.1 Modified specific star formation rate (log SFR - .7 log Stellar Mass) plotted against stellar mass for objects in the SDSS training sample. The horizontal line is the dividing line between star forming and quenched galaxies. 30
3.2 Absolute r band magnitude plotted against redshift for galaxies in the training set. The red points in the left panel represent passive galaxies, the blue points in the right panel represent star-forming galaxies. Completeness curves are shown for both the training set (black curve) and the test set (magenta curve). The magenta square shows the test set completeness derived from [27] (see text for discussion). Vertical black lines indicate the redshifts at which the test set is complete to the labeled mass. 31
3.3 Counts of galaxies $16.5 < r < 18.5$ in a circular aperture of radius 9 arcmin around hosts (as shown on left), plotted versus distance from the host (solid line). Also shown are counts around random points on the sky (dashed line). This figure serves to demonstrate the validity of the technique of measuring satellite statistics by statistically subtracting the background. Note that at large distances from the hosts, the counts come into agreement with the random background. 32
3.4 The correction factor, defined as the ratio of precision to recall, plotted against stellar mass for each of our two algorithms. The red squares and blue circles represent the correction factor derived from narrowly binning the data in the DT and NN algorithms, respectively. From these data, we fit a linear correction function for each algorithm (dashed lines). Note that the algorithms were tuned such that correction factors near unity were produced. 41

3.5 The quenched fraction of satellite galaxies as a function of satellite mass, derived from our two algorithms. Red squares denote quenched fractions measured using the *decision tree* algorithm, while blue circles denote those using the *neural network* algorithm (see §3.3). Our results indicate an elevated quenched fraction at low masses, in broad agreement between the two algorithms. 42
3.6 The quenched fraction of satellite galaxies as a function of satellite mass, plotted alongside previous results, as well as LG objects. As before, red squares denote quenched fractions measured using the *decision tree* algorithm, while blue circles denote those using the *neural network* algorithm (see §3.3). Taken together, these results show that low-mass satellites are significantly more susceptible to being environmentally quenched than high-mass satellites. 44

4.1 Example vessel images 50
4.2 A set of blood vessel images before (left) and after (right) alignment. The pre-drug-application images are placed in the image's green channel and the post-drug-application images are placed in the red channel. The separate green and red vessels in the left image shows that the pre- and post-drug-application images are misaligned, the more pervasive yellow in the right image comes from the green and red channels being aligned on top of each other. . 53
4.3 Three examples of the data augmentation process used for training and inference. The top image is an actual training image, and the bottom three are randomly transformed copies of that image. Each time an image is visited during the training process, it is first randomly transformed in a way that simulates creating new images with respect to the true invariances of the training images (e.g.: an image should have the same class as a copy of that image which is slightly shifted, rotated, or flipped). The left-most randomly generated image has been flipped horizontally, zoomed, and rotated slightly. The middle random image has been flipped both horizontally, vertically, and zoomed slightly. The right-most image has been flipped vertically, zoomed in, and translated down slightly. This random augmentation helps simulating a larger training set and prevent our model from overfitting. 56
4.4 The architecture for the best convolutional neural network we trained on these data. The blue prisms represent the 3-dimensional input images (two channels, width, and height) and the three dimensional output of each convolutional layer (filters, width, and height). The green prisms represent a sample receptive field for the subsequent convolutional layer. 63
4.5 Receiver operating characteristic curves for a binarized version of this classification problem (*no-hit* vs. *soft-hit* or *hard-hit*). ROC-AUC scores range between 0.5 and 1.0, with 0.5 indicating performance at chance and 1.0 indicating perfect classification (a standard which the best custom convolutional neural network we tried achieves on this binarized problem). 64

LIST OF TABLES

Page

2.1 AUC scores for different diversity measures based on 1000 pseudo-documents from PubMed. 17
2.2 AUC scores for pseudo-documents from specific journal pairs from PubMed. 18

3.1 Number of datapoints in resampled versions of the second validation set. . . 38
3.2 AUC and accuracy measures for neural net and random forest models on several versions of the second validation set. 38

4.1 Loss function weight values 54
4.2 Test Set Confusion Matrix for Average of Four Human Raters 61
4.3 Test Set Confusion Matrix for Linear Ensemble 61
4.4 Test Set Confusion Matrix for Pre-Trained Convolutional Ensemble 62
4.5 Test Set Confusion Matrix for Custom Convolutional Ensemble 63

LIST OF ALGORITHMS

		Page
1	Outer-Loop Hyperparameter Optimization	35

ACKNOWLEDGMENTS

I would like to thank

My parents Chris and Carol,
Whose unequivocal care and diligent examples taught me what it is to live from the heart
and inspired me to work hard from an early age,

My step parents, Christina and David,
Who showed me that love can take varied and wonderful forms,

My siblings Jason and Lara
and extended siblings Adrienne, Ben, Nate, Grace, and Micaela,
Whose support and ongoing tom-foolery helped shape me
into the whole human being that I am today,

Dae Il and George,
Who urged me to follow this dream and modeled its pursuit with style,

My many close friends at UCI including Sky, Gabe, Patricia, Mehdi, Kristin, Maeraj, Michael, Theano, Justin, Chris, Lowell, Olivia, Guggis, Malorie, Dimitris, Zach, Levi, Yue, Alex, Sylvinia, Shae, Coral, Maryna, Andy, Mike, Sonja, George, Craig, Ekin, Cesar, Chris, Drew, Andrew, Jasmine, Zhen, Mary, Nick, ZJ, Eugenia, Zach, Joanna, Robin, Corey, and countless others,
We kept each other fed, sane, happy, and thankfully never burned the house down,

Jimmy,
Who models the pure pursuit of knowledge with a true human goodness,

Moshe,
Who quashes fear, speaks truth, and loves fun,

John,
For your relentless pursuit of truth, for helping me find my voice, for changing me,

Kyle,
For years of unspeakable fun, for countless adventures, for following and leading me,
for your unyielding selflessness, and for knowing me all the way through,

My benefactors,
Who helped make this dream possible,

Steve Sellinger and the crew of The Vessel Triumph,
For an unexpected lift and the world's most welcome burger,

Cindy, Holly, and Karina,
Who keep us all honest,

My many excellent professors at UC Irvine,
Whose efforts make all our paths easier,

My committee,
A trio of first-rate academics,

My co-authors,
Who served as mentors and peers on my meandering journey,

My summer mentors, Dennis, John, Chengu, and Youngmin,
Who stoked my confidence and inspired me to push on,

Prof. Padhraic Smyth,
Who guided me into the world of machine learning,
and whose inveterate skepticism will always underpin my data analysis,

And finally **Prof. Pierre Baldi**,
Who understood me at a glance. Without your able guidance, persistent determination,
and genuine human kindness I never could have gotten here.
Thank you.

Sections of this dissertation are reprinted material as they appear (or are planned to appear) in the Proceedings of the ACM's Special Interest Group on Knowledge Discovery and Data Mining, the Proceedings of Technology Management for the Interconnected World, the Proceedings of the International Conference on Learning Representations, the IEEE/ACM Transactions on Computational Biology and Bioinformatics, and the Monthly Notices of the Royal Astronomical Society. The co-authors of those papers helped conceive and execute the research upon which this dissertation is based.

This work was supported by National Science Foundation IIS-1550705, Defense Advanced Research Projects Agency D17AP00002, National Institutes of Health (NIH) R01 CA180122 (PQD5), NIH UH3 TR000481, an NVDIA Corporation hardware award, the Google Faculty Research Award, the Chao Family Comprehensive Cancer Center (CFCCC) through NCI Center Grant award P30A062203, Intelligence Advanced Research Projects Activity via Dept. of Interior National Business Center contract D11PC20155.

CURRICULUM VITAE

Kevin M. Bache

EDUCATION

Doctor of Philosophy in Computer Science **2017**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Psychology **2007**
Harvard University *Cambridge, MA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2011–2015**
University of California, Irvine *Irvine, California*

Research Associate **2008–2011**
Kiehl Lab, Mind Research Network *Albuquerque, NM*

Research Associate **2006-2007**
Lazar Lab, Massachusetts General Hospital *Boston, MA*

TEACHING EXPERIENCE

Teaching Assistant **2016–2017**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images — **In Review**
IEEE/ACM Transactions on Computational Biology and Bioinformatics

The Cosmic Prevalence of Quenching in Low Mass Satellites — **In Submission**
Monthly Notices of the Royal Astronomical Society

Premotor functional connectivity predicts impulsivity in juvenile offenders — **2011**
Proceedings of the National Academy of Sciences

REFEREED CONFERENCE PUBLICATIONS

Hot Swapping for Online Adaptation of Optimization Hyperparameters. — **2015**
International Conference on Learning Representations

An exploratory study of interdisciplinarity and breakthrough ideas. — **2013**
Technology Management in the IT-Driven Services (PICMET)

Text-based measures of document diversity — **2013**
Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

ABSTRACT OF THE DISSERTATION

Machine Learning for the Sciences

By

Kevin M. Bache

Doctor of Philosophy in Computer Science

University of California, Irvine, 2017

Professor Pierre Baldi, Chair

Progress in the sciences depends critically on the analysis of ever-growing bodies of data. Many of these analysis patterns are inferential in nature; their goal is to infer the value of one or more parameters which bear some real-world meaning. Others are in essence discriminative; their goal is to build a black-box model with the strongest possible predictive power. For both of these analysis styles, machine learning offers a host of powerful tools to tackle historically unapproachable problems.

In this dissertation, I present three examples of machine learning tools applied to the sciences. The first offers a novel model of textual diversity applied to the science of science itself. The second, explores a series of discriminative models which probe the evolution of the cosmos. The third offers a novel convolutional neural architecture for discriminating effective from ineffective drug candidates.

Taken together, these studies offer a glimpse of the breadth and potency of the contributions that machine learning can offer to the sciences.

Chapter 1

Introduction

The process of scientific induction depends critically on the analysis of rapidly growing bodies of data. While this process has historically been governed by traditional statistics, scientists increasingly turn to machine learning to solve their most data-intensive tasks in fields across the breadth of the sciences: from physics [2] to biology [15], medicine [74] through the social sciences [42].

Many of these learned models lean probabilistic, in which the researcher is often interested in the values of specific parameters for their real-world technical meanings [39]. One of the most successful classes of probabilistic models is topic models [6]. A topic model decomposes a corpus of documents into a series of topics, each defined as a probability distribution over all known words in the corpus, and a series of topic mixture probabilities, one per document. These topics can then be used to summarize information about the corpus or specific documents within it.

At the other end of the machine learning spectrum lie discriminative models, in which interpretability is subsumed beneath the goal of pure black-box predictive quality. In many ways, the most successful modern predictive paradigm is deep learning, which finding applications

in industry and across the sciences [44]. In deep learning, a sequential series of nonlinear transformations of input data are learned all at once, preempting the need for developing separate, intensive preprocessing and learning steps.

In this dissertation, I present three studies which span much of the stylistic and topical breadth of machine learning. The first is a new model of textual diversity applied to the science of science. The second two are applications of deep learning to problems in the sciences.

1.1 Dissertation Outline and Contributions

1.1.1 Text-Based Measures of Document Diversity

In this section, we build a new model of textual diversity as a proxy for interdisciplinarity in the sciences. We first define what we mean by diversity; a set of objects is considered diverse if it contains large portions of members from highly-disparate sub-groups. Next we develop a means of quantifying this notion of diversity across a corpus of documents. Our approach is to first use a topic model to decompose a corpus of documents into series of topics and topic mixture portions. Next, we define a distance measure between each pair of topics based on their co-occurrence rates within documents in the corpus. To complete the diversity model, we use the topic distribution and topic distance measures to calculate the diversity of each document, confirming that the documents which our system identifies as diverse and non-diverse match human intuition. Finally, we define diverse and non-diverse pseudo-documents and find that our notion of diversity is better capable of differentiating the two than simpler, competing notions.

1.1.2 The Cosmic Prevalence of Quenching in Low Mass Satellites

In this section, we develop a series of random forest and neural network models to predict whether a galaxy is quenched (i.e.: no longer producing new stars) from its spectroscopic readings. These models allow us to study galaxy quenching in distant objects at lower masses than have traditionally been possible. We conclude that observations of distant objects match those of our local group of galaxies: galaxy quenching increases as mass goes down.

1.1.3 Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images

In this section, we develop a deep convolutional neural architecture to aid in drug discovery. The network is designed to distinguish between images of microvasculature networks grown in microphysiological systems which have and have not been disrupted by the application of an effective drug candidate. This system achieves super-human classification performance, and fills a gap in a new, high-quality, high-throughput drug screening pipeline.

Chapter 2

Text-Based Measures of Document Diversity

2.1 Introduction

The quantification of diversity has been widely studied in areas such as ecology [48], genetics [56], linguistics [45], and sociology [32]. The typical context is where one wishes to measure the diversity of a population, where a population consists of a set of individual elements that have been categorized into T types (such as species), with proportions $\pi = \{p_1, \ldots, p_T\}$ and $\sum_{i=1}^{T} p_i = 1$.

A relatively simple measure of diversity is *variety*, how many different species are present in a population, or the number of non-zero proportions in π. One can alternatively measure diversity as a function of the relative *balance* among the proportions (also referred to as 'evenness' in ecology [62] or 'concentration' in economics [30]), using measures such as Shannon entropy $H(\pi) = -\sum_{i=1}^{T} p_i \log p_i$ or variance-based quantities such as $\sum_{i=1}^{T} p_i(1-p_i) = 1 - \sum_{i=1}^{T} p_i^2$ (e.g., [75]). The intuition is that higher entropy or variance implies greater population diversity

(e.g., see [69]).

From a more general perspective, Stirling [80] proposed that there are three distinct aspects to diversity: *variety*, *balance*, and *disparity*. *Disparity* is the extent to which the categories that are present are different from each other, based for example on distance within a known taxonomy [78]. For example, a population with 5 beetles and 5 elephants would be considered more diverse than a population with 5 beetles and 5 spiders, given that beetles and elephants are more taxonomically distant than beetles and spiders. Stirling argued that each of these three properties is a necessary (but non-sufficient) component in any quantitative characterization of diversity, arriving at a relatively simple mathematical formulation for diversity, a formulation originally proposed in earlier work by Rao [67]:

$$div = \sum_{i=1}^{T} \sum_{j=1}^{T} p_i p_j \ \delta(i,j) = \pi^t \Delta \pi \tag{2.1}$$

where p_i, p_j are the proportions of category i and j in the population, $\delta(i,j)$ is the distance between categories i and j, Δ is a $T \times T$ matrix of such distances, and π^t is the transpose of the $T \times 1$ vector of proportions π.

This diversity measure div has a simple and intuitive interpretation as the expected distance between two randomly selected elements of the population. The probability of selecting a pair of elements with replacement from categories i and j is $p_i p_j$. Thus, div can be interpreted as the expected value of the categorical distance, $E[\delta(i,j)]$, where the expectation is with respect to the distribution of pairs of elements.

The contribution of this present paper is to investigate diversity in the context of text documents, using Rao's measure a starting point. In particular, we will use words as elements, topics as word categories, and documents as collections (or "populations") of words. Specifically, we address the following task: given a corpus of documents, assign a diversity score to each document, where this diversity score can be used to rank documents from most to

least diverse.

There are a number of different practical problems where quantifying the topical diversity of documents in this manner is potentially useful. One specific area of application is in science policy. There is broad interest among science policy experts in diversity and interdisciplinarity in scientific research. In particular, there is interest in the hypothesis that interdisciplinary research can lead to new discoveries at a rate faster than that of traditional research projects conducted within single disciplines. Indeed, the United States National Science Foundation (NSF) encourages interdisciplinary proposals, and has put out solicitations for proposals that include specific combinations of disciplines. One such example was the recent NSF program "Collaboration in Mathematical Geosciences" (CMG), which was focused on research at the intersection of mathematics and geoscience. In this context an automated diversity measure would be potentially helpful in evaluating the diversity of submitted proposals during the review process. Furthermore, being able to quantify the diversity of papers that resulted from funding under such a program, compared to papers funded by traditional single-discipline programs, would be useful as a component in overall evaluation of the effectiveness of interdisciplinary research programs.

Similarly in scientometrics and bibliometrics, there is significant interest in developing quantitative measures of interdisciplinarity for both individual scientific articles as well as collections of articles such as journals (e.g., [88]). Further afield, one can envision tools that allow researchers to explore and rank the diversity of individual papers and journals, and for administrators (such as department chairs, deans, and heads of research labs) to quantify the diversity of the research in their departments and labs relative to other institutions.

We begin in Section 2.2 by discussing related work. Section 2.3 outlines a number of possible diversity measures based on topic models. Section 2.4 describes the text corpora and the topic modeling approach we use in the paper. In Section 2.5 we describe a set of experiments based on pseudo-documents which serve as a proxy for ground truth and allow us to evaluate

the performance of different text-based diversity measures. Section 2.6 discusses several examples of both high and low diversity scientific articles and grant abstracts detected by our approach, and Section 2.7 concludes the paper.

2.2 Related Work

2.2.1 Interdisciplinarity in Scientometrics

There has been a significant amount of work in the field of scientometrics on quantifying notions of *interdisciplinarity* as reflected in the output of scientific research (e.g., via published scientific articles). The 2005 *National Academies Committee on Facilitating Interdisciplinary Research* defined interdisciplinarity from an operational viewpoint as a "mode of research that integrates .... concepts ... tools ... data ... from two or more bodies of knowledge or research practice" [64]. Diversity in this context (e.g., diversity of citations or diversity of text content) can be thought of as a broader construct than interdisciplinarity, but one which serves as a useful proxy for it. Indeed, diversity as defined via co-citation counts is the most widely-used approach to quantify interdisciplinarity in practice, based on the notion that disciplines that are co-cited more often by the same article are "closer" than disciplines that are less frequently co-cited. *Journal subject categories* are typically used to capture the notion of a *discipline*, typically using the manually-defined 244 ISI subject categories from Thomson Reuters, with articles being assigned to a subject category associated with the journal the article is published in (e.g., [64, 63, 66, 88]).

Rafols and Porter [63] used journal subject categorizations of citations to analyze how interdisciplinarity has changed between 1975 and 2005 for six specific subject-categories. They concluded that although the number of citations and co-authors per paper was increasing significantly over time, the degree of interdisciplinarity was increasing at a much slower rate,

as reflected by citation patterns between subject categories. As a component in their analysis, Rafols and Porter used Rao's diversity index based on a count matrix of D documents by T categories derived from citations: p_i was the proportion of citations made by an article to other articles that were published in journals belonging to subject category i, and $\delta(i, j)$ was defined as 1 minus the cosine distance between citation count vectors (across documents) of subject categories i and j.

Our work differs from this earlier work and related threads in scientometrics in two specific ways. First, in our approach the categories and distances, $\delta(i, j)$, are learned directly from the text content, rather than being based on manually predefined schema such as the ISI subject categories. There are obvious limitations to relying on pre-defined taxonomies, as pointed out by Rafols and Porter [64]. Subject categories can change over time and no longer necessarily reflect current disciplinary boundaries. In addition, in some contexts such as analysis of proposals and grants, there may be very limited or no categorizations available. For analysis of narrow domains (say the field of data mining and machine learning) existing categorization schemes may be too coarse-grained to be useful. In this context, a corpus-driven approach to learning the categories, such as the topic-based method we describe here, is a useful alternative, and in some cases may be the only option.

