Dark Matter in the Milky Way

Laura Jan Chang

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Physics
Advisor: Mariangela Lisanti

November 2020
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

(http://creativecommons.org/licenses/by-nc-sa/4.0/).
We have entered a data-driven era of astrophysics and cosmology, providing a wealth of datasets within which to search for the answers to some of the most fundamental open questions in the physics of our Universe. One of these questions is the nature of dark matter (DM)—while there is phenomenal agreement between the theories of DM and the data on cosmological scales, there remains much to be understood about DM on scales at and smaller than the size of galaxies.

This thesis explores the astrophysical and particle physics properties of dark matter in the Milky Way Galaxy. Chapters 2–4 center around indirect detection of DM, the field of research that seeks to detect the Standard Model particles which result from DM annihilation (or decay). The focus here is specifically on searching for signatures of DM annihilation in gamma-ray data from the *Fermi* Large Area Telescope. Chapters 5–6 are dedicated to understanding substructure in the Milky Way. Chapter 5 focuses on characterizing how well the standard Jeans dynamical mass modeling method performs at accurately capturing the DM content of dwarf galaxies, while Chapter 6 presents a novel machine learning-based approach to inferring the missing information from *Gaia* stellar data, which can then be used to search for evidence of stellar and DM substructure in the Milky Way.
Acknowledgements

It truly takes a village to complete a PhD thesis.

I must first thank my advisor, Mariangela Lisanti, for taking a chance on me when I first knocked on her office door four years ago. As with any sustained relationship between two people, we have had our share of ups and downs—what has remained constant throughout it all has been the high standard to which Mariangela holds her work, which has led me to ultimately produce better science than I may have deemed possible at times. I have grown as a scientist and as a person through my collaboration with her.

I must also thank my first advisor at Princeton, Waseem Bakr. His group was what drew me here in the first place, and although my research interests ended up evolving over time, I am deeply grateful that he convinced me to come to Princeton, a decision which I still do not regret.

The work presented in this thesis would not have been possible without the computational resources managed and supported by Princeton Research Computing, a consortium of groups including the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology’s High Performance Computing Center and Visualization Laboratory at Princeton University. I have been spoiled by the extensive computing resources and dedicated research computing staff at Princeton. I have also been fortunate to have the financial support of an NSF Graduate Research Fellowship and a Paul and Daisy Soros Fellowship for New Americans during most of my time as a graduate student, as well as the support of a Princeton DataX grant during my final year. Through the PD Soros Fellowship, I met a group of individuals who are doing incredible things, but are first and foremost wonderful people; they inspire me to think bigger as I search for my way to make a difference in the world. I am immensely grateful to have been invited into the PD
Soros family during my time in graduate school—it still feels just as surreal now as it did when I got the phone call three years ago.

Scientific research is not a linear process. Over the years, I have worked on several “failed” projects which ultimately did not make it into any publishable form. In retrospect, I am grateful for those “lost projects,” for I learned much more from them than from the rare projects that progressed more or less seamlessly. I am also grateful to the many collaborators I have worked with along the way, without whom this thesis would have looked very different: Malte Buschmann, Tim Cohen, Adriana Dropulic, Hongwan Liu, Siddharth Mishra-Sharma, Lina Necib, Bryan Ostdiek, Nick Rodd, and Ben Safdi. I have learned something from working with each and every one of them. I am also grateful to the other members of the Lisanti group—it has been fun being amongst them, and I look forward to seeing how the group continues to grow in the years to come.

While the focus of graduate school is undoubtedly research, I have met important mentors through other avenues along the way. I have learned a lot from Kasey Wagoner, whom I first met playing summer softball. Kasey goes above and beyond to make the Princeton Physics Department—and more broadly, the field of physics—a more diverse, inclusive, and equitable place. His dedication is truly inspiring, and the department is very lucky to have him. I have immense gratitude for Lyman Page and Herman Verlinde, both of whom I first became familiar with through serving on the Open House Committee. At various points throughout this journey, they have each stepped in to help me navigate tough situations. Without their patience, support, and guidance, I may never have made it to this point of writing my dissertation. It meant a lot to me to have Lyman as a prethesis committee member and Herman as a reader for my thesis. I can no longer recall when or how I first met Jo Dunkley, but every interaction I have had with her since, however long or short, has left me feeling
uplifted. She has been a sounding board and a source of warmth and encouragement for me, despite our never having worked on any research together. Having Jo on my thesis committee is truly special and not something I take for granted. I must also thank Chris Tully for being on my prethesis committee and Dan Marlow for being on my thesis committee.