The second major difference in our approach is our use of word counts rather than citation counts (which are the basis of most prior work in scientometrics on quantifying interdisciplinarity). We expect that using text content will complement citation-based approaches, as both words and citations carry useful signal. There has long been debate over whether citations accurately reflect the content of a scientific article [18, 11]—arguably the words in an article may provide a more accurate reflection of the author's intentions than the citations the author uses. A systematic approach to the use of *both* word-based and citation-based measures of diversity would also be worth exploring in future work—in this paper, however, we limit our attention to the exploration of word-based measures.

2.2.2 Diversity as Outlier Detection

Another field which is related to our current work is that of outlier detection. If we consider documents as being represented by T-dimensional vectors of counts, then one approach to quantifying diversity is to look for documents that are outliers in this T-dimensional space, using a multivariate outlier detection algorithm. Typically these algorithms rely on a notion of global or local density, e.g., by finding data points that have low-probability under a global distribution or that are relatively distant from their nearest neighbors.

In addition to the usual issues associated with estimating distances and densities in high dimensions, a further complication in diversity characterization is that we are seeking low-probability data points with the constraint that we are not interested in solutions where all of the probability mass is on a single component, i.e., where $p_i \approx 1, p_j \approx 0, j \neq i$. Equivalently, since the p_i are the components of a probability vector in a $T-1$ dimensional simplex, we can think of high diversity documents as points that lie in the interior of the simplex (in at least 2 of the dimensions) rather than at the edge.

Although it might be possible to develop a principled approach to characterizing diversity in this way, e.g., by a constraint-based approach to outlier detection, the use of Rao's measure bypasses both the problem of estimating a high-dimensional distribution and the problem of constraining points of interest to lie in the interior of the simplex. In particular, we can view Rao's measure as a form of outlier detection based on second-order information, focusing on pairwise dependencies among the columns of the count matrix, via the $\delta(i, j)$ term, combined with a term $p_i p_j$ that penalizes count vectors consisting of a single dominant component.

2.2.3 Diversity in Information Retrieval

A third potentially relevant source of prior work is in information retrieval and search where one wishes to generate a diverse list of search results in response to a user query (e.g., to avoid showing similar items in a list of search results). This work has a somewhat different motivation than the one we pursue in this paper. In the typical search context, diversity is closely aligned with making inferences about users' goals, i.e., trying to find a diverse group of documents such that the probability is maximized that at least one of the documents matches a user's implicit goals (e.g., [95]) or maximizing some notion of coverage (e.g., [33]). In contrast, the focus in this paper is on characterizing the inherent topical diversity of single documents, rather than finding a group of documents that best fulfill a user's information need.

2.3 Defining Topic-Based Diversity

In the general case we consider a count-matrix representation for a corpus of D documents, where each row indexed by $d, 1 \leq d \leq D$, represents a document, each column $j, 1 \leq j \leq T$, represents a category, and each entry indexed by (d, j) in the matrix represents how many elements in document d belong to category j. In particular, in this paper we focus on word tokens as the elements of a document, and a learned set of topics as the categories to which elements have been assigned.

We use the Latent Dirichlet Allocation (LDA) topic model with collapsed Gibbs sampling to learn T topics for the D documents in the corpus [34]. A single iteration of the collapsed Gibbs sampler consists of iterating through the word tokens in the corpus, sequentially sampling topic assignments for each word token in each document while keeping all other topic-word assignments fixed. Using the topic-word assignments from the final iteration of the Gibbs

sampler[1] , we create a $D \times T$ *document-topic* count matrix with entries n_{dj} corresponding to the number of word tokens in document d that are assigned to topic j.

In this context we can define Rao's diversity measure for each document d as

$$div^{(d)} = \sum_{i=1}^{T} \sum_{j=1}^{T} P(i|d)P(j|d)\delta(i,j) \tag{2.2}$$

where $P(j|d)$ is the proportion of word tokens in document d that are assigned to topic j (estimated as $\frac{n_{dj}}{n_d}$ where n_d is the number of word tokens in d) and $\delta(i,j)$ is a measure of the distance between topic i and topic j. Note that $\delta(i,j)$ is constant across all documents, and $P(i|d)$ and $P(j|d)$ vary from document to document.

The interpretation of Equation 2.2 is intuitive: if we randomly select a pair of words from document d (with replacement), then $div^{(d)}$ is the *expected topical distance between a pair of words in document d*. Thus, a document that has two topics that are far away from one another, each with a large proportion of the word tokens assigned to them, will have a high diversity score. Conversely, documents whose word tokens are assigned to topics that are all relatively close to one another, or whose word tokens predominantly fall into a single topic, will earn a lower diversity score.

There are a number of possible approaches to defining distances between topics $\delta(i,j)$. We explore below a number of different pairwise measures of similarity between topics, $s(i,j)$, as well as different methods of transforming these similarities into distances. We begin with topic similarity functions based on *topic co-occurrence* in documents, as defined by the $D \times T$ matrix of *document-topic* counts. An alternative approach that we also explore is topic similarity based on the similarity of *topic-word* distributions using the $W \times T$ *word-topic* count matrix.

[1] An alternative approach would be to average over multiple samples and use expected counts in the document-topic count matrix rather than actual counts from the final sample.

2.3.1 Topic Co-occurrence Similarity

A straightforward measure of topic similarity based on co-occurrence within documents is the cosine distance of columns in the $D \times T$ matrix of *document-topic* counts. This is defined as

$$s(i,j) \equiv \frac{\sum_d n_{di} n_{dj}}{\sqrt{\sum_d n_{di}^2}\sqrt{\sum_d n_{dj}^2}} \tag{2.3}$$

where i and j represent two column indices (two topics) and $\sum_d$ is a sum over all documents indexed by d.

Other similarity measures can also be used. For example, consider randomly selecting two word tokens with replacement from within a randomly selected document d in the corpus. Let $s(i,j) = P(w_1 = i, w_2 = j)$ be the probability that the first word token w_1 is assigned to topic i and the second word token w_2 is assigned to topic j:

$$\begin{aligned} P(w_1 = i, w_2 = j) &= \sum_d P(w_1 = i, w_2 = j|d)P(d) \\ &= \sum_d P(j|d)P(i|d)P(d) \end{aligned} \tag{2.4}$$

where $P(d)$ is the probability of a random word belonging to document d and is estimated using $\frac{n_d}{N}$ where N is the number of word tokens in the corpus. In estimating $P(j|d)$ and $P(i|d)$ above we use smoothed maximum a posteriori estimates, with hyperparameter values from the Dirichlet prior on the document-topic multinomials in the topic model. The use of smoothed estimates produces non-zero similarities $P(w_1 = i, w_2 = j)$ for all pairs of topics i and j, avoiding singularities in the corresponding distances $\delta(i,j)$ and diversity measures. The conditional version of the expression above, $P_C(w_2 = j|w_1 = i)$ can be viewed as a topic-based version of the contextual word distribution defined by Dillon et al. [20], defined as the probability that one word is present in a document given that another word is also in the

same document.

2.3.2 Topic-Word Similarity

An alternative strategy to using topic co-occurrence is to consider topic similarity based on topic-word distributions. Similarity can be defined in exactly the same manner as above, but now using the $W \times T$ *word-topic* count matrix instead of the $D \times T$ *document-topic* count matrix, where W is the number of words in the model's vocabulary. In the context of measuring diversity, it is interesting to consider whether the *document-topic* or *topic-word* similarity is likely to be more useful. One can imagine situations where two topics have relatively different distributions over words (low similarity in *topic-word* distributions), yet the same two topics co-occur relatively frequently across documents (high similarity in *document-topic*). From a diversity perspective, documents that contain these two topics should in principle not be diverse, yet the *word-topic* similarity measure would indicate that they are since their word distributions are different. In our experimental results we explore this further and report results using diversities computed from both the *document-topic* (DT) and *word-topic* (WT) matrices.

2.3.3 From Similarity to Distance

We empirically investigated two different transformations to convert each similarity measure into a distance measure: $\delta(i, j) = 1 - s(i, j)$ and $\delta(i, j) = 1/s(i, j)$. We also investigated the effectiveness of $\delta(i, j) = -\log s(i, j)$ but found that it did not provide a performance gain over the other transformations.

2.4 Data Sets and Topic Models

2.4.1 Data Sets

The PubMed Central Open Access dataset (PubMed) is comprised of articles published in biomedical journals which are freely available under a creative commons license [55]. We collected approximately 228k articles which were published between the dataset's inception in 1996 and our collection date in mid-2010. We focused our efforts on a subset of approximately 165k articles for which full text was available. Each document contained a title, the name of the journal in which it was published, its year of publication, and names of its authors. We eliminated approximately 20k documents which had either fewer than 600 words or more than 10,000 words, yielding a collection of approximately 145k documents.

Our second data set is a collection of 74k NSF Awards from 2007 to 2012 gathered from www.nsf.gov/awardsearch. Each record includes the title and abstract of the award, as well as various metadata such as the NSF Directorate, Division and Program that funded the award. We eliminated approximately 12k documents which had duplicate titles, followed by an additional 10k which had fewer than 70 words or more than 1,000, resulting in a final set of 52k documents.

As a third data set we used the Association of Computational Linguistics Anthology Network (ACL) [65], consisting of papers published in selected computational linguistics conferences. This corpus contains the full-text of approximately 19k papers appearing at these conferences over a time span of more than four decades, in addition to each document's title, year, and conference of publication. We eliminated approximately 7k documents which were published as workshop papers, and an additional 1k which had fewer than 600 words or more than 10,000 words, yielding a collection of approximately 11k documents.

2.4.2 Topic Modeling

We performed simple tokenization and topic modeling on each of the three text corpora using MALLET [51]. This involved splitting on whitespace, removing punctuation and lowercasing, and converting into a bag-of-words representation using MALLET's default stopword list.

We then learned an LDA topic model with a fixed symmetric prior β over the word-topic distributions, and optimized the prior α over the document-topic distributions. The β prior was set to 0.01 and we initialized the α prior over the document-topic distributions at $0.05\frac{N}{DT}$, where N is the number of tokens in the dataset, D is the number of documents in the dataset, and T is the number of topics defined in the model. We enabled hyperparameter optimization every 10 iterations, and ran each Gibbs sampler for a total of 5,000 iterations, keeping only the final sample in the chain. For each dataset, we learned models with T = 10, 30, 100 and 300 topics.

2.5 Pseudo-Document Experiments

2.5.1 Pseudo-Documents

A significant challenge in evaluation is that there is no ground-truth measure for a document's diversity. To address this problem, we created artificial 'pseudo-documents,' half of which were designed to have *high* diversity and half of which were designed to have *low* diversity.

We create each pseudo-document by combining two actual documents into one pseudo-document in the following fashion. We begin by manually selecting two journals A and B with relatively unrelated (e.g., *The Journal of Cell Biology* and *The Journal of Foot and Ankle Research*). A pseudo-document is created by randomly selecting one article from jour-

nal A and one article from journal B, which we denote as *parent documents*. A *child pseudo document* is then created by computing the average of each *parent document's* bag of topic counts, rounded to the nearest count. If the parent journals, A and B, are relatively dissimilar in content, we expect the resulting pseudo-documents to be relatively diverse. We can also create low-diversity pseudo-documents by repeating the above process but now selecting both parent articles from the same journal. By labeling pseudo-documents as having *high* or *low* diversity in this manner, we can create a proxy for ground truth diversity for evaluation purposes. This approach will not necessarily be perfect: for example, it is possible that if one of the journals contains documents that span diverse topics (relative to the corpus as a whole) some of the pseudo-documents labeled as low-diversity by this method could have relatively high actual diversity. However, even though such mislabeling could occur in theory, our assumption is that this pseudo-document approach will allow us to accurately measure *relative* performance across different diversity measures.

We manually selected ten pairs of journals from PubMed, where each pair appeared to have unrelated content (see Table 2.2 for a list of journal pairs). Using the process outlined above, for each pair of journals, we generated 50 high-diversity pseudo-documents and for each individual journal in the pair generated an additional 25 low-diversity pseudo-documents. Each parent document was drawn without replacement, meaning that no real document served as a parent of more than one pseudo-document across the entire set. This process yielded a total of 1,000 pseudo-documents, half of which were designed to have high diversity, and half of which were designed to have low diversity.

2.5.2 Experiments

We first tested whether our diversity scores could be used to differentiate the two classes of pseudo-documents.

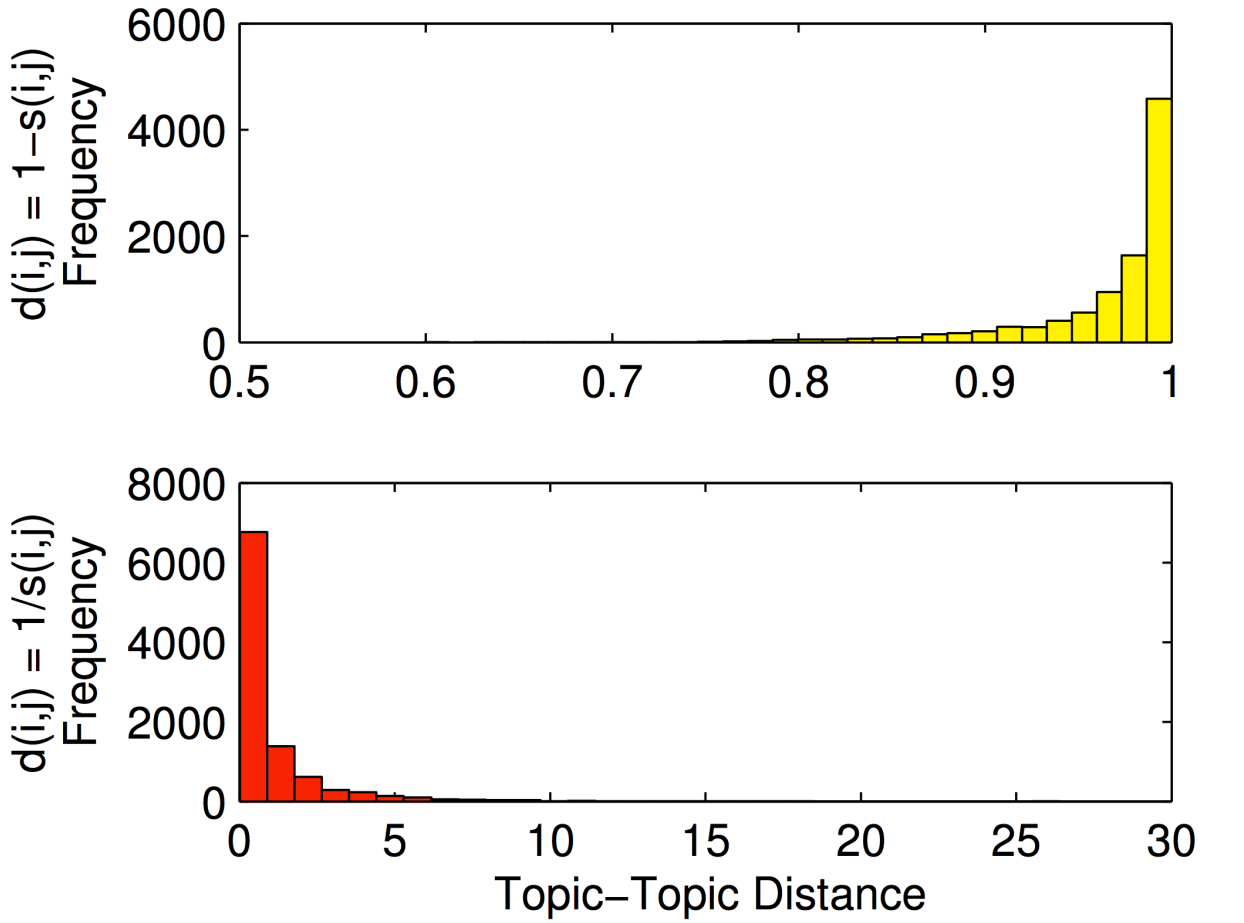

Figure 2.1: Histograms of topic-topic distances for $\delta(i,j) = 1 - s_c(i,j)$ and $\delta(i,j) = 1/s_c(i,j)$.

Abbreviation	Data Matrix	$s(i,j)$	$\delta(i,j)$	10 Topics	30 Topics	100 Topics	300 Topics
DT-PI	*Document-Topic*	Probabilistic	$1/s(i,j)$	0.923	0.911	0.955	0.950
DT-CI	*Document-Topic*	Cosine	$1/s(i,j)$	**0.926**	**0.929**	**0.964**	**0.964**
DT-P	*Document-Topic*	Probabilistic	$1-s(i,j)$	0.799	0.710	0.685	0.608
DT-C	*Document-Topic*	Cosine	$1-s(i,j)$	0.842	0.770	0.772	0.716
WT-PI	*Word-Topic*	Probabilistic	$1/s(i,j)$	0.828	0.722	0.801	0.771
WT-CI	*Word-Topic*	Cosine	$1/s(i,j)$	0.856	0.805	0.814	0.689
WT-P	*Word-Topic*	Probabilistic	$1-s(i,j)$	0.798	0.709	0.685	0.608
WT-C	*Word-Topic*	Cosine	$1-s(i,j)$	0.838	0.779	0.762	0.659
Abbreviation	**Diversity Formula for Document d**			**10 Topics**	**30 Topics**	**100 Topics**	**300 Topics**
Variety	$\sum_{i=1}^{T} 1_{[p(i\|d)>0]}$			0.681	0.667	0.648	0.643
Balance	$\sum_{i,j=1}^{T} p(i\|d)p(j\|d)$			0.797	0.709	0.685	0.608
Entropy	$-\sum_{i=1}^{T} p(i\|d)\log p(i\|d)$			0.812	0.738	0.707	0.646
Disparity	$\sum_{i,j=1}^{T} 1_{[p(i\|d),p(j\|d)>0]}\delta(i,j)$; $\delta(i,j)$ as in DT-CI			0.706	0.706	0.720	0.724

Table 2.1: AUC scores for different diversity measures based on 1000 pseudo-documents from PubMed.

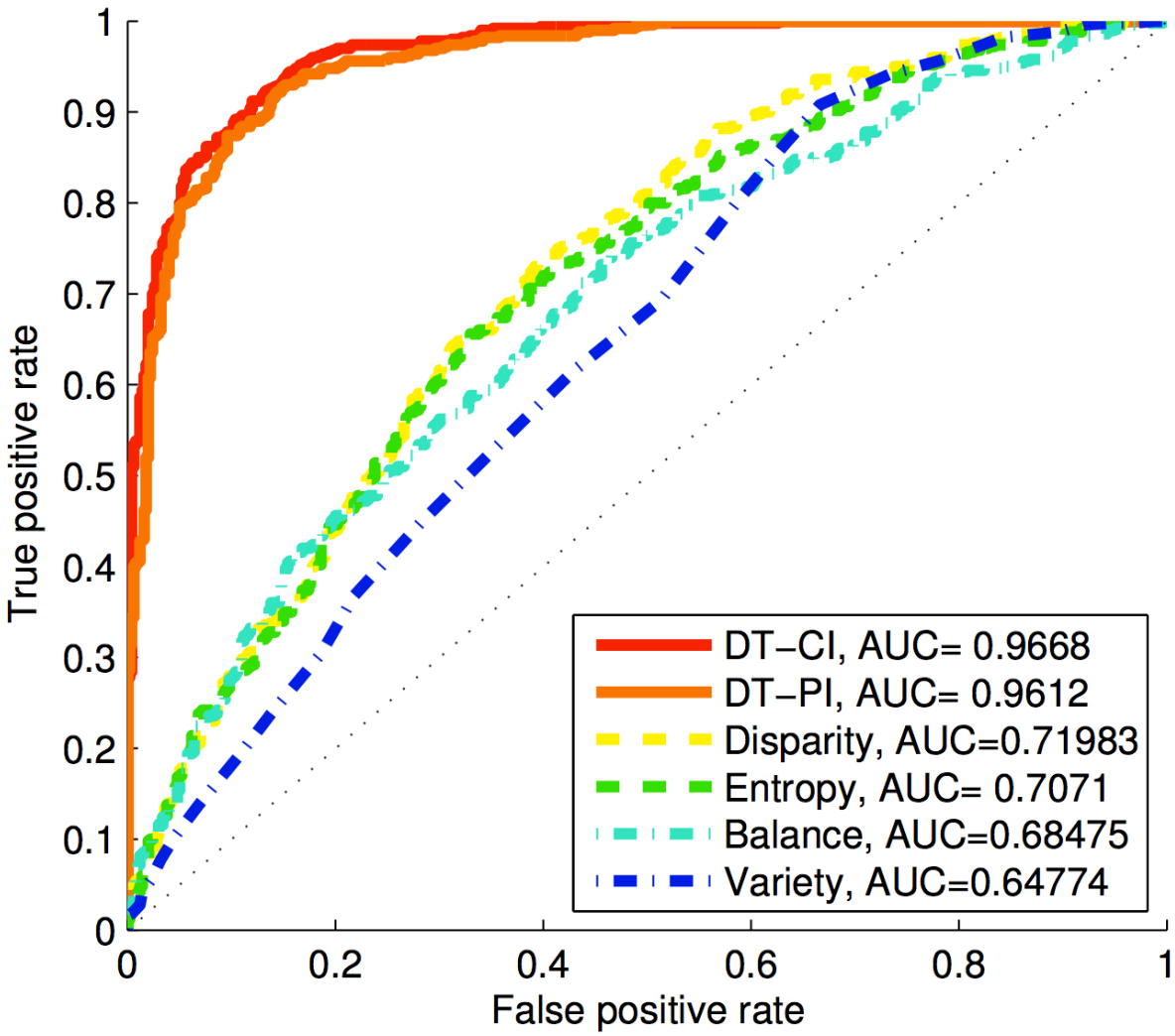

Figure 2.2: Pseudo-document ROC curves for PubMed data with 100 topics comparing Rao diversity to alternate methods. See also Table 2.1.

Journal Name Abbreviations	DT-PI	DT-CI	WT-PI	WT-CI	Variety	Bal	Ent	Disp
All Journal Pairs	0.955	**0.964**	0.801	0.814	0.648	0.685	0.707	0.720
Neuroimage || *BMC Public Health*	0.961	**0.967**	0.894	0.770	0.669	0.654	0.703	0.658
Eplasty || *Plant Mthds*	**0.963**	0.962	0.817	0.810	0.616	0.657	0.660	0.712
Clinical Orthp || *J Nucleic Acids*	**0.972**	**0.972**	0.892	0.854	0.621	0.616	0.642	0.735
J Cell Biol || *J Foot, Ankle Rsrch*	**0.996**	0.993	0.908	0.962	0.631	0.684	0.718	0.805
BMC Med Ethics || *BMC Immnlgy*	0.989	**0.997**	0.822	0.974	0.654	0.758	0.750	0.756
Intl J Emrgy Med || *Intl J Nanomed*	0.955	**0.978**	0.796	0.809	0.690	0.723	0.758	0.743
J Ethnbio, Ethnmed || *J Expl Botny*	0.962	**0.969**	0.781	0.825	0.744	0.666	0.712	0.786
Tbcco Indced Dis || *Neurl Devt*	0.960	**0.966**	0.840	0.888	0.713	0.723	0.735	0.812
Frntrs in Neuro || *Prtcle, Fibr Txclgy*	**0.888**	0.887	0.764	0.610	0.631	0.754	0.778	0.611
Thromb J || *Evlnry Bioinf Online*	0.984	**0.988**	0.849	0.828	0.643	0.758	0.785	0.706

Table 2.2: AUC scores for pseudo-documents from specific journal pairs from PubMed.

We started by learning a set of topic distances on the document-topic count matrix for the 145k PubMed documents. We then used this distance matrix to assign a diversity score to each pseudo-document using the method described in section 2.3. We computed an area under the curve (AUC) value for the ROC curve generated from the set of diversity scores produced by our method based on the designed ground truth 'high' and 'low' diversity values for each pseudo-document.

Table 2.1 lists AUC values for multiple diversity formulas across topic models with 10, 30, 100, and 300 topics. Chance performance will yield AUC values of 0.50, and perfect classification accuracy will yield an AUC of 1.

First, it is clear from these results that different distance measures yield significantly different results. For example, distance measures with $\delta(i, j) = 1/s(i, j)$ perform significantly better than distance measures with $\delta(i, j) = 1 - s(i, j)$ (see Table 2.1).

This is because $s(i, j)$ is close to 0 for most pairs of topics, with large values being on the order of 0.2. As a result, most distances are ≈ 1 when $\delta(i, j) = 1 - s(i, j)$ (see figure 2.1), making this method more akin to a "balance method" than Rao's diversity (as discussed in Section 1). On the other hand, when $\delta(i, j) = 1/s(i, j)$, small similarity values create very large distances, making the distance term appropriately dominant.

A second general observation from Table 2.1 is that distance formulas based on the *document-topic* matrix outperform distance formulas based on the *word-topic* matrix (see Table 2.1). This may indicate that topic co-occurrences in documents are generally more useful in characterizing diversity than are similarities in *topic-word* distributions. As mentioned in section 2.3.2, two topics with very different word distributions may still frequently co-occur within documents in the corpus, which is one possible explanation for why similarity based on *topic-word* distributions performs relatively poorly on this task.

A third observation is that Rao diversity significantly outperforms alternative approaches

```
TITLE: Collaborative Research: Differential Geometry and Statistics of Deformation Tensors
  p_i    [topic name]      top 5 words in each topic
  0.405 [ALGEBRA]          theory algebraic geometry study groups number
  0.207 [GEOSCIENCE]       earth history field time years
  0.180 [STATISTICS]       data statistical methods models analysis
  0.108 [MEETINGS]         conference mathematics researchers students graduate
  0.054 [EARTHQUAKES]      fault earthquake seismic deformation slip
  ...
  Score Term    d(i,j)       x p_i                  x p_j
  0.511 =       6.08         x 0.41 [ALGEBRA]       x 0.21 [GEOSCIENCE]
  0.146 =       6.66         x 0.41 [ALGEBRA]       x 0.05 [EARTHQUAKES]
  ...
  0.834 = Total Diversity Score
```

```
TITLE: Collaborative Research: Development and Application of Proteomics-based Research in
Archaeological Residue Analysis
  p_i    [topic name]      top 5 words in each topic
  0.522 [ARCHAEOLOGY]      archaeological site social region analysis
  0.190 [PROTEINS]         protein molecular structure biological binding
  0.071 [CELLS]            cell membrane proteins molecular development
  ...
  Score Term    d(i,j)       x p_i                  x p_j
  0.234 =       2.35         x 0.52 [ARCHAEOLOGY]   x 0.19 [PROTEINS]
  0.146 =       2.62         x 0.52 [ARCHAEOLOGY]   x 0.07  [CELLS]
  ...
  0.431 = Total Diversity Score
```

Figure 2.3: Two of the most diverse NSF grant proposals.

(see Figure 2.2 and Table 2.1). This supports Stirling's arguments [80] that taking each of *balance*, *variety*, and *distance* is important for measuring diversity, compared to methods such as entropy which don't take all three aspects into account.

Overall, Rao diversity with the distance measures we have termed 'DT-PI' or 'DT-CI' perform the best, where DT refers to a *document-topic* based similarity measure, P to probability-based similarity, C to cosine-based similarity, and I to the inverse transformation of similarity. In addition to yielding high pseudo-document classification accuracies, these methods also appear to be largely invariant to the number of topics in the model (see Table 2.1), and show consistent performance across pseudo-documents drawn from different pairs of journals (Table 2.2). Since the 'DT-PI' and 'DT-CI' methods are very close in performance overall, we use 'DT-CI' as our default measure of diversity from this point forward.

```
TITLE: Gas-Phase Studies of Organic Sigma-type Polyradicals
  p_i   [topic name]     top 5 words in each topic
  0.547 [CHEMISTRY]      chemistry synthesis organic reactions metal
  0.265 [MASS SPECTROMETRY]   mass chemistry nmr instrumentation spectrometer
  0.077 [FLUID DYNAMICS]      flow fluid transport particle heat
  0.043 [MAGNETISM]      magnetic materials spin properties field
  ...
  Score Term    d(i,j)      x p_i                         x p_j
  0.020 =       0.48        x 0.08 [FLUID DYNAMICS]       x 0.55 [CHEMISTRY]
  0.009 =       0.44        x 0.08 [FLUID DYNAMICS]       x 0.27 [MASS SPECTROMETRY]
  ...
  0.047 = Total Diversity Score
```

```
TITLE: Arithmetic Gross-Prasad conjecture for unitary Shimura varieties
   p_i   [topic name]    top 5 words in each topic
   1.000 [ALGEBRA]       theory algebraic geometry groups number
   -----
   0.000 = Total Diversity Score
```

Figure 2.4: Two of the least diverse NSF grant proposals.

2.6 Detecting Diverse Documents

In this section we show examples of the most diverse and least diverse documents detected by our algorithm for each of our three corpora: PubMed Open Access, NSF Grant Awards, and the ACL Anthology. For each corpus we built a topic model with 100 topics, and computed diversity scores using Rao diversity with the DT-CI distance measure as defined in Table 2.1. We scaled the distances $\delta(i, j)$ to have a mean value of 1 within each corpus, putting the distances and diversity scores on roughly the same scale across corpora. We also manually assigned names to topics to aid in interpreting the results.

Figure 2.3 shows two of the most diverse NSF awards (from a corpus of approximately 52k abstracts of awards) detected by the algorithm. The first award is a collaborative research project between mathematicians and geoscientists. As shown in Figure 2.3, the releatively large distances (6 times larger than the mean pairwise topic distance) between ALGEBRA and each of the GEOSCIENCE and EARTHQUAKE topics drive a significant portion of the total score. The distances between these topics is reflected in the description of the project in the abstract:

> This vast mathematical theory has been applied to geology in only a few instances. This project represents collaboration between two structural geologists and a mathematician.... [It] opens the door to further cross-fertilization among geology, mathematics, and other fields.

The second of the two awards in Figure 2.3 is considered diverse because of the combination of the topic ARCHAEOLOGY and the two biology-related topics PROTEINS and CELLS. Again, the relatively large distances (2.4 and 2.6) between these topics and their relative strength within the document yield a particularly high diversity score for this document.

The two examples of low-diversity documents in Figure 2.4 tell a different story. The first grant is somewhat narrowly focused, dominated by topics that are relatively close such as CHEMISTRY, MASS SPECTROMETRY, and FLUID DYNAMICS. The second grant is an example of a document that gets a topical diversity score of 0 because all of its words are assigned to the single topic of ALGEBRA.

Figure 2.5 shows two the most diverse articles from the PubMed corpus. The diversity score for the first article is dominated by the combination of the PSYCHIATRY and FUNGI topics, which have a distance of 16.91 times the mean topic distance. The diversity score of the second document is largely driven by the fact that the BONES/JOINTS topic is relatively distant from each of the HIV/AIDS and VIRUSES topics. Low diversity PubMed documents showed similar patterns to low diversity NSF grants.

Finally, Figure 2.6 shows examples of one *high* diversity document and one *low* diversity document from the ACL corpus. The *high* diversity document achieves its score because the SUMMARIZATION topic is usually associated with text, but here it co-occurs with a set of topics related to SPEECH RECOGNITION. Thus, this paper is unusual in that it applies summarization techniques to non-text data (as indicated in the title). The other paper in Figure 2.6 is a typical example of a low-diversity document which is composed of a

```
TITLE: Neuropsychiatric manifestation of confusional psychosis due to Cryptococcus
neoformans var. grubii in an apparently immunocompetent host: a case report
   p_i   [topic name]    top 5 words in each topic
  0.314  [CLINICAL MEDICINE]  patient case diagnosis lesions examination
  0.195  [PSYCHIATRY]    depression patients disorder symptoms mental
  0.131  [FUNGI]         fungal species albicans amp cbs
  0.120  [INFECTIOUS DISEASE] isolates infection tuberculosis strains resistance
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.432 =       16.91       x 0.20 [PSYCHIATRY]            x 0.13 [FUNGI]
  0.045 =       0.44        x 0.08 [PSYCHIATRY]            x 0.27 [INFECTIOUS DISEASE]
  ...
  0.598 = Total Diversity Score
```

```
TITLE: Operations about Hip in Human Immunodeficiency Virus-Positive Patients
   p_i   [topic name]    top 5 words in each topic
  0.264 [BONES/JOINTS]   bone patients joint knee fracture
  0.234 [HIV/AIDS]       hiv aids sexual infection drug
  0.189 [SURGERY]        surgery patients procedure postoperative patient
  0.043 [VIRUSES]        virus infection replication hiv influenza
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.404 =       6.53        x 0.26 [BONES/JOINTS]          x 0.23 [HIV/AIDS]
  0.047 =       4.11        x 0.26 [BONES/JOINTS]          x 0.04 [VIRUSES]
  ...
  0.541 = Total Diversity Score
```

Figure 2.5: Two of the most diverse PubMed OA articles.

```
TITLE: Summarizing Speech Without Text Using Hidden Markov Models
  p_i   [topic name]     top 5 words in each topic
  0.248 [SUMMARIZATION]  summary document rouge sentences content
  0.132 [SPEECH RECOGNITION]  speech recognition speaker training models
  0.089 [FINITE STATE MACHINES] state finite transducer transition automaton
  0.078 [EVALUATION]     results set precision performance score
  0.073 [PROSODY]        prosodic pitch speech phrase cue 0.417
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.136 =       7.55        x 0.25 [SUMMARIZATION]         x 0.07 [PROSODY]
  0.135 =       4.13        x 0.25 [SUMMARIZATION]         x 0.13 [SPEECH RECOGNITION]
  0.044 =       1.99        x 0.25 [SUMMARIZATION]         x 0.09 [FINITE STATE MACHINES]
  ...
  0.431 = Total Diversity Score
```

```
TITLE: Less is More: Significance-Based N-gram Selection for Smaller, Better Language Models
   p_i   [topic name]    top 5 words in each topic
   0.507 [LANGUAGE MODELS]    model language data training gram
   0.254 [PROBABILITY]   probability distribution number estimate entropy
   0.077 [ALGORITHMS]    algorithm time search number size
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.005 =       0.04        x 0.51 [LANGUAGE MODELS]       x 0.25 [PROBABILITY]
  0.003 =       0.07        x 0.51 [LANGUAGE MODELS]       x 0.08 [ALGORITHMS]
  ...
  0.021 = Total Diversity Score
```

Figure 2.6: High diversity (top) and low diversity (bottom) ACL articles.

combination of topics that are very close together.

2.7 Conclusions

We presented an approach for quantifying the diversity of individual documents in a corpus based on their text content. Empirical results illustrated the effectiveness of the method on multiple large corpora. This text-based approach for assigning diversity scores has several potential advantages over previous alternatives, such as methods that define diversity based on citations categorized into predefined journal subject categories. The text-based approach is more data-driven, performing the equivalent of learning journal categories by learning topics from text, and can be run on any collection of text documents, even without a prior categorization scheme. In addition, it produces human-readable explanations and can be easily generalized to score the diversity of other entities such as authors, departments, or journals (e.g., by aggregating counts across such entities).

A possible direction for future work is that of temporal document diversity, for example, using topics and topic-based distance measures that only depend on documents in the corpus with earlier time stamps. This would allow for distances and diversities that change over time and the detection of documents that are highly diverse relative to the time-period they were published in. An example would be early papers in bioinformatics, combining machine learning and biological concepts, which co-occur relatively frequently in the current literature but far less so 20 years ago.

Chapter 3

The Cosmic Prevalence of Quenching in Low Mass Satellites

3.1 Introduction

Local Group (LG) observations serve as a Cosmic Rosetta Stone, which can greatly aid our understanding of the distant universe. Observations of resolved stars in local group dwarf galaxies grant insight into their histories [36, 49, 87, 52, 38, 93, 94], providing a lens to the early Universe which witnessed their formation. [14, 31, 70, 7, 13].

Studies of the satellite systems of Milky Way analogs have revealed much about how the evolutions of its dwarf galaxies are affected by their environments [92, 19, 40, 96, 97, 60]. These studies demonstrate that massive (i.e. $\sim 10^9\,\mathrm{M}_\odot$) satellites become environmentally quenched inefficiently over long (several Gyr) timescales, although the mechanism by which the cessation of star formation occurs remains an area of active study. The high-mass dwarfs that make up the more distant universe stand in stark contrast to the LG's population of satellites, where dwarfs below $\approx 10^8\,\mathrm{M}_\odot$ in solar mass are nearly all passive, suggesting rapid

quenching timescales [29].

Further studies are needed to bridge the gap between Local Group and studies of the z=0 populations of dwarfs at large. Comparing these two populations necessitates probing lower-mass objects than is currently possible with large scale spectroscopic surveys, which can not probe satellites below $10^{8.5}\,\mathrm{M}_{\odot}$ in stellar mass [61]. Photometric studies, on the other hand, can provide leverage to connect low-mass LG systems to the field, particularly in their mass and phase-space distributions [58, 81]. These surveys are less useful in replicating the detailed star formation information obtained in spectroscopic studies. This is significant, as populations of galaxies at $10^{7}\,\mathrm{M}_{\star} \sim 10^{8}\,\mathrm{M}_{\star}$ are predicted to have quenched fractions that are highly dependent on their local environment [28, e.g.]. However, this differing behavior of high-mass and low-mass dwarfs is grounded only on local observations.