The staff truly are the heart and soul of Jadwin Hall, home to the Princeton Physics Department. I knew from the first day I stepped foot in that building that the staff members were special. I met Darryl Johnson during my very first few days as an unseasoned first year, wandering the halls of the A level trying to get my hands on a postage stamp. He was warm and welcoming, and while he did not have a stamp for me, I am forever grateful that he introduced himself to me that day. Too many theorists pass through Jadwin without getting to know the true gem of a human that Darryl is—throughout the past five years, seeing Darryl has never failed to brighten my day. I also met Julio Lopez during the first weeks of first year. Visits to the stockroom (fun in their own right) were made that much more fun by Julio’s presence. Chatting with him has always been a blast, and he has always been eager to help in any way possible—I am honored to have been let in on the secret stash of Kleenex. If Julio and Darryl ever make good on their promise to perform in the department recital, I will do everything in my power to be there. Kate Hare’s office was my safe haven during first year, and I have missed her since she left Princeton; I hope she is doing well. Barbara Mooring has retired and returned numerous times throughout my graduate career; it has been a treat to see her back each time and get the best hugs from her. Kate Brosowsky has been the Graduate Administrator for the majority of my time as a graduate student, and she has been unfailingly patient and positive in every interaction we have had. Kate might just be prouder than anyone else each time a student successfully graduates—we are truly lucky to
have her cheering us on and supporting us along the way. Visiting Toni Sarchi in her office has been a joy every time. I thank her from the bottom of my heart for being patient and kind while I took far too long to learn how to properly put together a Concur report (she probably still fixed many things for me even after I thought I had mastered the process). Regina Savadge has been an integral ally and supporter of the Women in Physics Group; without her help, we would have been very lost. Lisa Scalice is a force to be reckoned with, and while she and I have not crossed paths as much as I would have liked, I know the department is lucky to have her. Angela Lewis has always greeted me with a warm hello when we have run into each other; while she may not have realized it, the simple act of exchanging a smile with her has helped me cope with difficult, stressful days.

It has been thrilling to watch the Women in Physics Group begin to flourish in the past few years, and I know I am leaving it in enthusiastic and capable hands—I am excited to see how the group continues to develop in the years to come. A special shout-out goes to Sara Sussman, Stephanie Kwan, and Gillian Kopp, who stepped up as eager young first years to push the group forward as I began to take a step back from leadership roles. An extra special shout-out goes to Mallika Randeria, my former co-WiP leader, former officemate, confidante and friend. In the year since she left Princeton, I have missed her presence greatly; I look forward to the next chance we get to hang out in person again, whenever that may be. Another extra special shout-out goes to Sonia Zhang, my office neighbor, basically-across-the-street neighbor, friend and mutual catsitter. I have admired her tenacity in graduate school, and will be sad to no longer be living in the same city. Graduate school would have been much lonelier without Mallika and Sonia.

The Women in STEM Leadership Council has given me the opportunity to get involved in broader, university-wide initiatives. Through it, I have met some truly
inspiring women scientists, including Dean Vanessa Gonzalez-Perez. Over the years, I have enjoyed many invigorating conversations with Vanessa, about everything from activism to knitting. I thank her for being an inspiration and also for always including me on emails about knitting events. I would also like to thank Shannon Swilley Greco, science education extraordinaire and force to be reckoned with, for putting up with me as her co-chair of the planning committee for CUWiP 2017 at Princeton. I had a lot of fun and learned a great deal from her through the process. I would have loved to have another opportunity to plan CUWiP with her, but I know that CUWiP 2022 at Princeton is in great hands with Shannon at the helm—with the experience of hosting a resoundingly successful conference under her belt—and a much larger group of graduate students and postdocs already eager to help out compared to the team we had the first time around.

Many of people who have made it possible for me to get to this point have had no affiliation with Princeton University. I am grateful to Mukund Vengalattore, who first gave me the opportunity to do intensive physics research as an undergraduate. Mukund taught me how to think like a researcher, and it was through working in his lab that I learned the foundational skills that later transferred over to the research presented in this thesis. I am indebted to Liam McAllister, whose honors freshman physics class inspired me to become a physics major. Throughout college, I was a frequent visitor of Liam’s office, seeking advice on a variety of life and career topics. Even now, Liam continues to mentor me; I am not sure why he has chosen to make as much time for me as he has over the years, but I am incredibly grateful that he has—without him, there is simply no way I would be where I am now. I thank my dear friend Jack Jiang, whom I first met in Liam’s class, for being a remote comrade throughout this journey. I also thank my old friends Fei and Ni, who have supported me in their own ways through it all. The path to the doctorate involves a long process
of self-discovery. I have come a long way since the day I first set foot on this journey, and for that I owe Dr. Kuzman endless gratitude. She saw strength in me when I could not summon it myself, and never allowed me to give up.

This thesis is dedicated to my family, without whom none of this would have been possible. To Mom and Dad, who raised me to never doubt that I should and could be a scientist, and who taught me to strive to be disciplined, meticulous, and hardworking, but most importantly, happy. They have loved and supported me through all the twists and turns in my journey, and continue to do so as I head on to the next stage of my career. For this I am eternally grateful. To Hannah, who has paved the way for me. Knowing that she is just a phone call away with sisterly advice has brought me comfort during difficult times. To Kyle, who has never failed to welcome me into his home and ask me whether I have found dark matter yet. To Zoe, who isn’t afraid to dream big and love bigger. It has been a joy to watch her grow; I hope I can be a good role model for her as she charts her own course through life. To Daphne, the real physicist. I miss her daily, but know that she has been with me every step of the way. This thesis is as much hers as it is mine.

Finally, I would like to thank the person without whom this thesis would probably not exist, Siddharth Mishra-Sharma. His unwavering support, encouragement, and love have kept me afloat over the past several months. His companionship over the past several years has changed my life for the better. His patience and work ethic have consistently been an inspiration to me. With him by my side, I have learned to believe in myself and gathered the courage to chase my dreams. We have made countless memories together, and I cannot wait to witness how we continue to grow together. I will forever be grateful that I met my very best friend in graduate school. Thank you for everything, Sid.
For Daphne.
I live and work with three basic assumptions:

(1) There is no problem in science that can be solved by a man that cannot be solved by a woman.

(2) Worldwide, half of all brains are in women.

(3) We all need permission to do science, but, for reasons that are deeply ingrained in history, this permission is more often given to men than to women.