Confirming that this feature of LG satellites is a general property of satellite galaxy evolution is a crucial step in understanding how the LG fits into the larger picture of galaxy evolution, and determining if the formation and evolution of the LG is "typical" of similar systems. The significance of the representativeness, or conversely the uniqueness, of the Local Group is magnified by the observation that at high redshift (z=7), the Local Group progenitor spanned a linear distance of 7 comoving Mpc, corresponding to a volume of $350\,\mathrm{Mpc}^3$ [8]. This makes it a valuable cosmological tool, as Local Group archeology probes a region of similar size to the Hubble Ultra Deep Field, while also comprising objects significantly fainter than would be feasible to detect at high redshift. However, the utility of the LG as a cosmological tool depends on how representative the LG's evolutionary history is to similar systems.

In this work, we present a novel methodology for extending spectroscopic studies through machine learning to lower masses than they are capable of probing directly. Our technique enables us to reach masses on the order of those of LG satellites in service of studying the dependence between the evolution of low-mass dwarfs and their environment, and determining how representative the LG is of similar systems across cosmic time. The paper is

organized as follows: in Section 3.2, we describe the observational data we use, in Section 3.3, we describe our procedure for identifying and classifying satellites, in Section 3.5, we give our results and discuss the implications of our findings.

3.2 Observations

The observational data used in this study is compiled from several different surveys:

The *training set* was drawn from the Sloan Digital Sky Survey (SDSS), using data from value-added galactic catalogs of [5] and [10]. These data were used to assign each point of four dimensional color-color space a value that maps to the probability that a galaxy residing at that point in color space is quenched.

From the SDSS, we select for the training sample galaxies of stellar mass $7 < \log \mathrm{M}_\star$. Galaxies are labeled star-forming or quenched based on where they lie in SFR-stellar mass space. If they lie above the line

$$\mathrm{SFR} = -0.7 \times \log \mathrm{M}_\star - 7.7 \tag{3.1}$$

they are considered star forming, otherwise they are taken to be quenched. Star formation rates are taken from the catalog, and are derived from emission lines and the D4000 spectral index. For the purposes of this study, it is critical that the star formation rates used in this sample are derived from galactic spectra, as we will be seeking to link spectroscopic properties to ones derived from photometric observations of faint objects.

Additional training set data was drawn from the Galaxy and Mass Assembly (GAMA) survey, a spectroscopic survey of $\sim 300,000$ galaxies covering $\sim 286\mathrm{deg}^2$ of equatorial sky. Spectra were taken using an AAOmega multi-object spectrograph on the Anglo-Australian telescope.

The survey is complete down to $r < 19.8$, making it 2 magnitudes deeper than SDSS in spectroscopy. Since we are primarily interested in faint objects, and our algorithm weighs low-mass objects more heavily than high-mass objects, this added depth is beneficial.

The *test set* is drawn from the Canada France Hawaii Telescope Legacy Survey (CFHTLS), a photometric survey comprised of a "wide" component and a "deep" component. For the purposes of this study, we consider only three of our of the four "wide" fields, corresponding to those with spectroscopic overlap, fields W1 (72 square degrees), W3 (49 square degrees), and W4 (25 square degrees). The chosen fields overlap with the NASA-Sloan Atlas, which combines Galaxy Evolution Explorer (GALEX) photometry with SLOAN data; this overlap allows us to select photometric satellites of spectroscopically confirmed hosts.

In selecting our test set, we apply a number of quality cuts to CFHTLS objects, most importantly requiring they be flagged as galaxies by SEXTRACTOR. At this point, we apply no cut on distance to nearby object; below we will discuss how our test set is divided into photometric satellites and background objects. The photometric satellites form our primary scientific sample; for this reason it is important to understand our completeness, both in an absolute sense and the extent to which completeness depends on galaxy properties, particularly color.

To evaluate how complete our test set is, we first examine how complete our training set is. Figure 3.2 illustrates this; the red and blue points represent the passive and star-forming galaxies in our training sample, respectively. The black dashed line represents the region of parameter space where the training set is complete. As a fiducial measure, we transpose the completeness curve to fainter magnitudes to bring it into agreement with the Next Generation Virgo Cluster Survey (NGVS, see [27]), a study that used the MegaCam instrument to observe Virgo cluster objects (i.e. objects with known redshift). We then adjusted our completeness limit to account for the difference in depth between the NGVS and the CFHTLS, and adjusted again to account for the differential completeness between MegaCam bands.

This ensures that a galaxy observed in, e.g., the r band would not be unobserved in the i band.

These scalings give us a final completeness magnitude of $r \sim 19.7$, two magnitudes fainter than SDSS. Using our assumed relationship between magnitude and mass, we can compute the redshifts our study is complete to as a function of mass; we are able to observe galaxies of mass $10^9\,\mathrm{M}_\star$ out to $\mathrm{z} = 0.05$, of which there are 75 hosts meeting our criteria in our survey area; galaxies of mass $10^8\,\mathrm{M}_\star$ out to $\mathrm{z} = 0.027$, giving us 26 hosts; and galaxies of mass $10^7\,\mathrm{M}_\star$ out to $\mathrm{z} = 0.013$, giving us just 4 hosts. We would be able to track galaxies down to satellites of mass $10^6\,\mathrm{M}_\star$ to $\mathrm{z} = 0.006$, however there are no such hosts in our sample. These critical redshifts are noted in Figure 3.2.

3.3 Satellite Characterization

The procedure of assigning dwarf galaxies to satellite systems can be broken down into two parts: identifying the families of satellite galaxies, and characterizing their star formation. In this section, we discuss each of these in turn.

3.3.1 Satellite Finding

The first step in obtaining satellite galaxy samples is host selection. As our goal is to select hosts similar to the Milky Way, we select hosts from the NASA-Sloan Atlas between $10^{10.4}\,\mathrm{M}_\star$ and $10^{11}\,\mathrm{M}_\star$ in stellar mass. Hosts are retained only if they fall on our three CFHTLS fields. A redshift cut is then applied to the hosts so that only hosts closer than z=.04 end up in the final sample (see previous section). We then select all objects within 150 projected kpc to be photometric satellites. To account for edge effects, we draw a circle of radius 150 kpc around each galaxy and calculate the ratio of the area of a circle that lies within the field to the

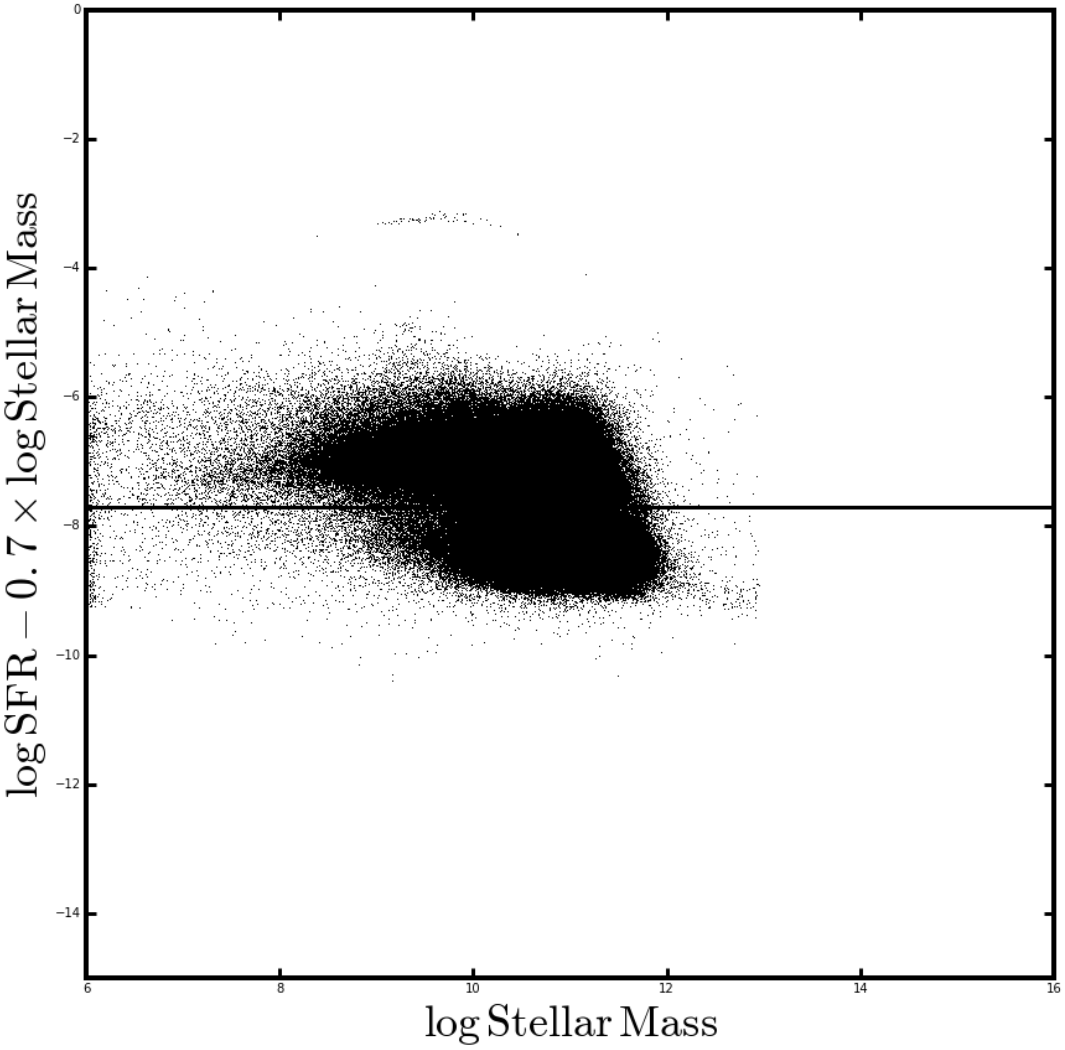

Figure 3.1: Modified specific star formation rate (log SFR - .7 log Stellar Mass) plotted against stellar mass for objects in the SDSS training sample. The horizontal line is the dividing line between star forming and quenched galaxies.

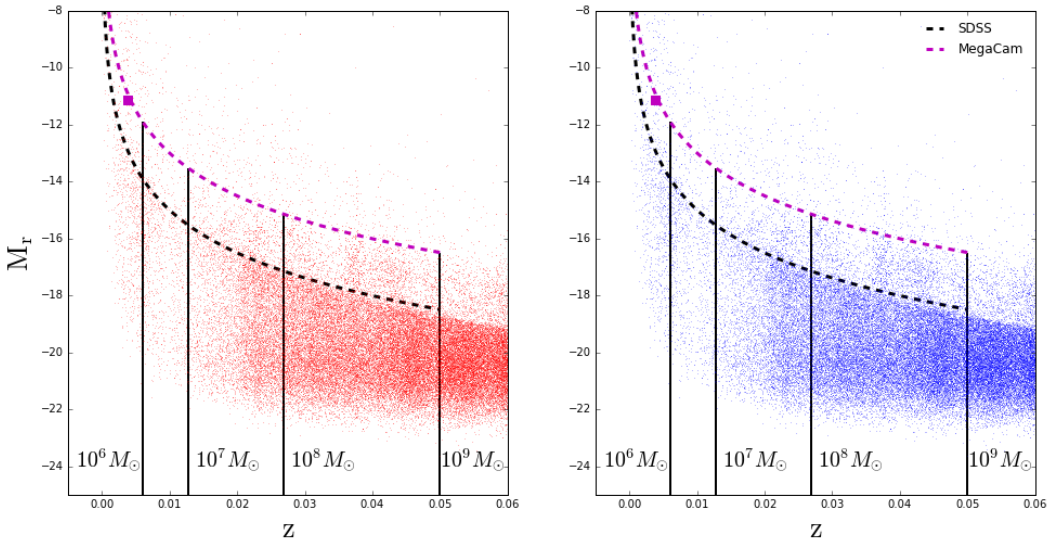

Figure 3.2: Absolute r band magnitude plotted against redshift for galaxies in the training set. The red points in the left panel represent passive galaxies, the blue points in the right panel represent star-forming galaxies. Completeness curves are shown for both the training set (black curve) and the test set (magenta curve). The magenta square shows the test set completeness derived from [27] (see text for discussion). Vertical black lines indicate the redshifts at which the test set is complete to the labeled mass.

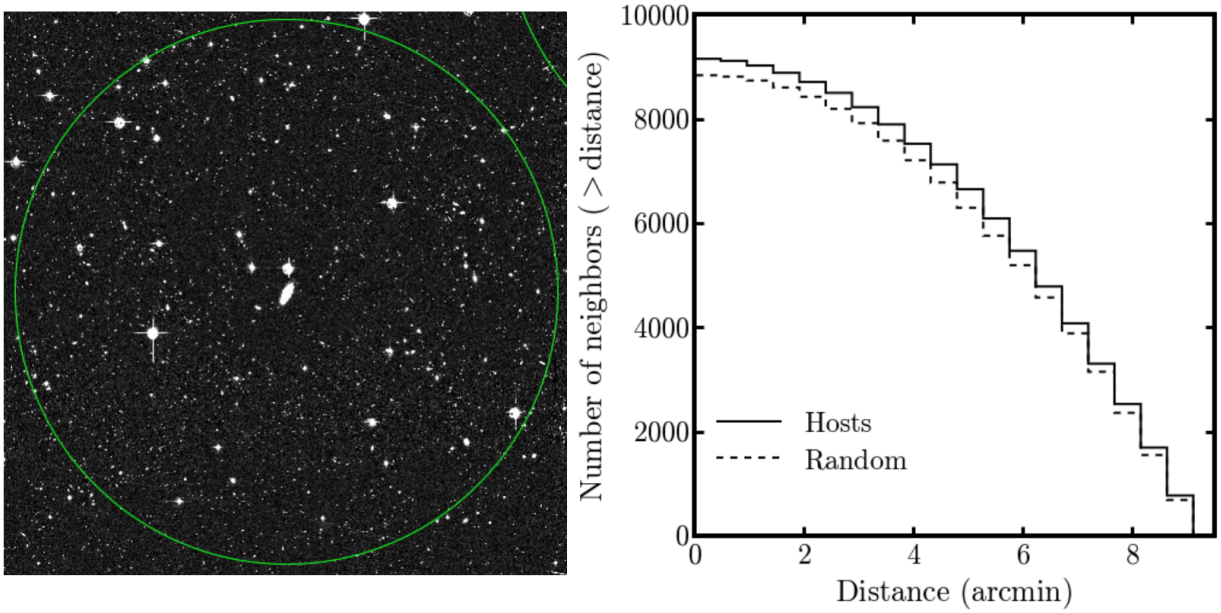

Figure 3.3: Counts of galaxies $16.5 < r < 18.5$ in a circular aperture of radius 9 arcmin around hosts (as shown on left), plotted versus distance from the host (solid line). Also shown are counts around random points on the sky (dashed line). This figure serves to demonstrate the validity of the technique of measuring satellite statistics by statistically subtracting the background. Note that at large distances from the hosts, the counts come into agreement with the random background.

full area of the circle, calling this ratio F_{area}. Hosts with F_{area} lower than 0.75 are discarded; for the remainder, $1/F_{\text{area}}$ serves a correction factor for the missing area lying outside of the field.

Once satellites are identified, we assign a mass to them by assuming that the mass of each satellite is a power law function of its apparent r band magnitude, where the log-space intercept of the function is set by redshift. Our investigation confirms that, in the training set, mass is indeed well-described by a power law in r band magnitude. By interpolating the redshift-intercept relationship, we can define a power law relation between stellar mass and r magnitude for satellites of any redshift, where we assume the redshift of each satellite is approximately equal to that of its host.

The fundamental challenge of assigning satellite group membership to photometric objects is that it is not known where they lie in velocity space, and thus whether they may be associated with a given host. There are two ways of addressing this issue: making use of photometric redshifts derived from SED fitting, or by systematically subtracting out an assumed background population based on observations of "blank" fields. It is this second method that we make use of. For each of the three CFHTLS fields we use in this study, we take use 500 random "pencil beam" pointings within the field to compute the on-sky background density. We use the 40th percentile two-dimensional density to account for the skewedness in the distribution of densities due to random pointings in the direction of low-z, i.e. foreground, overdensities, as the background density for each. Figure 3.3 serves as a proof of concept for counting satellites in this way, the right panel shows an excess in counts of objects as a function of on-sky distance around our hosts as compared to around random locations. This excess is limited to ~ 9 arcmin, corresponding to a phyiscal distance of $100 - 300$ kpc at the redshifts we are concerned with.

3.3.2 Star Formation Modeling

Our modeling of star formation is motivated by a desire to classify photometric satellites into star formation categories in a data-driven way based on their position in color-color parameter space. We take a machine-learning approach to the problem using two classes of models, one built around random forests and one which uses a neural network. Both of these are calibrated on the training set, then applied to the test set. We first discuss the shared data preparation procedure for these two model classes followed by more detailed descriptions of the each model type.

We combined photometric data from 54,288 galaxies gathered in the SDSS with data from 70,850 galaxies gathered in the GAMA for a total of 125,138 data points. Of these, we removed 5,191 datapoints which contained unphysical color values, leaving a total of 119,947 datapoints for the main analysis. 30.4% of these datapoints represent quenched galaxies, the remaining 69.6% represent star forming galaxies. These datapoints were randomly shuffled and split into three sub-datasets: a training set containing 95,959 datapoints (80% of the total), a first validation set to use for tuning hyperparameters containing 11,994 datapoints (10%), and a second validation dataset for estimating out-of-sample errors containing the remaining 11,994 datapoints (10%). The random shuffling and splitting was done in such a way that the portion of quenched galaxies in each dataset remained approximately constant.

Hyperparameter Optimization

The success of machine learning algorithms often depends critically on the values of a number of model- and dataset-specific hyperparameters [76]. While the express purpose of training a machine learning model is to learn the *parameter* values which minimize the model's error on a given training dataset, the values of its hyperparameters are taken to be defined a-priori and are not modified by the model's training process. Therefore, in order to train the best

final model possible, it's critical to optimize these hyperparameters for each problem and dataset under consideration. For both our random forest and neural network models, we train a number of distinct versions of the model, each using different hyperparameter values. The general algorithm that we use to optimize the hyperparameters of both of these models is Outer-Loop Hyperparamter Optimization (algorithm 1).

Algorithm 1 Outer-Loop Hyperparameter Optimization

```
HPOpt ← Chosen Hyperparameter Optimizer
for Iteration i ← 1...K do
    Hyperparams α_i ← HPOpt.NextHyperParamSet
    Train model on training data with hyperparams α_i
    Make predictions for validation data with model
    HPOpt.RecordValidationError
end for
Best model is model with lowest validation error
```

It is worth highlighting that this method requires a full training and testing cycle within each iteration of the hyperparameter selection loop, making it considerably more computationally expensive than training a single model in isolation. The benefit is that the predictive performance of the best models trained using this method will often be considerably better than models trained using default hyperparameters only.

Random Forest Models

We first trained a series of random forest models to try to predict whether a galaxy was quenched or not from its four-band color information. A random forest is comprised of a set of small decision trees [9], each trained on a subset of the training datapoints and features. At prediction time, each tree in the forest makes an independent prediction, and the predictions for each datapoint are averaged across trees to obtain the prediction for the forest as a whole.

We used the `RandomForestClassifier` implemented in Python's `scikit-learn` package to train our random forest models. We also used `scikit-learn`'s `RandomizedSearchCV` class to

perform our hyperparameter search by selecting hyperparameter sets uniformly at random from the range of legal values that we defined. These included: `max_depth`, the maximum legal depth for each tree in the random forest (range: 1 to 7 or `None` for no max depth), `max_features`, the maximum number of features which may be used to train each tree (range: 1 to 4), `min_samples_split`, minimum number of datapoints which a node in the tree must contain in order to be considered for splitting (range: 1 to 31), `min_samples_leaf`, minimum number of datapoints required before a node in a tree can be considered a leaf node (range: 1 to 31), and finally `criterion`, whether to split nodes using information gain or the gini coefficient.

Since each tree in a random forest is only trained on a subset of the training data, the remaining data in the training set can be used to get an accurate out-of-sample error estimate, rendering our first validation dataset unnecessary. As such, we folded the first validation set into the training set, to yield an augmented training set size of 107,953 datapoints.

We trained a total of 500 random forest model, the best of which, as judged by area under a receiver operating characteristic curve (AUC) on out-of-sample data[1], achieved a final out-of-sample AUC on the second validation dataset of 0.898 and an accuracy of 84.0%.

Neural Network Models

We next trained a series of feed forward neural network models to predict whether a galaxy was quenched or not from its four-band color data. Each neural network had its hyperparamters chosen by `Spearmint`, a Bayesian optimization framework [76] for hyperparameter selection. Unlike the `RandomizedSearchCV` procedure that we used to optimize the hyperparameters for the Random Forest models, `Spearmint` records the final validation set

[1]AUC is a measure of classification accuracy which is robust to unbalanced class proportions. A model which makes random choices will achieve an AUC value of 0.5, while a model which achieves perfect classification accuracy will receive an AUC value of 1.0.

performance of each hyperparameter set and uses them to bias future hyperparameter draws toward areas of hyperparameter space for which it expects that model to perform will.

The hyperparameter ranges that we considered were: the number of hidden layers in the feed-forward neural network (range: 2 to 10), number of hidden units per layer (range: 4 to 400), the dropout probability for each layer [79] (range: 0.0 to 0.8), the batch size for each stochastic update (range: 16 to 512), the base-10 log of the stochastic gradient learning rate (range: -5.0 to 0.0), the per-epoch learning rate decay factor (range: 0.90 to 1.0), the momentum coefficient for the stochastic gradient update rule [83] (range: 0.0 - 1.0), whether to use Nesterov [57] or classical momentum updates, and whether to replace the stochastic momentum update rule the with AdaDelta update rule [98].

Once the hyperparameter values for a given neural network model had been chosen, we trained the model for 20 epochs on the training data set[2]. After each epoch, we recorded the partially trained model's AUC score on the first validation dataset. We took the best validation set AUC after any training epoch as the metric of interest for hyperparamter optimization, reporting this value back to `Spearmint` at the end of the training phase for each neural network. All the neural network models were built in Keras [16] with a Theano [86] backend and trained on NVIDIA GPUs.

In total, we trained 500 distinct neural network models. The best of the 500 models—as judged by AUC on the first validation dataset—was a model with 5 hidden layers, 388 neurons in each layer, dropout probability of 0.1232 at each hidden layer, batch size of 86, and AdaDelta updates with a learning rate multiplier of 1.0. The model was trained on the training data until reaching a maximum AUC on the first validation dataset after 6 epochs. This trained model achieved an AUC value of 0.900 and an accuracy of 84.0% on the hitherto-unseen data from the second validation set.

[2]Here, an "epoch" means a full pass through the training dataset

Table 3.1: Number of datapoints in resampled versions of the second validation set.

Log Mass Range	Original	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 1.5$
$[6, 8)$	287	1,677	4,022	5,715
$[8, 10)$	6,444	3,522	4,058	3,721
$[10, 12)$	5,263	6,795	3,914	2,558

Table 3.2: AUC and accuracy measures for neural net and random forest models on several versions of the second validation set.

Dataset	NN AUC	RF AUC	NN Acc	RF Acc
Original	0.900	0.898	0.840	0.840
$\gamma = 0.5$	0.872	0.871	0.804	0.806
$\gamma = 1.0$	0.885	0.885	0.830	0.836
$\gamma = 1.5$	0.887	0.884	0.840	0.844

3.3.3 Mass Distribution Sensitivity Analysis

Since the mass distribution for the objects in the training set may differ significantly from the mass distribution of the true objects of interest, we next conducted a sensitivity analysis to see how well these models would perform on data drawn from a different underlying mass distribution. For this purpose, we partitioned the second validation set into objects with log stellar masses in the ranges [6, 8), [8, 10), and [10, 12). We then sampled a series of datapoints (with replacement) from each of these bins in the second validation set to create new versions of the second validation set with different mass distributions. If the number of datapoints sampled from the bins [6, 8), [8, 10), and [10, 12) are labeled n_6, n_8, and n_{10} respectively, we sampled datapoints such that $n_6/n_8 = n_8/n_{10} \equiv \gamma$ for γ values of 0.5, 1.0, 1.5 (see Table 3.1).

Finally, we used the models which had been trained on the un-resampled training data to make predictions for each of the resampled test sets (see Table 3.2). The test set AUC is slightly lower for the resampled datasets, but remains above 0.87 even for datasets with grossly different mass distribution than the training data.

3.3.4 Algorithm Efficacy

In evaluating the algorithm, the two values we are concern with are the "precision" (often called purity) and "recall" (often called completeness). In the case of our algorithm, precision is the probability of a galaxy identified as quenched by the algorithm is truly quenched. Recall is the probability that a truly quenched galaxy is identified as quenched. Since the SDSS gives us spectroscopic information on the galaxies in our training set, we may compute both precision and recall directly. Furthermore, in the case where objects have binary labels (i.e. "quenched" and "star-forming"), it can be shown that

$$F_q = f_q \times (P/R) \tag{3.2}$$

where F_q is the true quenched fraction, f_q is the quenched fraction determined by the algorithm, P is the precision, and R is the recall. There are two things to note from this equation: Regardless of the values of P and R, if $P = R$ then the measured quenched fraction is the same as the true quenched fraction. More importantly, regardless of the values of P and R, if those values are known, F_q may be determined from f_q using the value of P/R, which we will refer to as the "correction factor," or C. Note that this can be easily generalized to a scenario where F_q, f_q, P, and R and all functions of mass, enabling us to define a mass-dependent correction factor $C(\mathrm{M}_\star)$.

In Figure 3.4 we show the correction factor derived from our two algorithms plotted against stellar mass. The red squares represent points derived from the DT algorithm, while the blue circles represent the same for the NN algorithm. For each algorithm, we assign a fit to serve as our modeled $C(\mathrm{M}_\star)$. While both the NN and DT models are well-described by a linear fit, we emphasize that, in principal, $C(\mathrm{M}_\star)$ could take any form and the true quenched fraction would be recoverable from the data. We tuned our algorithm such that $C(\mathrm{M}_\star)$ was near unity over the dynamic range we consider. Choosing $C(\mathrm{M}_\star)$ in this way minimizes both

the correction itself and the extent to which systematic errors will be amplified in correcting the data.

3.4 Precision and Recall

In §3.3, we claim that the true fraction of quenched galaxies is given by the measured fraction of quenched galaxies multiplied by a correction factor which is the ratio of the algorithm's precision to its recall. This claim is proven here. Let N be the number of galaxies in a set. Let q be the number of galaxies in that set that are truly quenched, while Q is the number of galaxies that are measured to be quenched. Let Qq be the number of galaxies that are both truly quenched and measured to be quenched. We then have the precision, given by

$$P = \frac{Qq}{Q} \tag{3.3}$$

and the recall, given by

$$R = \frac{Qq}{q} \tag{3.4}$$

Thus

$$\frac{P}{R} = \frac{q}{Q} \tag{3.5}$$

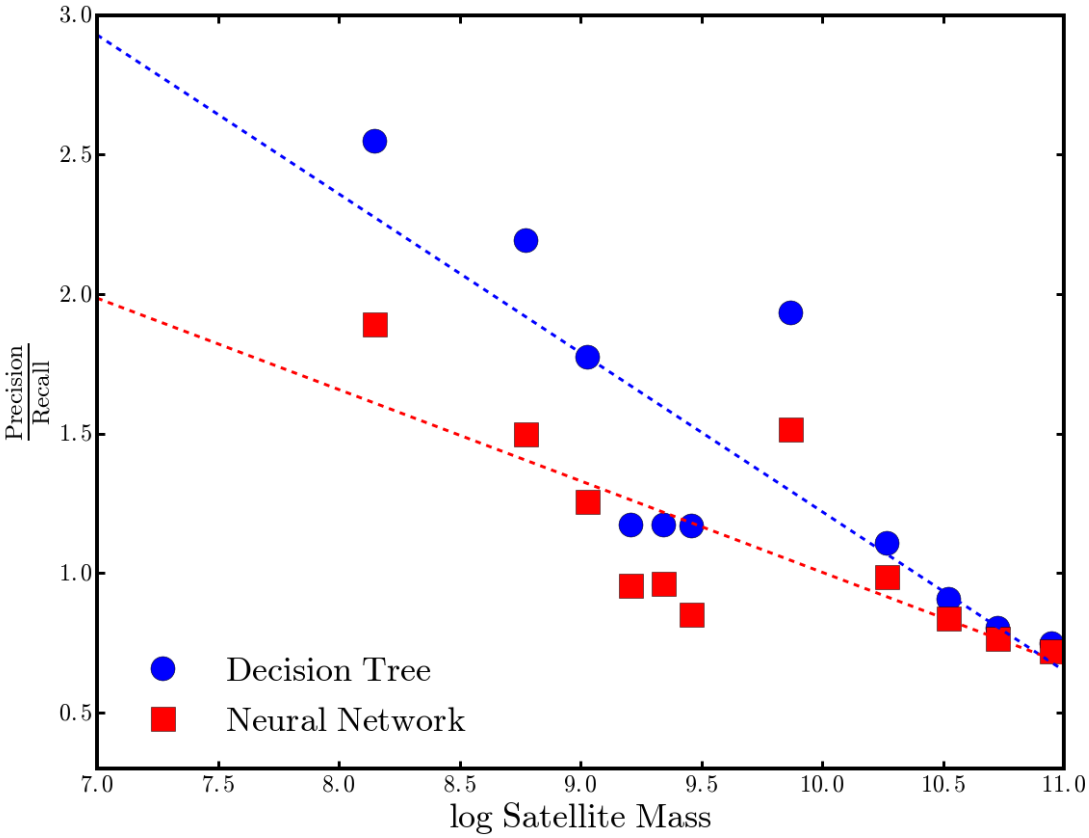

Figure 3.4: The correction factor, defined as the ratio of precision to recall, plotted against stellar mass for each of our two algorithms. The red squares and blue circles represent the correction factor derived from narrowly binning the data in the DT and NN algorithms, respectively. From these data, we fit a linear correction function for each algorithm (dashed lines). Note that the algorithms were tuned such that correction factors near unity were produced.

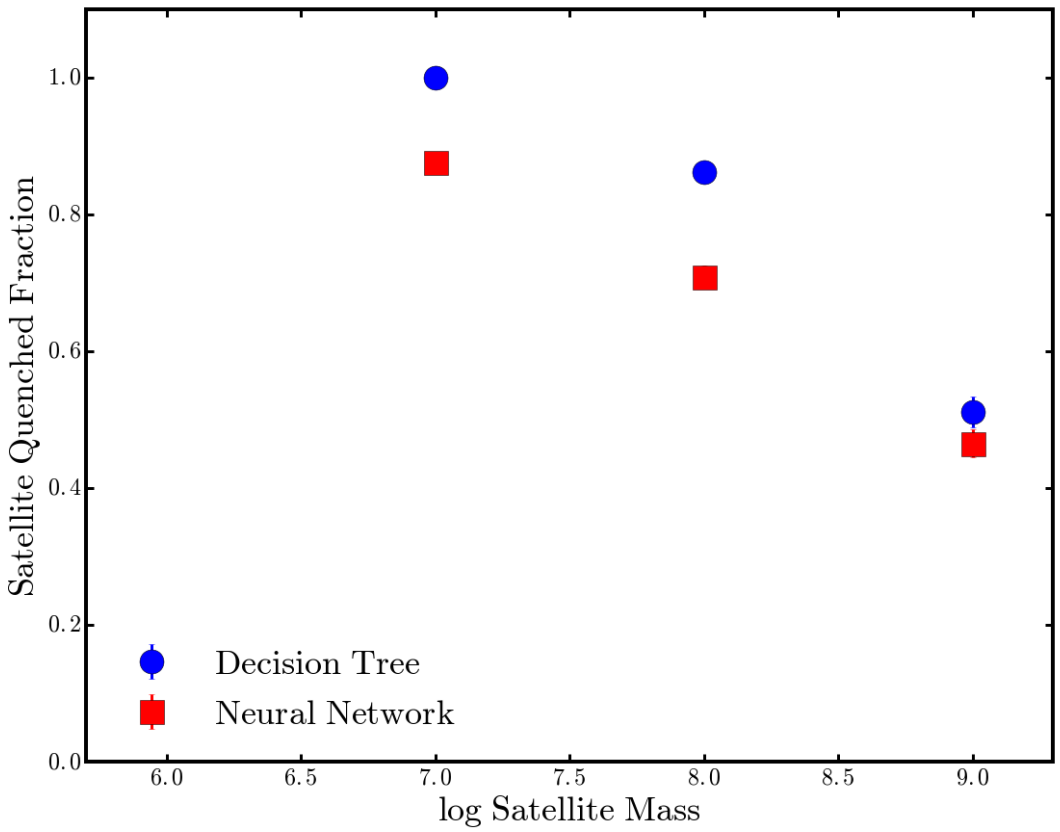

Figure 3.5: The quenched fraction of satellite galaxies as a function of satellite mass, derived from our two algorithms. Red squares denote quenched fractions measured using the *decision tree* algorithm, while blue circles denote those using the *neural network* algorithm (see §3.3). Our results indicate an elevated quenched fraction at low masses, in broad agreement between the two algorithms.

The true fraction of galaxies that are quenched is

$$f_{quen} = \frac{q}{N} = \frac{Q}{N} \times \frac{q}{Q} = f_{quen,measured} \times \frac{P}{R} \tag{3.6}$$

3.5 Results and Discussion

In Figure 3.5 we show the fraction of satellite galaxies that are quenched, as determined by both of our algorithms as a function of satellite mass. The NN algorithm is shown as red squares, while the DT algorithm is shown as blue circles. Both algorithms show a trend of decreasing quenched fraction with increasing stellar mass. The errors reported in Figure 3.5 are binomial errors on the satellite quenched fraction; we can estimate the errors associated with the algorithm itself to be $\sigma =\sim 0.1$, which would bring the data from both algorithms into relative agreement. With these errors in place, the trend is still seen.

In comparing Figures 3.4 and 3.5, there is a possible point of concern: the trend noted above seems to be driven by the functional form of $\mathrm{C_m}$. If our methodology failed entirely, the algorithms would, in essence, randomly assign "quenched" or "star-forming" labels to galaxies, and the dependence of quenched fraction on stellar mass would be entirely driven by C_m. While at first this may appear to be the case, our high-mass data, is in reasonable agreement with previous results; this serves as confirmation that our methodology "matches up" with known results at high mass. Furthermore, the functional form of the algorithm applied to the evaluation portion of the training set irrespective of environment shows a strong dependence on stellar mass, with low mass galaxies being quenched at a rate of about 20%, highly different than our test set result, suggesting we are indeed capturing an environmental effect.

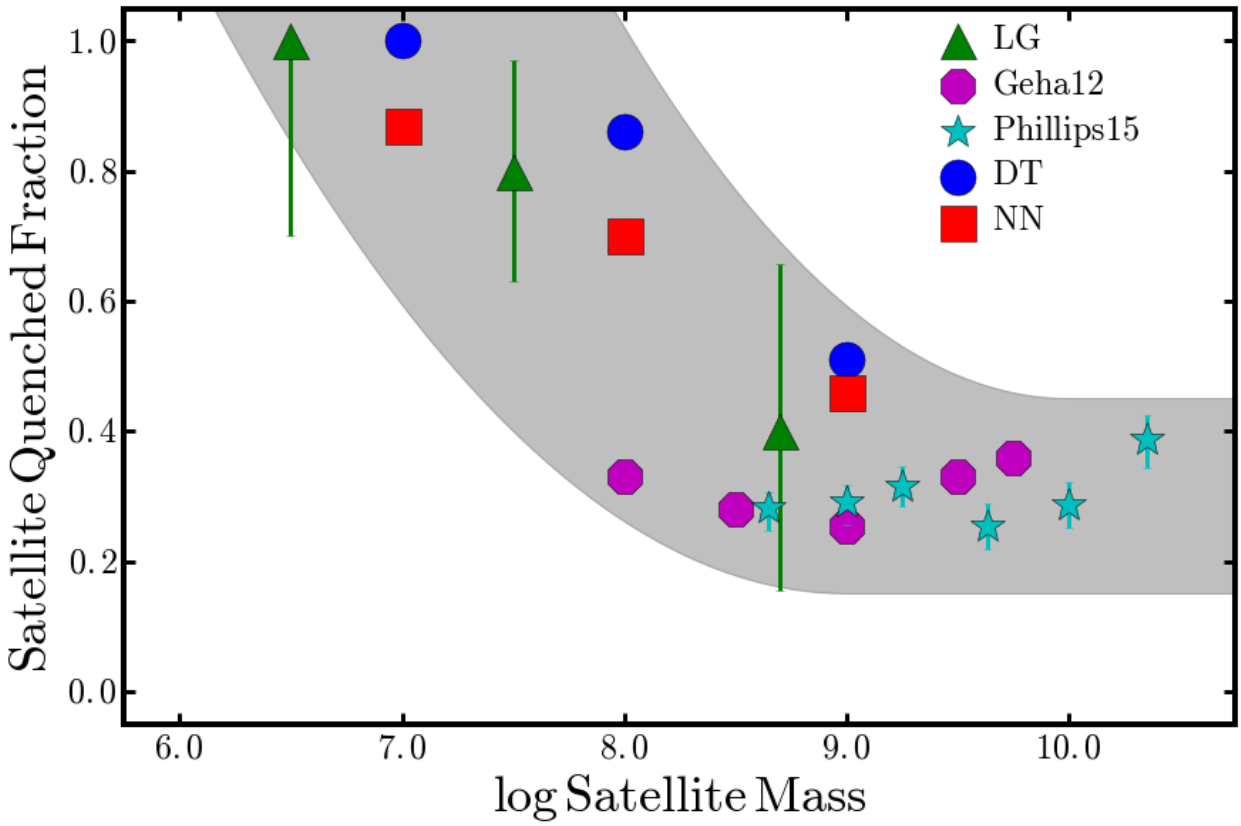

Figure 3.6: The quenched fraction of satellite galaxies as a function of satellite mass, plotted alongside previous results, as well as LG objects. As before, red squares denote quenched fractions measured using the *decision tree* algorithm, while blue circles denote those using the *neural network* algorithm (see §3.3). Taken together, these results show that low-mass satellites are significantly more susceptible to being environmentally quenched than high-mass satellites.