—Vera Rubin, “Bright Galaxies, Dark Matters”
Contents

Abstract ................................................................. iii
Acknowledgements ....................................................... iv

1 Introduction ............................................................. 1
    1.1 A Brief History of Dark Matter ................................. 2
        1.1.1 Early Foundation ............................................ 2
        1.1.2 Galaxies and Galaxy Clusters ............................... 5
        1.1.3 Cosmological Imprints ...................................... 11
        1.1.4 Additional Evidence ........................................ 14
    1.2 Dark Matter Zoology .............................................. 15
        1.2.1 Cosmological Constraints ................................... 15
        1.2.2 Thermal WIMP Dark Matter ................................. 17
        1.2.3 The Zoo of Candidates ...................................... 20
    1.3 Thesis Outline .................................................. 24

2 Dark Matter in the Sky: Indirect Detection ......................... 26
    2.1 Photons from Dark Matter Annihilation ....................... 26
    2.2 Template Fitting ................................................ 32
    2.3 Non-Poissonian Template Fitting ............................... 34
    2.4 Search Targets in the Gamma-Ray Sky ....................... 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>Background Mismodeling</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>Search for Dark Matter Annihilation in the Milky Way Halo</td>
<td>50</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>3.2</td>
<td>Analysis Procedure</td>
<td>53</td>
</tr>
<tr>
<td>3.3</td>
<td>Results and Discussion</td>
<td>59</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Dark Matter Annihilation Limit</td>
<td>59</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Galactic Foreground Modeling</td>
<td>61</td>
</tr>
<tr>
<td>3.3.3</td>
<td>The GeV Excess</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Conclusions</td>
<td>66</td>
</tr>
<tr>
<td>3.5</td>
<td>The Region-of-Interest</td>
<td>68</td>
</tr>
<tr>
<td>3.6</td>
<td>Signal Injection and Recovery</td>
<td>70</td>
</tr>
<tr>
<td>3.7</td>
<td>Extended Results</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>Characterizing the Nature of the Unresolved Point Sources in the</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Galactic Center: An Assessment of Systematic Uncertainties</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Foreword: A Closer Look at the Inner Galaxy</td>
<td>83</td>
</tr>
<tr>
<td>4.2</td>
<td>Introduction</td>
<td>88</td>
</tr>
<tr>
<td>4.3</td>
<td>Statistical Methodology</td>
<td>93</td>
</tr>
<tr>
<td>4.3.1</td>
<td>NPTF Procedure</td>
<td>93</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Simulated Data Maps</td>
<td>98</td>
</tr>
<tr>
<td>4.4</td>
<td>Anatomy of a Source-Count Function</td>
<td>100</td>
</tr>
<tr>
<td>4.5</td>
<td>Dark Matter and the GCE</td>
<td>107</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Dark Matter-Only GCE</td>
<td>108</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Fractional Dark Matter Recovery</td>
<td>110</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Signal Injection Tests on Monte Carlo</td>
<td>112</td>
</tr>
<tr>
<td>4.6</td>
<td>Diffuse Mismodeling</td>
<td>114</td>
</tr>
</tbody>
</table>
5.5 Conclusions ............................................................. 201
5.6 Light Profile Fitting Procedure ................................. 206
5.7 Prior Selection and Joint Analysis .............................. 209
  5.7.1 Implementing Narrow Priors .................. 209
  5.7.2 Joint Analysis .................................................. 212

6 Mining Stellar Substructure in the Milky Way 218
  6.1 Introduction ......................................................... 218
  6.2 Machine Learning Primer .................................... 222
  6.3 Clustering in 5D .................................................. 223
  6.4 Machine Learning the Sixth Dimension .............. 232
    6.4.1 Galaxia Mock Gaia Catalog ...................... 232
    6.4.2 The Case of the Helmi Streams ................. 235
    6.4.3 Building a Neural Network ...................... 236
    6.4.4 Optimizing the Neural Network ................. 240
    6.4.5 Predicting the Network Confidence .......... 247
  6.5 Conclusions and Outlook .................................. 251

Bibliography ............................................................ 256
Chapter 1

Introduction

Nearly ninety years since Fritz Zwicky first inferred from kinematic measurements within the Coma Cluster that additional, unobserved massive objects must account for the majority of the mass within the galaxy cluster [1, 2]—roughly marking the birth of the modern study of dark matter (DM)—the precise nature of dark matter remains elusive and continues to be one of the big open questions in particle physics and cosmology. Dark matter is the central theme running throughout this thesis, and this Introduction will cover some of the important building blocks of dark matter research.

This Introduction is organized as follows: Section 1.1 briefly reviews some of the historically significant developments in the study of dark matter; in particular, there will be a focus on presenting some of the empirical evidence for the existence of dark matter. Section 1.2 provides an overview of today’s “dark matter zoo,” introducing thermal WIMPs and highlighting some of the other classes of dark matter particle candidates that are interesting and viable. Section 1.3 outlines the remainder of this thesis.
1.1 A Brief History of Dark Matter

While many people attribute the advent of dark matter research to Zwicky’s work in the 1930s [1, 2], astronomers had been postulating the existence of invisible, massive bodies long before then, as early as the 19th century. Some of the key developments in the history of dark matter research are highlighted here, while a more in-depth review of the history and “prehistory” of dark matter can be found in Ref. [3].