Our work extends previous results examining the quenched fraction of satellites in the $z \sim 0$ Universe. In Figure 3.6 we compare our results to this work, also including HI observations of Local Group satellites. Taking all the data together, the data paints a coherent picture of satellite quenching over a substantial dynamical mass range: high mass galaxies ($\mathbf{M_\star} >\sim 10^9\,\mathbf{M_\odot}$) are quenched fairly inefficiently ($f_{quench} \sim .4$). This has been seen in previous studies, but is confirmed by our work. At intermediate masses ($\mathbf{M_\star} \sim 10^8\,\mathbf{M_\odot}$), galaxies undergo a transition, where they begin to become quenched with high efficiency. At low masses ($\mathbf{M_\star} <\sim 10^8\,\mathbf{M_\odot}$), galaxies are nearly uniformly quenched. These trends are shown by the gray band in Figure 3.6. Our study provides evidence that the increase in quenching efficiency at low satellite mass seen in the LG is a general feature of satellite galaxy evolution.

Our results suggest the idea that below $10^9\,\mathbf{M_\star}$, the depth of a galaxy's potential well is insufficient to retain its gas upon falling onto a central galaxy; the pressure exerted by the circumgalactic medium of the central is greater than the internal gravitational pressure exerted by the gravitating mass of the satellite. Since dwarf galaxies are dominated by their dark matter, this result has implications for how galaxies populate dark matter halos. Since the quenching of low-mass satellites appears common, such galaxies are unlikely to be found in deep potential wells, which provides a constraint on the scatter or the stellar mass-halo mass relation.

On the other hand, our result serves as an indication of the validity of the assumption that the satellite systems in the Local Group are not strongly biased in their star formation histories. Note that the galaxies in our sample are of similar surface brightness to the classical MW dwarfs, meaning that bias attributable to surface brightness is not a concern over the mass ranges we are probing. This is an important point, as local observations can provide strong leverage on addressing questions of high-Z galaxy formation and evolution, provided that they are free of bias. Our study provides a novel means of addressing this issue, which already provides tantalizing hints at the universality of quenching in low-mass satellites,

and which paves the way for future studies using deeper observations, such as the Large Synoptic Survey Telescope (LSST) survey. We estimate that an identical study carried out using LSST would provide data from more than 500 hosts reaching satellites of mass 10^6 and below, and tens of thousands of systems at which the critical mass window ($10^8\,\mathrm{M}_\star$) may be investigated.

Chapter 4

Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images

4.1 Introduction

The total cost of bringing a new drug from discovery to approval has exhibited a steady, exponential rise over the past five decades [72]. One contributing factor to this phenomenon, dubbed Eroom's law (Moore's law backwards), appears to be the failure of traditional, pre-clinical models to accurately simulate many of the more complex features of their clinical successors. These pre-clinical, *in vitro* studies serve to quickly and cheaply identify compounds that exhibit promising effects for further study *in vivo*. However, traditional 2D monolayer culture systems (i.e.: petri dishes) lack many features that are present *in vivo*, such as 3D cellular structure, heterogeneous cellularity, cell-cell interactions, the presence of a complex extracellular matrix (ECM), biomechanical forces (e.g. shear forces generated by

fluid flow), and the presence of perfused vasculature [26]. Animal studies, on the other hand, are too complex to analyze and expensive to substitute for *in vitro* pre-screening, and often fail to identify potential human toxicity due to physiological differences between humans and the animal model [47]. In short, a compound that appears effective in traditional, pre-clinical studies may fail spectacularly in the human body, further contributing to the costly societal burden of failed clinical trials [1].

Microphysiological systems (MPSs), or "organ-on-a-chip" platforms, promise to help close the gap between *in vitro* and *in vivo* drug screens [22, 21, 82], and have seen rapid, recent development [37, 73, 50, 12, 23], supported in part through private-public partnerships fostered under the auspices of the National Center for Advancing Translation Science [46]. These MPSs make significant strides toward more accurately modeling the pertinent properties of *in vivo* biological environments for drug discovery, however many remain in a proof-of-concept stage and require complex peripheral equipment and accessories to operate and maintain.

We have demonstrated an MPS for growing vascularized, perfused microtissues [53, 91]. This platform produces highly robust and uniform vascular networks which are suitable for screening anti-tumor compounds [77] and in large-scale drug discovery studies [59], all while requiring little additional training for the user and no added equipment beyond a standard incubator. We have shown that the survival of these miniature tissues is dependent on nutrients delivered through living vasculature. Importantly, by accurately identifying drugs that target tumor cells, the vascular networks that supply them, or both, the system has proven much better at mimicking human drug responses than previous models. In our studies using FDA-approved or clinical trial compounds to target the vasculature, we have found that anti-angiogenic compounds such as sorafenib and axitinib induce regression on sprouting vessels, but do not have profound effect on mature, interconnected vascular networks. Therefore, they often show a milder effect on the vasculature. On the other hand, non-specific, anti-vascular compounds such as bortezomib and vincristine aggressively fragment the vascular

network. In brief, this system exhibits exceptional potential for developing more targeted, effective anti-vascular and anti-angiogenic compounds to target the tumor vasculature without adverse effects on normal tissue.

A remaining obstacle to deploying this system for truly large-scale anti-angiogenic and anti-vascular drug screening is the need to have human experts determine whether each compound is effective in targeting the vasculature network. Effects are categorized as *no-hit*s (i.e. the compound had no effect on the vasculature network), *soft-hit*s (i.e. the compound moderately disrupted the vasculature network or induced vascular regression), or *hard-hit*s (i.e. the compound had a devastating effect on the vasculature network) from a primary screening (see figure 4.1). Once identified from the initial screen, *soft-hit* and *hard-hit* compounds can be further analyzed in a dose-response screen to identify the half maximal inhibitory concentration (IC50), optimized for molecular structure, and subsequently characterized for their pharmacokinetics *in vivo*. *Soft-hit* compounds are treated as anti-angiogenic while *hard-hit* compounds are treated as anti-vascular.

In the past, human raters have made this determination by manually analyzing each pair of before- and after-drug-application images and quantifying their total vessel length difference using AngioTool [101]. However, this workflow is imprecise—e.g. in its insensitivity to anti-angiogenic compounds that do not significantly affect total vessel length of a fully mature vascular network and its reliance on subjective human judgment—and low throughput—for its need to carefully tune several dataset-specific parameters in the software and the time it takes a human to look at each image.

Automatic classification of these images via machine learning could provide an attractive replacement to the slow and error-prone process of requiring human ratings. In this paradigm, a set of carefully hand-labeled images would be fed to a classifier which could "learn" to distinguish between classes.

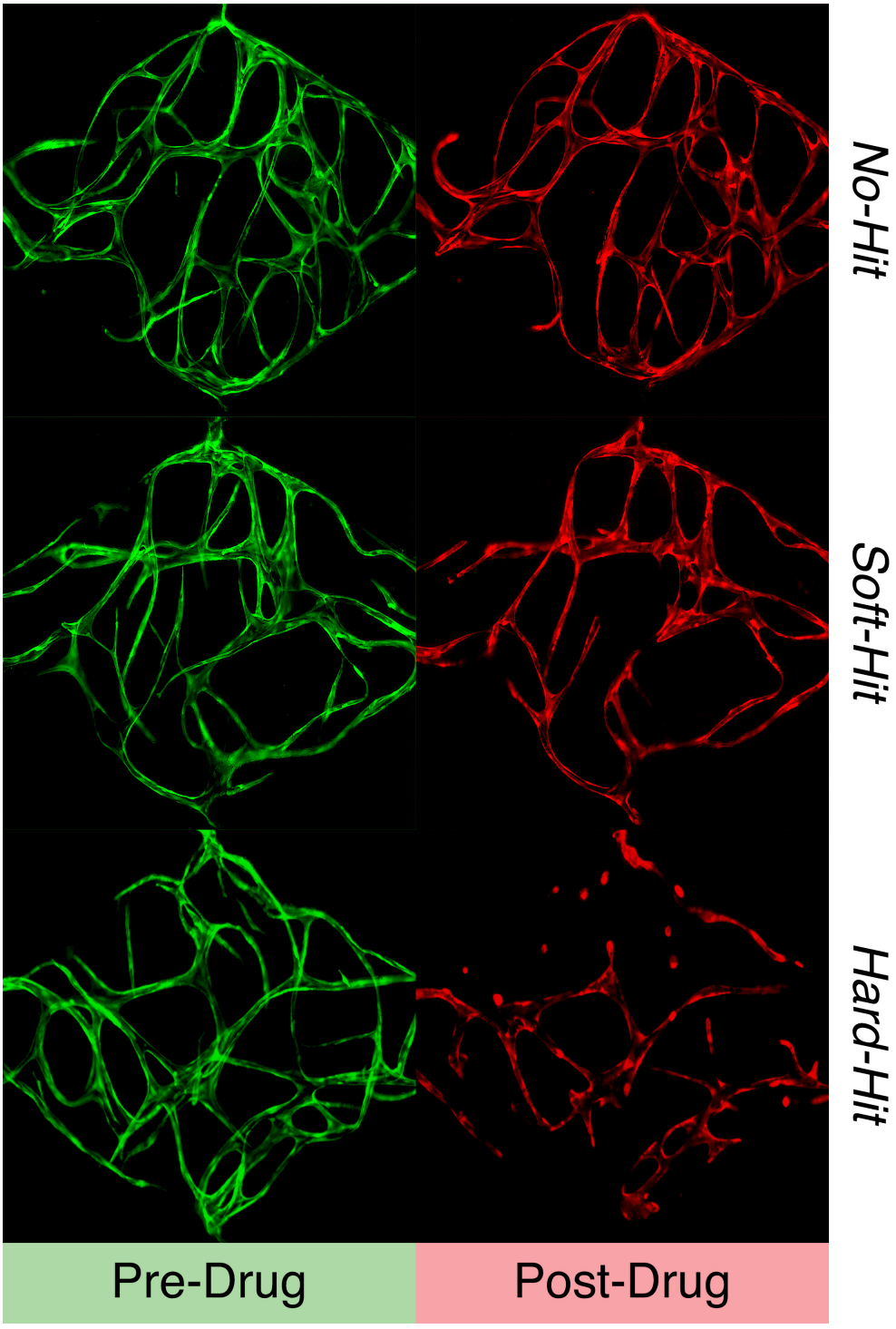

Figure 4.1: Example vessel images

A convolutional neural network is a type of machine learning model that is particularly suited to applications in computer vision. Not only do they offer state-of-the-art performance in

general image classification tasks (e.g. [84]), they have also proven effective for biological applications, with past work demonstrating convolutional networks capable of detecting cardiovascular disease [89], spinal metastasis [90], and skin cancer [24] from medical images.

In this paper, we develop a convolutional neural network to automatically classify images of vasculature networks formed in our MPS into *no-hit*, *soft-hit*, and *hard-hit* categories. The accuracy of our best model is significantly better than our minimally-trained human raters and requires no human intervention to operate. This model is a first step toward automation of data analysis for high-throughput drug screening.

4.2 Methods

4.2.1 Data Collection

Drug studies were performed in the MPS as previously described [59, 77]. Briefly, the cell-ECM suspension was loaded into the platform and cultured for 7 days to allow the vascular network to develop inside the tissue chambers. Each tissue unit was exposed to various compounds obtained from the National Cancer Institute (NCI) Approved Oncology Compound Plate or purchased from Selleck Chemicals. Time course images of vascular network before and after drug treatment were taken using a Nikon Ti-E Eclipse epifluorescent microscope with a 4x Plan Apochromat Lambda objective. For close-up imaging of the tissue chambers, a 1.5x intermediate magnification setting was used.

4.2.2 Preprocessing

Each image in our dataset was between 1000 and 1300 pixels wide. Images of this size contain far more information than is needed for deep image classification (e.g.: [41] classifies natural

images taken from 1000 classes with 256×256 pixels images), so we downsampled images to create 4 separate constant-size datasets: one each of 128×128px, 192×192px, 256×256px, and 320×320px. Next, we z-normalized each image, subtracting the mean pixel intensity and dividing by the standard deviation of the pixel intensities within that image to obtain images with 0-centered pixel values and unitary standard deviation. This normalization helps our models to converge more quickly and uniformly across random initializations. After all this, we concatenated the pre-drug-application and post-drug-application images to obtain a single, 2-channel image.

Image Alignment

We would like the pre-drug-application and post-drug-application images to spatially align as closely as possible. If they do not, then our model would be required to learn an extra invariance: that the channel images need not be aligned. Because the pre- and post-drug-application images were captured three days apart, it is not in general possible to ensure that the two images will be perfectly aligned (e.g. the later image might be shifted or rotated slightly compared to the original). To combat this effect, we implemented a rigid alignment preprocessing step to align the post-drug image to the pre-drug image using the warpAffine method in OpenCV3[25]. For each image, we tried three sets of transformations:

1. A single Euclidean (translation + rotation) transformation on the full-resolution image.
2. A Euclidean transformation on a smaller (32x32px) copy of the image followed by a Euclidean transformation on the full-resolution image.
3. A translation-only transformation on a smaller (32x32px) copy of the image followed by a Euclidean transformation on the full-resolution image.

From these three, we selected the transformed version which yielded the highest possible

correlation coefficient between the pre- and transformed post-drug image. See figure 4.2 for two examples of this alignment process in action.

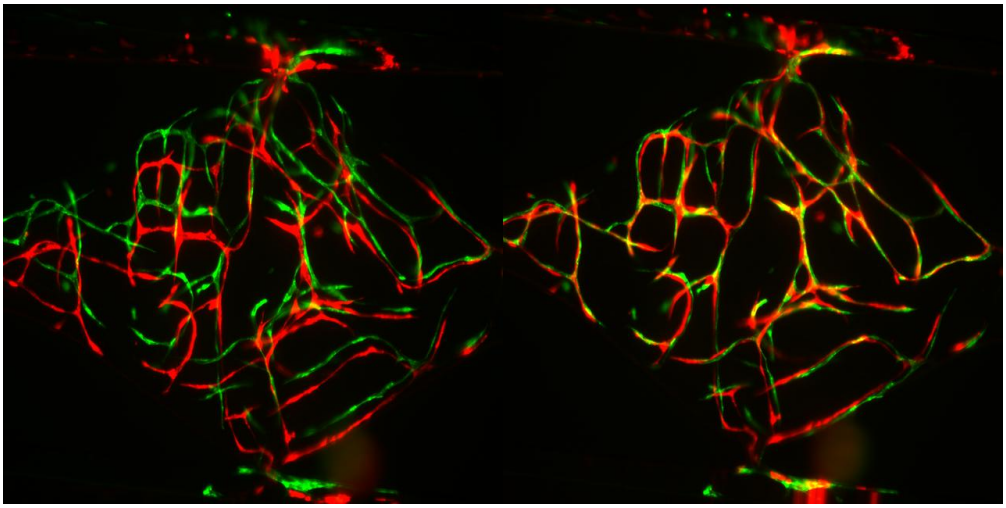

Figure 4.2: A set of blood vessel images before (left) and after (right) alignment. The pre-drug-application images are placed in the image's green channel and the post-drug-application images are placed in the red channel. The separate green and red vessels in the left image shows that the pre- and post-drug-application images are misaligned, the more pervasive yellow in the right image comes from the green and red channels being aligned on top of each other.

4.2.3 Human Ratings

Two human experts rated each of the 277 images, comparing disparate ratings where necessary to come to a consistent set of gold-standard ratings. 164 images were labeled as 0 or *no-hit* (59.2%), 52 were labeled as 1 or *soft-hit* (18.8%), while 61 were labeled as 2 or *hard-hit* (22.0%). These ratings are used throughout the remainder of this paper.

We also obtained ratings from 4 additional humans: undergraduate research assistants who were trained to recognize each image class and who had been assigned this task in the past. Raters were presented with the full set of 277 images in randomized order and were asked to provide an integer class assignment for each using the following instructions: "How much of an effect did the drug have? (0 for no effect, 1 for solid effect, 2 for devastating effect)".

4.2.4 Loss Weighting

For the purposes of drug discovery, false negatives are potentially much costlier than false positives. A false positive (i.e.: predicting that an image from an ineffective drug was actually effective) will result in secondary screening in which the ineffectiveness of the drug may be confirmed. A false negative (i.e.: predicting that an image taken from an effective drug did not actually have any effect) may result in a potentially useful compound being overlooked in this and any future drug trials. To help control our model's false-negative rate, we employed a weighted cross-entropy loss function of the form:

$$\text{loss}(y_i, \hat{y}_i|W) = -\sum_{c=0}^{c=2} W_{c_{itrue},c}\, y_{ic} \log(\hat{y}_{ic})$$

where i indexes over datapoints, c over classes, y_{ic} is an indicator variable that takes the value of 1 if the true class of datapoint i is c and 0 otherwise, c_{itrue} represents the true label of datapoint i (i.e.: 0, 1, or 2), and the weights $W_{c_{itrue},c}$ are drawn from the hand-tuned confusion weighting matrix shown in table 4.1. Note that if all elements of this weight matrix were set to 1.0, then our weighted cross-entropy loss would reduce to standard cross-entropy.

Table 4.1: Loss function weight values

	$Y_{ipred} = 0$	$Y_{ipred} = 1$	$Y_{ipred} = 2$
$Y_{itrue} = 0$	0.8	0.8	0.8
$Y_{itrue} = 1$	2.0	1.0	0.8
$Y_{itrue} = 2$	2.0	0.8	1.0

This loss function penalizes false negatives at twice the default value. In addition, it penalizes the treatment of all true *no-hit* images at 0.8 times the default value and reduces the penalties for confusing soft- and hard-hits to the same amount. We arrived at these weights through trial and error and use them for all experiments presented in this paper.

4.2.5 Training Procedure

We partitioned the full dataset of 277 images into a test set consisting of 25% of the images (69 images) and a training+validation set consisting of 75% of the images (208 images). We employed 4-fold cross validation on the training+validation set, training on 75% of its datapoints (156 images) and tracking validation loss on the remaining 25% (52 images). Unless otherwise noted, we trained on each fold for a total of 200 epochs. All linear and neural models presented in this paper were built in Keras [16] with a Theano [86] backend and trained on NVIDIA GPUs. We selected the model from each fold which attained the lowest validation-set loss value across all training epochs.

We combined the best models from each fold into a 4-model ensemble of models. We averaged the predictions across all 4 models in the ensemble to attain final predictions for each set of hyperparameters on the test dataset.

Data Augmentation

Since our training set is rather small, we employed random data augmentation during training. In each pass over the data, each training image was randomly rotated between -5 and 5 degrees clockwise, translated between -5% and 5% vertically and horizontally, zoomed in between 0 and 10%, and possibly flipped horizontally and vertically, with each transformation value selected uniformly at random from the legal range. Empty pixels that resulted from the random rotation and translation were filled with the values from their nearest existing neighbor pixel. Figure 4.3 shows three randomly transformed versions of one training image. This random data augmentation scheme with continuous parameters yields an infinitude of variations for each 156-image training set and helps prevent our models from overfitting to the specific details of our training data.