1.1.1 Early Foundation

In the early 20th century, Lord Kelvin sought to estimate the amount of mass within some volume around the sun that would be consistent with the observed velocities of stars [4]. For simplicity, he took this to be the volume enclosed by a spherical shell with radius $3.09 \times 10^{16}$ km (corresponding to “the distance at which a star must be to have parallax one one-thousandth of a second” [4]), and assumed that the total mass within that volume was $10^9 M_\odot$. Using Gauss’s Law, he found that a star on the shell would experience gravitational acceleration $\sim 1.37 \times 10^{-13}$ km/s$^2$. If such a star had started at rest and accelerated (with constant acceleration) over the course of 25 million years, it would end up with a velocity of $10^8$ km/s; in reality, the gravitational acceleration would be a function of the position of the star, and therefore he concluded that “[the velocity of the star] would, in fact, have many chances at being much greater than 108 kilometres per second, and many chances also of being considerably less” [4]. These numbers were comparable to the line-of-sight velocities measured at the time, leading him to conclude:

Thus it is quite possible, perhaps probable, that there may be as much matter as a thousand million suns within the distance corresponding to parallax one one-thousandth of a second ($3.09 \cdot 10^{16}$ kilometres). But it
seems perfectly certain that there cannot be within this distance as much matter as ten thousand million suns; because if there were, we should find much greater velocities of visible stars than observation shows.[1] [...] [There] may be as many as 1000 million stars within the distance $3.09 \cdot 10^{16}$ kilometres; but many of them may be extinct and dark, and nine-tenths of them though not all dark may not be bright enough to be seen by us at their actual distances. [4]

These “dark” stars that Lord Kelvin postulated were merely stars with very low luminosity. While he made no mention of dark matter as we now understand it, it is quite interesting that his simple argument using the dynamics of stars to constrain the amount of mass in the local universe is reminiscent of more sophisticated methods that were later used—and are still employed today—to dynamically estimate the amount of dark matter in galaxies and galaxy clusters. Chapter 5 of this thesis, for example, is focused on dynamical mass modeling in dwarf galaxies.

Some years later, Jacobus Kapteyn took inspiration from the kinetic theory of gases to model the motion of stars in the Milky Way [6]. Kapteyn deduced that there must be rotation in the galaxy within the galactic plane. He also used the measured dispersion of radial velocities to estimate the effective mass of stars, defined as “the total mass within a definite volume, divided by the number of luminous stars” [6]. Kapteyn noted that the effective masses derived using this method were comparable to the effective masses of binary star systems found independently in a paper published two years prior [7]. He went on to note that this could give an estimate for the amount of dark matter—by which he meant non-luminous, gravitationally interacting matter, much like the modern definition of dark matter—and concluded:

1 The term “dark stars” in modern literature refers to a unique class of theoretical stellar objects that are powered not by fusion but instead by the annihilation of Weakly Interacting Massive Particles (WIMPs) [5].
We therefore have the means of estimating the mass of dark matter in the universe. As matters stand at present it appears at once that this mass cannot be excessive. If it were otherwise, the average mass as derived from binary stars would have been very much lower than what has been found for the effective mass. [6]

Not long after, James Jeans revisited Kapteyn’s analysis, relaxing the assumption that Kapteyn had made about isothermal equilibrium, assuming instead that the stellar system was in steady state [8]. Jeans found that the velocities of stars could be explained if and only if the sun was not near the center of the galaxy as Kapteyn had assumed, but rather about 1 kpc from the Galactic Center (though he acknowledged that a precise measurement of the distance from the sun to the Galactic Center would require a dedicated study). Under these assumptions, taken together with the aforementioned binary star measurements, Jeans found that the average mass of a star in the solar neighborhood to be approximately $2.4 M_\odot$, concluding that there were “two dark stars to each bright star” [8].

One more decade later, Kapteyn’s student Jan Oort performed an updated, extensive analysis with several takeaways, one of which was an updated estimate of the local dark matter density [9]. Oort found the total local density to be $0.092 M_\odot$/pc$^3$. Compared to Kapteyn’s results, corresponding to $0.099 M_\odot$/pc$^3$ [6], Oort stated, “this agreement is unexpectedly good; in fact, it seems probable that part of it should be attributed to a chance coincidence, as the velocity and density data used by Kapteyn differ rather widely from those used in the present paper” [9]. By comparison, two other analyses found the value to be $0.143 M_\odot$/pc$^3$ [8] and $0.217 M_\odot$/pc$^3$ [10].\footnote{Ref. [10] has been rather difficult to track down, as it was published in an internal publication of Uppsala University. I have sought to find the correct reference to the best of my ability, using Refs. [9, 11] as guidance, but have not been able to find an archived version of the article itself.} In order to turn this into an estimate of the local dark matter density, Oort extrapolated...
the stellar mass function from measured stars down to low luminosities to obtain an estimate for the total luminous mass in the solar neighborhood, concluding that “the total mass of nebulous or meteoric matter near the sun is less than $0.05 \ [M_\odot/\text{pc}^3]$ or $3 \cdot 10^{-24} \text{g/cm}^3$; it is probably less than the total mass of visible stars, possibly much less” [9].

To summarize, several different studies throughout the 1920s found that dark matter could contribute no more than roughly half of the matter in the solar neighborhood. While several of these studies explicitly used the term “dark matter,” scientists primarily believed that the missing mass would be accounted for by stars much fainter than those which had been measured. This is very different from what scientists mean when they use the term “dark matter” today, which is a proper noun used to refer to the novel type of matter with very specific cosmological properties that makes up the majority of the matter density in our Universe.

1.1.2 Galaxies and Galaxy Clusters

While the early works mentioned in the previous section focused on dynamics in the solar neighborhood, an important step in the transition to the modern way of thinking about dark matter was to study dynamics on galactic scales, outside of the Milky Way. To this day, our understanding of dark matter is most accurate on the largest scales, with much left to be understood and resolved on smaller scales [12, 13].

Coma Cluster

One of the seminal works along this direction was Zwicky’s study in 1933 of unusually large velocity dispersions in the Coma Cluster [1]. Zwicky noticed that within Coma, there were galaxies with differences in apparent velocity of at least 1500 to 2000 km/s. The fact that Coma had much larger velocity dispersion than other galaxy clusters...
had already been observed by Edwin Hubble and Milton Humason [14], whose redshift data Zwicky used. Hubble and Humason had remarked about NGC 4865 (a galaxy in Coma with particularly large velocity), “[its] velocity, however, is the one outstanding discrepancy among some 28 velocities in 8 clusters or groups” [14]. While Hubble and Humason concluded that NGC 4865 must be a field galaxy, Zwicky took a different approach and applied the virial theorem to the data.

Running under the assumption that the system was in steady state, and that the cluster contained 800 galaxies with mass $10^9 M_\odot$ and had a radial extent of one million light years, Zwicky found that in order to describe the data, the total mass in the Coma Cluster would have to be at least 400 times its luminous mass. If this finding were correct, he stated that “we would get the surprising result that dark matter is present in much greater amount than luminous matter” [1]. While it is now an accepted fact that dark matter dominates the matter content of the Universe, Zwicky’s skepticism is understandable as his findings seemed at odds with the previous works of Kapteyn [6], Jeans [8], Lindblad [10], and Oort [9], each of whom had concluded that dark matter and luminous matter accounted for comparable fractions of the total mass in the solar neighborhood. Zwicky also explored the possibility that the system was in fact not virialized, and therefore the galaxies in Coma would gradually disperse over time, reaching velocities around 1000 to 2000 km/s; this scenario, however, “hardly agrees with the facts gained from experience, since the variation of eigenvelocities of isolated nebulae does not exceed 200 km/s” [1].

Zwicky later revisited this calculation, making minor numerical adjustments to his assumptions, and found this time that the mass-to-light ratio of the Coma Cluster would be $\Upsilon = 500$ [2]. He still viewed this result as a problem, contrasting it with the mass-to-light ratio of $\Upsilon = 3$ from Kapteyn’s work from local stellar kinematics [6], calling this discrepancy “so great that a further analysis of the problem is in order” [2].
Among the reasons Zwicky’s work is considered so revolutionary in the study of dark matter is this very discrepancy—while the prior studies had found an $\mathcal{O}(1)$ amount of missing mass, which could easily be explained away by low-luminosity stellar objects that were evading detection, Zwicky was the first to publish findings that suggested there was a vast amount of missing mass in galaxy clusters, which could not be so easily explained by missing stars.

It is worth noting that Zwicky was not the only one who noticed around this time that the dynamics in galaxy clusters implied a vast amount of non-luminous mass. In 1936, Sinclair Smith analyzed data from the Virgo Cluster, and found its dynamical mass to be $10^{14} M_\odot$ [15]. Assuming there were 500 galaxies in Virgo, the average mass for each galaxy was then $2 \times 10^{11} M_\odot$. This was discrepant with what Hubble had found using the mass-luminosity relationship, which was that the average mass of a galaxy was around $10^9 M_\odot$ [16]. Smith posited that these two numbers may not actually be contradictory of each other, and that it was “possible that both figures are correct and that the difference represents a great mass of internebular material within the cluster” [15]. In present-day language, this “internebular material” would be the dark matter.

**Rotation Curves**

In the decades following Zwicky’s Coma Cluster results, the astrophysics community was divided in its opinion of the solution to the missing mass problem, or even if there was a problem to be solved at all. The measurement of galactic rotation curves—the circular velocities of stars or gas in a galaxy as a function of distance from the center of the galaxy—provided an important piece of evidence that led to a shift in the mentality when it came to the existence and the importance of understanding dark matter. While there were several important precursors, I will focus here on the
pioneering work of Vera Rubin and Kent Ford in 1970 [17]. Ford had built a state-of-the-art spectrometer, and Rubin had been fascinated with directly measuring rotation in galaxies. Rubin and Ford set out to measure the rotational velocities of stars and gas in M31, expecting that objects in the outer regions of the galaxy would orbit with lower velocities since most of the visible mass in the galaxy was concentrated near the center, and therefore the circular velocity \( v_c(r) = \sqrt{GM(r)/r} \) should drop as \( r^{-1/2} \) towards the outskirts of the galaxy. To their surprise, as they measured farther and farther out from the center of M31, they never saw the decrease in velocity that