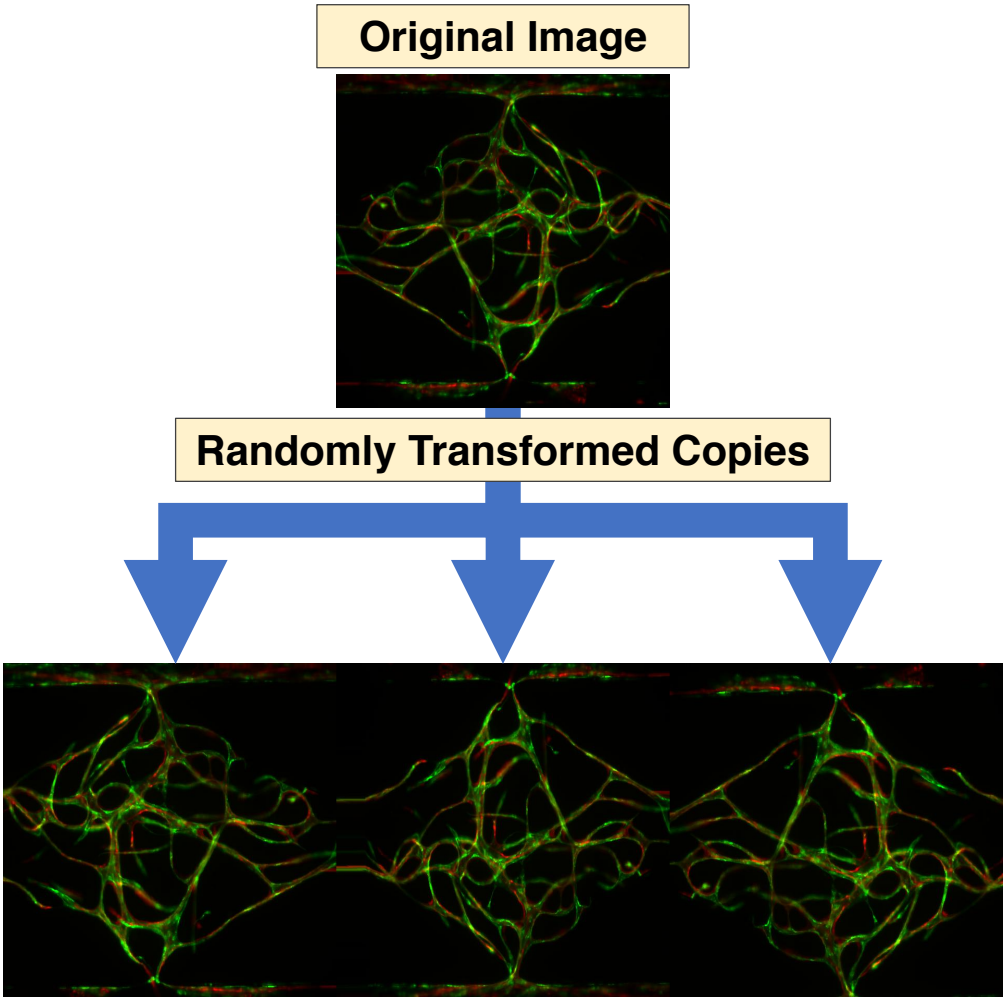

Figure 4.3: Three examples of the data augmentation process used for training and inference. The top image is an actual training image, and the bottom three are randomly transformed copies of that image. Each time an image is visited during the training process, it is first randomly transformed in a way that simulates creating new images with respect to the true invariances of the training images (e.g.: an image should have the same class as a copy of that image which is slightly shifted, rotated, or flipped). The left-most randomly generated image has been flipped horizontally, zoomed, and rotated slightly. The middle random image has been flipped both horizontally, vertically, and zoomed slightly. The right-most image has been flipped vertically, zoomed in, and translated down slightly. This random augmentation helps simulating a larger training set and prevent our model from overfitting.

At inference time, we randomly generated five versions of each validation or test image and averaged the model's predictions for each image over all five of its randomly-generated copies.

4.2.6 Linear Models

We first tried to classify the data with a simple, linear model. For this purpose, we treated the raw pixel intensity values of the concatenated before and after images as the features for this model and trained a series of separate logistic regression models on input images of sizes 128×128, 192×192, 256×256, and 320×320. For each of these image sizes, we fit models with L1 and L2 (i.e. ElasticNet) regularization strengths of 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, and 1e-6 on the weights matrix.

Optimization was completed using a batch size of 32 and and an AdaDelta optimizer [98] with a hand-tuned learning rate of 0.1 and per-epoch learning rate decay factor of 0.98.

4.2.7 Convolutional Neural Network Models

Where the logistic models from section 4.2.6 are limited to a linear class separation boundary in feature space, feed-forward neural networks with even a single hidden layer are theoretically capable of fitting any possible decision boundary[17]. This additional representational power is often useful for complex classification tasks, making deep neural networks some of the most useful models for modern machine learning applications.

Convolutional neural networks offer a slight refinement over feed-forward neural nets by introducing a weight-sharing scheme into certain 'convolutional' layers[43]. These layers learn translation-invariant filters which, when applied as part of an image classification neural network model, achieve state-of-the-art classification performance on a variety of tasks [84, 68, 100].

Standard convolutional architectures for image classification include a series of convolutional layers followed by one or more fully connected layers [43, 41, 84]. Each convolutional and fully connected layer is followed by a rectified linear unit (ReLU) nonlinearity [54] and max

pooling layers are interspersed through some subset of the convolutional layers to repress non-maximal responses and reduce the number of parameters in subsequent layers. Dropout may also be used on some of the convolutional and fully connected layers to help prevent overfitting [79].

Overall, convolutional neural networks offer a well-established process for performing high-quality image classification.

4.2.8 Hyperparameter Search for Convolutional Architectures

Building a convolutional neural network requires specifying a large number of hyperparameters, such as the number of convolutional and fully-connected layers in the network, the size of each layer, dropout probabilities etc. The number of possible hyperparameter combinations grows exponentially with the number of hyperparameters, so a thorough grid search of hyperparameter combinations quickly becomes unwieldy [4].

Instead, we employ a Gaussian-process-based meta-model which maps from a set of chosen hyperparameters to an estimate of the out-of-sample accuracy attained by a model trained with the given hyperparameters[76]. This meta-model of hyperparameter fitness is used in an outer-loop hyperparameter optimization process (see algorithm 1). First, the meta-model proposes a hyperparameter set to try. For each hyperparameter set, we follow the same training procedure as that detailed in section 4.2.6, using 4-fold cross-validation on the training+validation set, building a 4-model ensemble from the best version of the model for each fold (across epochs and as judged by validation-set accuracy), and averaging each model's validation- and test-set predictions over 5 randomly generated versions of each input image. At the end of training, we report the validation-set accuracy (averaged across all 4 folds) as the objective value attained for the given hyperparameter set. This objective value is used to update the meta-model of hyperparameter quality and the process repeats.

4.2.9 Pre-Trained Convolutional Architecture

Given the small size of our training dataset, we next tried a large convolutional architecture that had been pre-trained on a large, general purpose image recognition problem. For this purpose we picked the InceptionV3 architecture [85] as implemented in Keras [16] with weights that had been pre-trained on the ImageNet classification challenge [71]. The full convolutional portion of the InceptionV3 model contains 21,611,968 parameters and some 216 layers. We instantiated the model without including the final fully-connected layers, opting not to fine-tune its convolutional weights, but to train two fully connected and one 3-class softmax layer anew for our classification problem while using the convolutional portion of the InceptionV3 model as an elaborate, fixed computer vision preprocessing routine.

While fixing our convolutional architecture fixed many of the hyperparameters of our model, several still remained. These were: the input image size (192×192px, 256×256px, or 320×320px; we skip the 128×128px version because InceptionV3 requires input images to be at least 139px×139px), the number of neurons in the first fully connected layer (16, 32, 64, 128, 256), dropout probability for the dropout layer immediately after the first fully connected layer (0.0 to 0.99), the number of neurons in the second fully connected layer (16, 32, 64, 128, 256, 512, 1024), dropout probability for the dropout layer immediately after the second fully connected layer (0.0 to 0.99), the optimization batch size (16 to 64), $\log_{10}$ of the learning rate (-3.0 to 0.0), $\log_{10}$ of the L1 penalty applied to the weights of the network (-9.0 to -1.0), $\log_{10}$ of the L2 penalty applied to the weights of the network (-9.0 to -1.0).

4.2.10 Custom Convolutional Architecture

Though the Inception architecture employed in section 4.2.9 has proven very useful for general-purpose image classification, the images of microscopic blood vessel networks used in this task have their own structure that does not necessarily match the constraints of general

object recognition[1].

For this purpose, we also trained a series of custom convolutional architectures specifically for this blood-vessel classification task. We constrained our architecture to contain several convolutional layers followed by two fully connected layers.

The hyperparameter ranges that we considered were: the input image size (128×128px, 192×192px, 256×256px, or 320×320px), the number of convolutional layers in the model (2 to 7), the number of convolutional filters at the start of the convolutional cascade (16, 32, 64, 128 or 256), the number of convolutional filters at the end of the cascade (16, 32, 64, 128 or 256; filter counts were linearly interpolated across the 2 to 7 convolutional layers between the number of filters at the start and the number of filters at the end), the number of convolutional layers between max pooling layers (1 to 4; the first pooling layer was fixed after the second convolutional layer), the size of the max pooling receptive fields (2 to 7; stride was fixed to match pooling size), dropout probability for a dropout layer immediately after the convolutional layers (0.00 to 0.99), the number of neurons in the first fully connected layer (16, 32, 64, 128, 256), dropout probability for the dropout layer immediately after the first fully connected layer (0.0 to 0.99), the number of neurons in the second fully connected layer (16, 32, 64, 128, 256, 512, 1024), dropout probability for the dropout layer immediately after the second fully connected layer (0.0 to 0.99), the optimization batch size (1 to 8), $\log_{10}$ of the learning rate (-3.0 to 0.0), $\log_{10}$ of the L1 penalty applied to the weights of the network (-9.0 to -1.0), $\log_{10}$ of the L2 penalty applied to the weights of the network (-9.0 to -1.0).

[1]For example, detecting eyes is very important for detecting the myriad animal types in ImageNet, but irrelevant for our task.

4.3 Results

4.3.1 Human Rating Results

The four human raters found the vessel rating task difficult compared to the expert raters, matching the gold-standard ratings 72.9%, 76.5%, 69.3% and 83.0% of the time. The rounded average of all four raters' ratings (i.e.: 0, 1, or 2) matched the gold standard ratings 85.9% of the time. (See table 4.2 and section 4.4 for further details).

Table 4.2: Test Set Confusion Matrix for Average of Four Human Raters

	$Y_{ipred} = 0$	$Y_{ipred} = 1$	$Y_{ipred} = 2$
$Y_{itrue} = 0$	86%	14%	0
$Y_{itrue} = 1$	27%	65%	9%
$Y_{itrue} = 2$	0	0	100%

4.3.2 Linear Model Results

The best linear ensemble that we found for this task, as judged by validation set accuracy (67.3%) used an input image size of 256px×256px and $\log_{10}$ of L1 and L2 regularization strength equal to -2.0. This ensemble did not perform better than chance on the test set, achieving a three-class test set accuracy of 56.2% (the largest class made up 62.3% of the test set; see the confusion matrix in table 4.3 for details).

Table 4.3: Test Set Confusion Matrix for Linear Ensemble

	$Y_{ipred} = 0$	$Y_{ipred} = 1$	$Y_{ipred} = 2$
$Y_{itrue} = 0$	84%	2%	14%
$Y_{itrue} = 1$	91%	0	9%
$Y_{itrue} = 2$	60%	7%	33%

4.3.3 Pre-Trained Convolutional Neural Network Results

We explored a total of 100 hyperparameter sets for the pretrained convolutional architecture using the procedure explained in section 4.2.8. The best model, as judged by three-way validation-set accuracy (87.0%), used 320px×320px input images, its first fully connected layer after the InceptionV3 convolutional stack contained 256 neurons, its second fully connected layer contained 1024 neurons, and the final dropout probability before the 3-way softmax layer was 0.27[2].

The optimization was completed with a batch size of 16, $\log_{10}$ of the learning rate of -1.24, a per-epoch learning rate decay factor of 0.98, $\log_{10}$ of L1 shrinkage of -9.0, and $\log_{10}$ of L2 shrinkage of -1.0.

A 4-model ensemble based on this architecture achieved a three-class accuracy value of 87.0% on the hitherto-unseen test (see the confusion matrix in table 4.4 for details).

Table 4.4: Test Set Confusion Matrix for Pre-Trained Convolutional Ensemble

	$Y_{ipred} = 0$	$Y_{ipred} = 1$	$Y_{ipred} = 2$
$Y_{itrue} = 0$	98%	2%	0
$Y_{itrue} = 1$	45%	36%	18%
$Y_{itrue} = 2$	0	7%	93%

4.3.4 Custom Convolutional Neural Network Results

We explored a total of 1000 hyperparameter sets for our custom convolutional architecture, the best of which, as judged by three-class validation-set accuracy (96.6%), is a 21-layer convolutional neural network, the architecture for which is illustrated in figure 4.4.

The optimization was completed with a batch size of 1, $\log_{10}$ of the learning rate of -1.91,

[2]This model contained 21,611,968 fixed parameters that had been pre-trained on ImageNet data and 33,820,931 fully connected parameters that were trained on the vessel data.

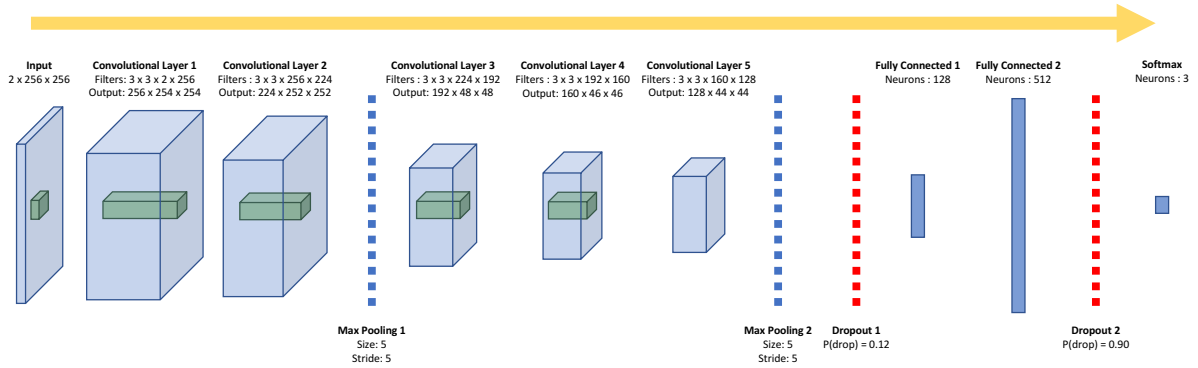

Figure 4.4: The architecture for the best convolutional neural network we trained on these data. The blue prisms represent the 3-dimensional input images (two channels, width, and height) and the three dimensional output of each convolutional layer (filters, width, and height). The green prisms represent a sample receptive field for the subsequent convolutional layer.

a per-epoch learning rate decay factor of 0.98, $\log_{10}$ of L1 shrinkage of -9.0, and $\log_{10}$ of L2 shrinkage of -9.0.

A 4-model ensemble based on this architecture achieved a three-class accuracy value of 95.7% on the hitherto-unseen test set with no false negatives (see the confusion matrix in table 4.5 for details).

Table 4.5: Test Set Confusion Matrix for Custom Convolutional Ensemble

	$Y_{ipred} = 0$	$Y_{ipred} = 1$	$Y_{ipred} = 2$
$Y_{itrue} = 0$	100%	0	0
$Y_{itrue} = 1$	0	82%	18%
$Y_{itrue} = 2$	0	7%	93%

4.4 Discussion

In this paper, we present a new classification problem—to distinguish effective from ineffective drug compounds through automatic analysis of vascularization images. A cursory glance at figure 4.1 might tempt the casual reader to conclude that this problem is fairly

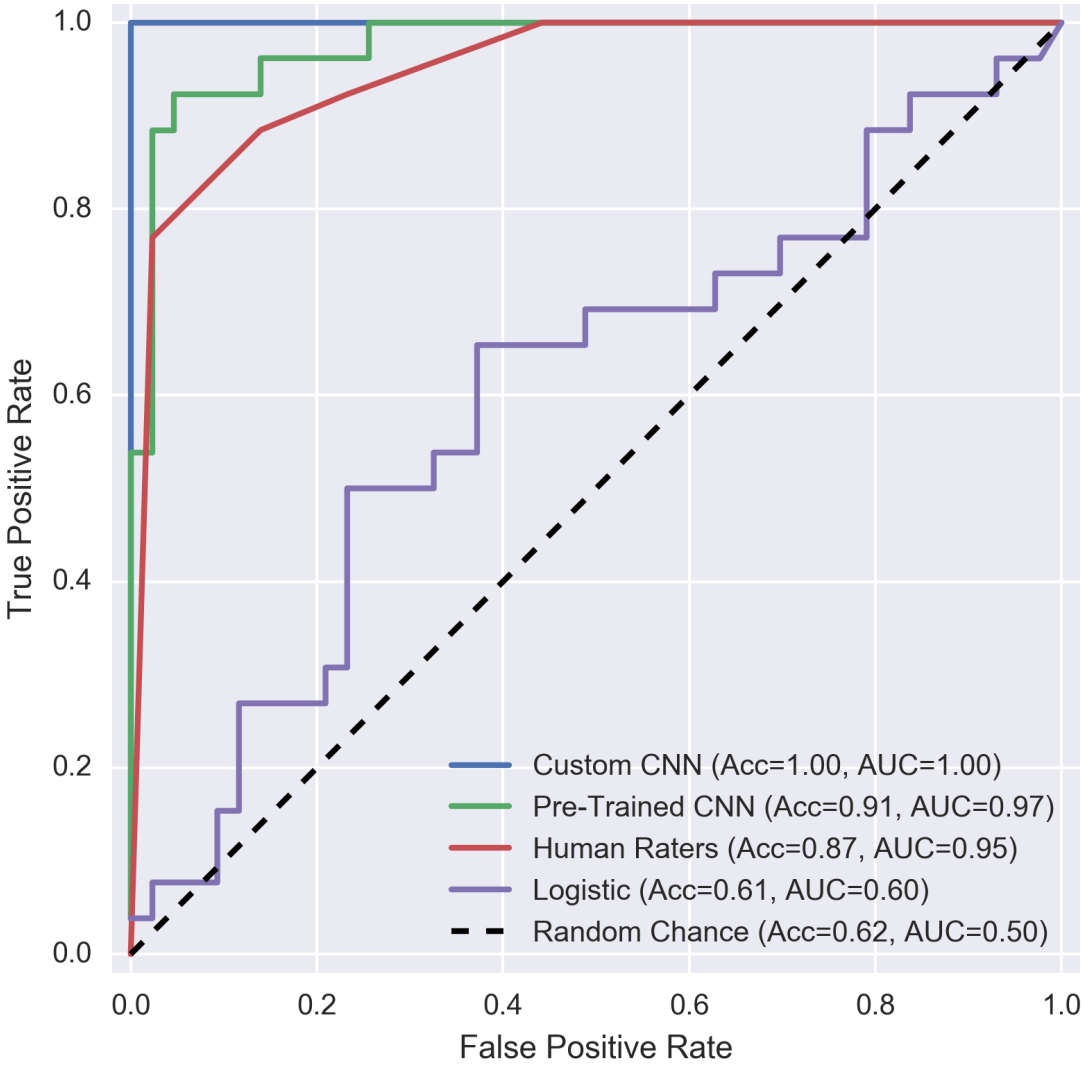

Figure 4.5: Receiver operating characteristic curves for a binarized version of this classification problem (*no-hit* vs. *soft-hit* or *hard-hit*). ROC-AUC scores range between 0.5 and 1.0, with 0.5 indicating performance at chance and 1.0 indicating perfect classification (a standard which the best custom convolutional neural network we tried achieves on this binarized problem).

straightforward and that it could be solved by simply counting the number of pixels which are present in the pre-drug-application image but missing in the post-drug-application image. The results listed in section 4.3 suggest that this is not the case. While a linear model

would prove effective in such a problem regime, the logistic models which we trained utterly failed to distinguish between effective and ineffective compounds (three-way accuracy: 56.2%; majority class size: 62.3%). The difficulty appears to be driven by the nuances of the classification problem, which cannot be captured in a simple linear decision boundary in pixel space. For example, the death of a *bridge-to-nowhere* vessel should be treated as less important than the death of a vessel on a major thoroughfare in the vasculature network. To further highlight its difficulty, even an ensemble of four trained human raters had some difficulty with this task (three-way accuracy: 85.9%).