Figure 1.1: From Ref. [17]. The rotation curve of M31. The open data points are from measuring the [NII] emission line in the innermost regions of M31, while the solid circles(squares) are from measurements of OB associations in the NE(SW) quadrants of the galaxy. The solid curve is constructed from a fifth-order polynomial for \( R < 1.6 \) kpc and a fourth-order polynomial for \( R > 2.6 \) kpc (obtained from a least-squares fit required to remain flat near 24 kpc), joined smoothly in the intermediate region. The dashed curve is chosen arbitrarily such that the inferred density is positive everywhere (which is not true of the black curve for the chosen galaxy model in Ref. [17]).
they had expected. Ford was even quoted as saying, “We kept going farther and farther out and had some disappointment that we never saw anything,” while Rubin said, “I do remember my puzzling at the end of the first couple of nights that the spectra were all so straight” [19]. Figure 1.1 shows this puzzling rotation curve that they observed: out to a distance of over 20 kpc, the measured velocities do not drop significantly. While Rubin and Ford presented this observation, they did not seek to explain the nature of the discrepancy from their expectations. They proceeded to use their spectrograph to measure galaxy after galaxy, finding that the flattened rotation curve was ubiquitous across the many galaxies they observed [18, 20] (Fig. 1.2 shows some examples of the rotation curves they measured), and several other groups found similar results over the course of several years [21, 22, 23, 24, 25, 26, 27].
In fact, the optical results from Rubin and Ford did not extend as far out from the centers of galaxies as radio measurements, and in particular it was later shown that the baryonic disk of M31 could explain the data from Ref. [17] (the scale length of M31’s disk was recently shown to be $\sim 5.3$ kpc [30]). Figure 1.3 shows that the Rubin and Ford measurements are consistent with an exponential disk model for M31, whereas later measurements of 21-cm HI line emission showed that farther out from the center of the galaxy, there were inconsistencies between the data and the disk model [28, 29].
In 1974, the prevalent flat rotation curves caught the attention of two separate groups of theorists—Jaan Einasto, Ants Kaasik, and Enn Saar; and Jerry Ostriker, Jim Peebles, and Amos Yahil—each of which published a groundbreaking paper arguing convincingly that vast amounts of mass were missing from galaxies and galaxy clusters [31, 32]. In particular, Ostriker, Peebles, and Yahil argued that galaxies were in fact at least a factor of 10 more massive than their luminous mass suggested; correspondingly, the mean density of the Universe, back then calculated from the observed number density and average mass of galaxies, would be underestimated by roughly the same amount. They also inferred from the flatness of the rotation curves that in local spiral galaxies, the total mass must approximately follow the form $M(r) \propto r$ within the range $20 \text{kpc} \lesssim r \lesssim 500 \text{kpc}$, “similar to that in the outer parts of isothermal gas spheres” [32]. Compared to the case in the 1930s, the missing mass problem was taken much more seriously within the astrophysics community in the era of flat rotation curves: in 1979, Sandra Faber and John Gallagher wrote an extensive review on the state of the field [33].

1.1.3 Cosmological Imprints

On even larger scales, cosmology has provided some of the most robust evidence for the existence of dark matter. Since the discovery of the cosmic microwave background (CMB) in 1965 [34, 35], increasingly sophisticated cosmological measurements have served as high-precision probes of dark matter properties. Another key development was the advent of numerical simulations in the late 1970s–early 1980s [36, 37, 38, 39, 40], which allowed physicists to begin exploring the cosmological signatures of different types of dark matter. To this day, cosmological observations and numerical simulations continue to provide some of the highest-precision insight into the nature
of dark matter, and both areas will continue to see drastic improvements in the next \(~10\) years.

Even in the earliest \textit{N}-body simulations, it was clear that the temperature of the dark matter particles would have implications on the formation of large-scale structure: “hot” (relativistic) dark matter particles would form structure in a “top-down” fashion—large structures would form first and later fragment into smaller structures—whereas “cold” (nonrelativistic) dark matter particles would form structure in a “bottom-up” way—overdensities on small scales would seed the formation of hierarchically larger structures. Simulators could compare their results with the then-new CfA redshift survey of galaxies in the local Universe \cite{41}. For example, Simon White, Carlos Frenk, and Marc Davis simulated a cosmology dominated by neutrinos, and found that the predicted structure did not match the data, stating “[...] the discrepancy between the large coherence length of neutrino-dominated universes and the small scale of observed galaxy clustering makes it appear unlikely that neutrinos provide the missing mass” \cite{40}. Increasingly, the community realized that numerical simulations with cold dark matter matched the galaxy clustering data very well \cite{42, 43}—far better than hot dark matter did. More generically, it is known today that when constructing a model of dark matter, it is important to check against constraints from the matter power spectrum—the spectrum of the density fluctuations of matter in the Universe. This is discussed more in Section 1.2.1.

As previously alluded to, the discovery and subsequent increasingly precise measurements of the CMB have provided evidence for the existence as well as insight into the abundance of dark matter in the Universe. In the early Universe, photons were tightly coupled to baryons in a hot, dense plasma. The radiation pressure of the photons counteracted the pull on the baryons from gravitational potential wells; the compression and expansion of the plasma resulted in acoustic oscillations and temper-
Figure 1.4: The temperature angular power spectrum from Planck 2018 data [44]. The red line is the best-fit assuming a baseline $\Lambda$CDM cosmology [45]; in particular, the baryon density is set to $\Omega_b h^2 = 0.0223828$ and the CDM density is set to $\Omega_c h^2 = 0.1201075$. The gray solid(dashed) line shows the resulting $TT$ power spectrum for a model with $\Omega_c h^2 = 110\%(90\%)$ of the best-fit value—with $\Omega_b h^2$ appropriately adjusted to maintain the same total matter density—and all other parameters set to their best-fit values. All of the calculated power spectra were computed using the publicly available CAMB code [46].

ature fluctuations. When photons decoupled from baryons and began free-streaming at the time of recombination, they carried the information of these fluctuations with them—the fluctuations are imprinted onto the temperature distribution of the CMB we observe today.

The most relevant observable for the purpose of understanding the contribution of dark matter is the temperature angular power spectrum ($TT$ power spectrum) of the CMB. The power spectrum is a series of peaks and troughs in Fourier space (see Fig. 1.4), and the locations and heights of the peaks depend sensitively on the contents of the Universe. The location of the first peak (and subsequent peaks) is sensitive to the spatial curvature of the Universe, and therefore the total energy contents
of the Universe—the current measurements show that the Universe is very close to perfectly flat (i.e., close to critical energy density). The location of the second peak encodes information about the baryon abundance, while the amplitudes of odd peaks relative to even peaks measures the relative abundance of non-baryonic dark matter to baryonic matter\(^3\). The most recent measurements of the \(TT\) power spectrum from Planck have yielded baryon density \(\Omega_b h^2 = 0.0224 \pm 0.0001\) and cold dark matter density \(\Omega_c h^2 = 0.120 \pm 0.001\) \([45]\).

1.1.4 Additional Evidence

While the historical overview covered in Sections 1.1.1–1.1.3 touched upon a number of the important pieces of evidence for the existence of dark matter, it was by no means a comprehensive list. Many other observations have provided evidence as well, and the fact that so many different and complementary measurements have pointed in the same direction—that dark matter must not only exist, but be far more abundant than ordinary matter within the Standard Model—has broadly convinced the astrophysics and particle physics communities to accept that we do indeed live in a Universe filled with dark matter.\(^4\) Some important examples of the additional sources of evidence come from gravitational lensing. Ref. \([52]\) provides a review of evidence for dark matter from lensing searches which is slightly outdated but nonetheless instructive (in particular, the Dark Energy Survey Collaboration has published newer results on weak lensing \([53]\)). Notably (and famously), all of the evidence for dark matter to-date has been gravitational in nature, though there are many creative ongoing and proposed

\(^3\) The statements made here have been heuristic in nature; the interested reader can find a series of excellent tutorials providing more intuition on the physics behind the CMB power spectrum at Ref. \([47]\) or simulate the effects of altering the cosmological makeup of the Universe on the power spectrum at Ref. \([48]\).

\(^4\) There are, of course, exceptions. Notably, Modified Newtonian Dynamics (MOND) was first proposed by Mordehai Milgrom in a series of three papers \([49, 50, 51]\), and continues to be regarded as an alternative theory to dark matter. Proponents of MOND nevertheless acknowledge the need for some modifications to our standard set of theories in order to explain the observational data.
experiments seeking to directly detect dark matter, not through gravitational effects; if a direct detection of dark matter were to be made in the upcoming years, the field of dark matter physics could shift dramatically (see, for example, Ref. [54] and references within).

1.2 Dark Matter Zoology

This section is not meant to be an extensive review of all of the viable models of DM. Rather, the goal of the section is to illustrate the wide array of possibilities in constructing a theory of dark matter. There is viable parameter space in many aspects of the properties of DM, for example: whether DM is a fundamental particle or not; whether one species makes up the entirety of DM or there are multiple DM species; what the mass scale of the DM is; and how it interacts with itself, other potential dark particles, and Standard Model (SM) particles.\(^5\)

1.2.1 Cosmological Constraints

When constructing a DM model, it is important to consider the existing constraints on the theory parameter space. In particular, there are stringent cosmological constraints that must be satisfied, coming from measurements of the CMB power spectrum and the matter power spectrum \(P(k)\). On the theory front, different DM models may lead to different predictions for these observables—famously, theories of warm dark matter (WDM) lead to suppression of structure on small scales, which translates into a “cutoff in the (matter) power spectrum” at large wavenumbers. This suppression is shown by the dashed blue line in Fig. 1.5, along with the canonical cold dark matter (CDM) result in solid blue. Another model, atomic dark matter (ADM), is shown

\(^5\) Much of this section is inspired by Tongyan Lin’s lectures at TASI 2018; her excellent lecture notes go much more in depth into dark matter models and direct detection [54].
Figure 1.5: From Ref. [55]. Prediction of the dimensionless matter power spectrum, \( \Delta^2(k) \equiv 4\pi(k/2\pi)^3P(k) \), for three separate DM scenarios: CDM (solid blue; transfer function from Ref. [56]), WDM with \( m_{\text{WDM}} = 8 \text{ keV} \) (dashed blue; model from Ref. [57]), and ADM (dotted blue; model from Ref. [58]). The gray shaded region roughly denotes the limit beyond which smaller scales cannot currently be probed.

in dotted blue—in this model, there are “dark acoustic oscillations” (analogous to baryon acoustic oscillations) in addition to damping at small scales. On the data front, Fig. 1.6 shows that current measurements of the linear matter power spectrum—CMB measurements at large scales, galaxy clustering at intermediate scales, and Lyman-\( \alpha \) forest and cosmic shear measurements at small scales—are in remarkable agreement with the predictions from the canonical \( \Lambda \)CDM paradigm. Notably, neither of the alternative models depicted in Fig. 1.5 appreciably alters the linear matter power spectrum on scales that can currently be probed (hence, they can still be viable models), but generically speaking, deviations from the \( \Lambda \)CDM paradigm may have more drastic effects that can be ruled out using current cosmological measurements.
Figure 1.6: From Ref. [59]. (Top) The 3D linear matter power spectrum at $z = 0$, probed by Planck CMB power spectra [44], DES cosmic shear data [60], SDSS data measuring the clustering of luminous red galaxies [61], and Lyman-$\alpha$ forest measurements from the extended Baryon Oscillation Spectroscopic Survey (eBOSS) of SDSS [62]. The solid black line shows the Planck 2018 best-fit linear matter power spectrum [45], while the dotted line shows the inclusion of non-linear effects. (Bottom) The deviations of the data from the Planck 2018 best-fit spectrum.