Convolutional neural networks, however, appear equal to the challenge. Convolutional neural networks have already demonstrated super-human classification performance on general computer vision tasks [35], and the pattern holds for this new classification problem. Where a cadre of four undergraduate raters achieved a three-way accuracy of 85.9% on this dataset, a convolutional ensemble based on the InceptionV3 architecture [85] and pre-trained on ImageNet data [71] achieved three-way accuracy of 87.0% (though it committed more false negatives than the human raters). A custom convolutional architecture, however, achieves a robust 95.7% three-way accuracy for drug-hit classification, while committing no false negatives. This pattern repeats itself if we reduce our 3-way classification problem to a binary problem by aliasing together the *soft-hit* and *hard-hit* categories (see figure 4.5).

The success of this convolutional model is driven in part by carefully tuning our loss function to discourage false negatives (see section 4.2.4), but also by the great steps we took to control overfitting in our model. One such regularization strategy was to augment our limited training dataset to virtually infinite size via randomly transforming images during each training pass (see section 4.2.5). Judicious use of dropout also contributed to the result. In fact, the hyperparameter optimization scheme that we used automatically picked a model with a large final layer (512 neurons) and a high dropout probability (0.90). Dropout can be interpreted as implicitly performing a geometric average over an ensemble of regularized

subnetworks [3], so this model can be interpreted as implicitly averaging over an enormous ensemble of diverse sub-networks.

These regularization strategies were important, as our final network contained 2,485,827 learned parameters and 15 optimized hyperparameters, more than enough capacity to memorize the identity of 208 training+validation datapoints. However, our network still exhibits excellent generalization power, with test accuracy of 95.7% only barely lagging behind the hyperparameter optimized 96.6% validation accuracy which in turn closely follows the training accuracy of 98.1%. This tendency toward strong generalization performance is often seen in deep networks, and cannot yet be fully explained by any known regularization mechanism or learning theory[99].

4.5 Conclusion

In this paper, we have developed a convolutional neural network to improve the data analysis processes for high-throughput drug screening using our MPS. This network can classify new images near instantaneously and surpasses human accuracy on this task. A larger scale drug screening can be achieved by coupling this classifier and an automated microscope camera system to capture images before and after drug treatment.

Bibliography

[1] J. Arrowsmith and P. Miller. Trial watch: phase ii and phase iii attrition rates 2011-2012. *Nature Reviews Drug Discovery*, 12(8):569–569, 2013.

[2] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

[3] P. Baldi and P. J. Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems*, pages 2814–2822, 2013.

[4] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[5] M. R. Blanton, D. J. Schlegel, M. A. Strauss, J. Brinkmann, D. Finkbeiner, M. Fukugita, J. E. Gunn, D. W. Hogg, Ž. Ivezić, G. R. Knapp, R. H. Lupton, J. A. Munn, D. P. Schneider, M. Tegmark, and I. Zehavi. New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *AJ*, 129:2562–2578, June 2005.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] M. S. Bovill and M. Ricotti. Where are the Fossils of the First Galaxies? I. Local Volume Maps and Properties of the Undetected Dwarfs. *ApJ*, 741:17, Nov. 2011.

[8] M. Boylan-Kolchin, D. R. Weisz, J. S. Bullock, and M. C. Cooper. The Local Group: the ultimate deep field. *MNRAS*, 462:L51–L55, Oct. 2016.

[9] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] J. Brinchmann, S. Charlot, S. D. M. White, C. Tremonti, G. Kauffmann, T. Heckman, and J. Brinkmann. The physical properties of star-forming galaxies in the low-redshift Universe. *MNRAS*, 351:1151–1179, July 2004.

[11] R. N. Broadus. An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science*, 34(2):132–135, 2007.

[12] J. A. Brown, V. Pensabene, D. A. Markov, V. Allwardt, M. D. Neely, M. Shi, C. M. Britt, O. S. Hoilett, Q. Yang, B. M. Brewer, et al. Recreating blood-brain barrier physiology and structure on chip: A novel neurovascular microfluidic bioreactor. *Biomicrofluidics*, 9(5):054124, 2015.

[13] T. M. Brown, J. Tumlinson, M. Geha, E. N. Kirby, D. A. VandenBerg, R. R. Muñoz, J. S. Kalirai, J. D. Simon, R. J. Avila, P. Guhathakurta, A. Renzini, and H. C. Ferguson. The Primeval Populations of the Ultra-faint Dwarf Galaxies. *ApJ*, 753:L21, July 2012.

[14] J. S. Bullock, A. V. Kravtsov, and D. H. Weinberg. Reionization and the Abundance of Galactic Satellites. *ApJ*, 539:517–521, Aug. 2000.

[15] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. Scratch: a protein structure and structural feature prediction server. *Nucleic acids research*, 33(suppl 2):W72–W76, 2005.

[16] F. Chollet. keras. https://github.com/fchollet/keras, 2015.

[17] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[18] D. Davies. Citation idiosyncrasies. *Nature*, 228:1356, 1970.

[19] G. De Lucia, S. Weinmann, B. M. Poggianti, A. Aragón-Salamanca, and D. Zaritsky. The environmental history of group and cluster galaxies in a Λ cold dark matter universe. *MNRAS*, 423:1277–1292, June 2012.

[20] J. Dillon, Y. Mao, G. Lebanon, and J. Zhang. Statistical translation, heat kernels and expected distances. In *Proceedings of the Uncertainty in AI Conference (UAI 2007)*, pages 93–100, 2007.

[21] E. W. Esch, A. Bahinski, and D. Huh. Organs-on-chips at the frontiers of drug discovery. *Nature reviews Drug discovery*, 14(4):248–260, 2015.

[22] M. Esch, T. King, and M. Shuler. The role of body-on-a-chip devices in drug and toxicity studies. *Annual review of biomedical engineering*, 13:55–72, 2011.

[23] M. B. Esch, H. Ueno, D. R. Applegate, and M. L. Shuler. Modular, pumpless body-on-a-chip platform for the co-culture of gi tract epithelium and 3d primary liver tissue. *Lab on a Chip*, 16(14):2719–2729, 2016.

[24] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[25] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

[26] K. M. Fabre, C. Livingston, and D. A. Tagle. Organs-on-chips (microphysiological systems): tools to expedite efficacy and toxicity testing in human tissue. *Experimental Biology and Medicine*, page 1535370214538916, 2014.

[27] L. Ferrarese, P. Côté, J.-C. Cuillandre, S. D. J. Gwyn, E. W. Peng, L. A. MacArthur, P.-A. Duc, A. Boselli, S. Mei, T. Erben, A. W. McConnachie, P. R. Durrell, J. C. Mihos, A. Jordán, A. Lançon, T. H. Puzia, E. Emsellem, M. L. Balogh, J. P. Blakeslee, L. van Waerbeke, R. Gavazzi, B. Vollmer, J. J. Kavelaars, D. Woods, N. M. Ball, S. Boissier, S. Courteau, E. Ferriere, G. Gavazzi, H. Hildebrandt, P. Hudelot, M. Huertas-Company, C. Liu, D. McLaughlin, Y. Mellier, M. Milkeraitis, D. Schade, C. Balkowski, F. Bournaud, R. G. Carlberg, S. C. Chapman, H. Hoekstra, C. Peng, M. Sawicki, L. Simard, J. E. Taylor, R. B. Tully, W. van Driel, C. D. Wilson, T. Burdullis, B. Mahoney, and N. Manset. The Next Generation Virgo Cluster Survey (NGVS). I. Introduction to the Survey. *ApJS*, 200:4, May 2012.

[28] S. P. Fillingham, M. C. Cooper, A. B. Pace, M. Boylan-Kolchin, J. S. Bullock, S. Garrison-Kimmel, and C. Wheeler. Under pressure: quenching star formation in low-mass satellite galaxies via stripping. *MNRAS*, 463:1916–1928, Dec. 2016.

[29] S. P. Fillingham, M. C. Cooper, C. Wheeler, S. Garrison-Kimmel, M. Boylan-Kolchin, and J. S. Bullock. Taking care of business in a flash: constraining the time-scale for low-mass satellite quenching with ELVIS. *MNRAS*, 454:2039–2049, Dec. 2015.

[30] M. O. Finkelstein and R. M. Friedberg. The application of an entropy theory of concentration to the Clayton act. *Yale Law Journal*, 76:677, 1966.

[31] K. Freeman and J. Bland-Hawthorn. The New Galaxy: Signatures of Its Formation. *ARA&A*, 40:487–537, 2002.

[32] J. Gibbs and W. Martin. Urbanization, technology, and the division of labor: International patterns. *American Sociological Review*, pages 667–677, 1962.

[33] J. Gillenwater, A. Kulesza, and B. Taskar. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 710–720. Association for Computational Linguistics, 2012.

[34] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[36] P. Hodge. Populations in Local Group galaxies. *ARA&A*, 27:139–159, 1989.

[37] D. Huh, B. D. Matthews, A. Mammoto, M. Montoya-Zavala, H. Y. Hsin, and D. E. Ingber. Reconstituting organ-level lung functions on a chip. *Science*, 328(5986):1662–1668, 2010.

[38] R. A. Ibata, G. F. Lewis, A. R. Conn, M. J. Irwin, A. W. McConnachie, S. C. Chapman, M. L. Collins, M. Fardal, A. M. N. Ferguson, N. G. Ibata, A. D. Mackey, N. F. Martin, J. Navarro, R. M. Rich, D. Valls-Gabaud, and L. M. Widrow. A vast, thin plane of corotating dwarf galaxies orbiting the Andromeda galaxy. *Nature*, 493:62–65, Jan. 2013.

[39] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[40] G. Kauffmann, C. Li, W. Zhang, and S. Weinmann. A re-examination of galactic conformity and a comparison with semi-analytic models of galaxy formation. *MNRAS*, 430:1447–1456, Apr. 2013.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[42] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[43] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[44] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[45] S. Lieberson. Measuring population diversity. *American Sociological Review*, pages 850–862, 1969.

[46] C. A. Livingston, K. M. Fabre, and D. A. Tagle. Facilitating the commercialization and use of organ platforms generated by the microphysiological systems (tissue chip) program through public–private partnerships. *Computational and Structural Biotechnology Journal*, 14:207–210, 2016.

[47] L. A. Low and D. A. Tagle. Tissue chips to aid drug development and modeling for rare diseases. *Expert Opinion on Orphan Drugs*, 4(11):1113–1121, 2016.

[48] A. Magurran and A. Magurran. *Ecological Diversity and its Measurement*, volume 168. Princeton University Press, Princeton, NJ, 1988.

[49] M. L. Mateo. Dwarf Galaxies of the Local Group. *ARA&A*, 36:435–506, 1998.

[50] A. Mathur, P. Loskill, K. Shao, N. Huebsch, S. Hong, S. G. Marcus, N. Marks, M. Mandegar, B. R. Conklin, L. P. Lee, et al. Human ipsc-based cardiac microphysiological system for drug screening applications. *Scientific reports*, 5:8883, 2015.

[51] A. K. McCallum. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet, 2002.

[52] A. W. McConnachie. The Observed Properties of Dwarf Galaxies in and around the Local Group. *AJ*, 144:4, July 2012.

[53] M. L. Moya, Y.-H. Hsu, A. P. Lee, C. C. Hughes, and S. C. George. In vitro perfused human capillary networks. *Tissue Engineering Part C: Methods*, 19(9):730–737, 2013.

[54] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[55] U. N. L. o. M. National Center for Biotechnology Information. Pubmed central open access initiative. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, 2010.

[56] M. Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973.

[57] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[58] A. M. Nierenberg, M. W. Auger, T. Treu, P. J. Marshall, and C. D. Fassnacht. Luminous Satellites of Early-type Galaxies. I. Spatial Distribution. *ApJ*, 731:44, Apr. 2011.

[59] D. T. Phan, X. Wang, B. M. Craver, A. Sobrino, D. Zhao, J. C. Chen, L. Y. Lee, S. C. George, A. Lee, and C. Hughes. A vascularized and perfused organ-on-a-chip platform for large-scale drug screening applications. *Lab on a Chip*, 2017.

[60] J. I. Phillips, C. Wheeler, M. Boylan-Kolchin, J. S. Bullock, M. C. Cooper, and E. J. Tollerud. A dichotomy in satellite quenching around L* galaxies. *MNRAS*, 437:1930–1941, Jan. 2014.

[61] J. I. Phillips, C. Wheeler, M. C. Cooper, M. Boylan-Kolchin, J. S. Bullock, and E. Tollerud. The mass dependence of satellite quenching in Milky Way-like haloes. *MNRAS*, 447:698–710, Feb. 2015.

[62] E. C. Pielou. *An Introduction to Mathematical Ecology*. Wiley-Interscience, 1969.

[63] A. L. Porter and I. Rafols. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745, 2009.

[64] A. L. Porter, D. J. Roessner, and A. E. Heberger. How interdisciplinary is a given body of research? *Research Evaluation*, 17(4):273–282, 2008.

[65] D. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26, 2013.

[66] I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010.

[67] C. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

[68] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[69] C. Ricotta and L. Szeidl. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243, 2006.

[70] M. Ricotti and N. Y. Gnedin. Formation Histories of Dwarf Galaxies in the Local Group. *ApJ*, 629:259–267, Aug. 2005.

[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[72] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

[73] K. Schimek, M. Busek, S. Brincker, B. Groth, S. Hoffmann, R. Lauster, G. Lindner, A. Lorenz, U. Menzel, F. Sonntag, et al. Integrating biological vasculature into a multi-organ-chip microsystem. *Lab on a Chip*, 13(18):3588–3598, 2013.

[74] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, (0), 2017.

[75] E. Simpson. Measurement of diversity. *Nature*, page 688, 1949.

[76] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[77] A. Sobrino, D. T. Phan, R. Datta, X. Wang, S. J. Hachey, M. Romero-López, E. Gratton, A. P. Lee, S. C. George, and C. C. Hughes. 3d microtumors in vitro supported by perfused vascular networks. *Scientific Reports*, 6, 2016.

[78] A. Solow, S. Polasky, and J. Broadus. On the measurement of biological diversity. *Journal of Environmental Economics and Management*, 24(1):60–68, 1993.

[79] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[80] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.

[81] L. E. Strigari and R. H. Wechsler. The Cosmic Abundance of Classical Milky Way Satellites. *ApJ*, 749:75, Apr. 2012.

[82] M. L. Sutherland, K. M. Fabre, and D. A. Tagle. The national institutes of health microphysiological systems program focuses on a critical challenge in the drug discovery pipeline. *Stem cell research & therapy*, 4(1):I1, 2013.

[83] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 1139–1147, 2013.

[84] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[86] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[87] E. Tolstoy, V. Hill, and M. Tosi. Star-Formation Histories, Abundances, and Kinematics of Dwarf Galaxies in the Local Group. *ARA&A*, 47:371–425, Sept. 2009.

[88] C. Wagner, J. Roessner, K. Bobb, J. Klein, K. Boyack, J. Keyton, I. Rafols, and K. Börner. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1):14–26, 2011.

[89] J. Wang, H. Ding, F. Azamian, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi. Detecting cardiovascular disease from mammograms with deep learning. *IEEE Transactions on Medical Imaging*, 2017.

[90] J. Wang, Z. Fang, N. Lang, H. Yuan, M. Su., and P. Baldi. A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. *Computers in Biology and Medicine*, Submitted for Publication.

[91] X. Wang, D. T. Phan, A. Sobrino, S. C. George, C. C. Hughes, and A. P. Lee. Engineering anastomosis between living capillary networks and endothelial cell-lined microfluidic channels. *Lab on a Chip*, 16(2):282–290, 2016.

[92] S. M. Weinmann, F. C. van den Bosch, X. Yang, and H. J. Mo. Properties of galaxy groups in the Sloan Digital Sky Survey - I. The dependence of colour, star formation and morphology on halo mass. *MNRAS*, 366:2–28, Feb. 2006.

[93] D. R. Weisz, A. E. Dolphin, E. D. Skillman, J. Holtzman, K. M. Gilbert, J. J. Dalcanton, and B. F. Williams. The Star Formation Histories of Local Group Dwarf Galaxies. I. Hubble Space Telescope/Wide Field Planetary Camera 2 Observations. *ApJ*, 789:147, July 2014.

[94] D. R. Weisz, A. E. Dolphin, E. D. Skillman, J. Holtzman, K. M. Gilbert, J. J. Dalcanton, and B. F. Williams. The Star Formation Histories of Local Group Dwarf Galaxies. II. Searching For Signatures of Reionization. *ApJ*, 789:148, July 2014.

[95] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *Proceedings of the 20th international conference on World wide web*, pages 237–246. ACM, 2011.

[96] A. R. Wetzel, J. L. Tinker, C. Conroy, and F. C. van den Bosch. Galaxy evolution in groups and clusters: satellite star formation histories and quenching time-scales in a hierarchical Universe. *MNRAS*, 432:336–358, June 2013.

[97] C. Wheeler, J. I. Phillips, M. C. Cooper, M. Boylan-Kolchin, and J. S. Bullock. The surprising inefficiency of dwarf satellite quenching. *MNRAS*, 442:1396–1404, Aug. 2014.

[98] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[99] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[100] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.

[101] E. Zudaire, L. Gambardella, C. Kurcz, and S. Vermeren. A computational tool for quantitative analysis of vascular networks. *PloS one*, 6(11):e27385, 2011.