1.2.2 Thermal WIMP Dark Matter

A canonical example of a CDM candidate is the thermal Weakly Interacting Massive Particle (WIMP). While the work presented in this thesis is largely model-agnostic, whenever model-dependent statements are made, the underlying assumption is that dark matter is entirely made up of thermal WIMPs. “Thermal” refers to the fact that the DM is thermally produced in the early Universe: initially, the DM (denoted as $\chi$) was in thermal equilibrium with the Standard Model, and therefore the for-
ward and backward reactions of $\chi \chi \leftrightarrow \text{SM SM}$ occurred at the same rate. As the Universe expanded and cooled down, it became increasingly rare that two DM particles would annihilate. Once the annihilation rate $\Gamma_{\text{ann.}} = n_\chi \langle \sigma v \rangle_{\text{ann.}}$ (where $n_\chi$ is the number density of DM particles and $\langle \sigma v \rangle_{\text{ann.}}$ is the velocity-averaged annihilation cross-section) became on the order of the Hubble rate $H$, i.e. $\Gamma_{\text{ann.}} \sim H$, the forward reaction became extremely Boltzmann-suppressed, and the DM density became “frozen”—this is known as the thermal freeze-out mechanism, and is the procedure that sets the relic abundance of DM measured today for a thermal candidate.

A particle with weak-scale annihilation cross-section $\langle \sigma v \rangle \sim 3 \times 10^{-26} \text{cm}^3/\text{s}$ and weak-scale mass $m_\chi \sim 10 \text{GeV}–10 \text{TeV}$ gives the correct DM relic abundance that is observed today. In practice, many other candidates—whether thermal or nonthermal—can give the correct relic abundance, but WIMPs have been the dominant DM paradigm for many decades, in part because theoretically-motivated frameworks such as supersymmetry naturally give rise to WIMP DM candidates. Supersymmetry posits that for every SM particle, there exists another particle (referred to as the “superpartner” of the particle), where for bosons the superpartner is a fermion and vice versa [63, 64]. However, because superpartners have yet to be discovered at the Large Hadron Collider (LHC) and WIMPs also have yet to be conclusively detected, there has been a shift in paradigm in recent years away from the WIMP “lamppost”.

There are three generic search strategies for DM candidates with $2 \to 2$ interactions such as WIMPs, schematically represented by the three colored arrows in Fig. 1.7:

1. **Indirect Detection** Although DM annihilation $\chi \chi \to \text{SM SM}$ is heavily suppressed today, it can still occur in regions with high DM density. The goal of indirect detection is to detect the stable SM byproducts of this process, such as
photons or leptons. Chapter 2 covers more details on indirect detection, which is the basis of the work presented in Chapters 3–4.

2. **Direct Detection** Direct detection relies on detecting the effects of DM particles scattering off of SM particles, $\chi \text{SM} \rightarrow \chi \text{SM}$, such as nuclei or electrons. The constraints on DM parameters derived from null results in direct detection searches depend sensitively on the assumptions about the local density and velocity distribution of DM in the Milky Way. The work presented in Chapter 6 seeks to identify stellar substructure (and, in turn, DM substructure) in the Milky Way, which can substantially change our assumptions about local the DM density and velocity distribution, and therefore the constraints from direct detection.

3. **Collider Searches** The search strategy with perhaps the longest history has been to produce DM particles by colliding SM particles, $\text{SM SM} \rightarrow \chi \chi$, and to detect the signatures of DM particles having been created (such as missing energy from the final states). Historically, the hope and belief of many particle physicists was that the LHC would discover supersymmetry very quickly, and thereby discover the DM particle or evidence of what it must be. Thus far, however, the LHC has not found conclusive evidence for DM, inspiring physicists
to come up with many other avenues for detecting its existence through other non-gravitational means.

1.2.3 The Zoo of Candidates

Stepping away from WIMPs, there is still a wide open field of possibilities for dark matter candidates, with particle masses spanning nearly 90 orders of magnitude. Fig. 1.8 symbolically sketches out the range of possible DM masses. On the lower end, constraints on the DM mass come from phase-space density arguments: within a dark matter halo, the local dark matter density can be written as

$$n(x) = \int \frac{d^3p}{(2\pi)^3} f(x,p).$$

(1.1)

For fermionic DM, this quantity is bounded from above due to the Pauli exclusion principle:

$$n_{\text{fermion}}(x) \lesssim \frac{g}{(2\pi)^3} \frac{4\pi}{3} p_{\text{max}}^3,$$

(1.2)

where $g$ is the degeneracy factor (for instance, accounting for different spin states) and $p_{\text{max}}$ is the maximum possible momentum. This translates into a lower bound on the DM mass:

$$\left(\frac{g}{(2\pi)^3} \frac{4\pi}{3} v_{\text{max}}^3\right)^{-1} \lesssim m_{\text{fermion}},$$

(1.3)

where $v_{\text{max}}$ is the maximum velocity and $m_{\text{fermion}}$ is the mass of the fermionic DM particle. Kinematic measurements of galaxies can be used to derive quantitative limits. There are several versions of this bound, including the famous Tremaine-Gunn bound [65]; more recent analyses have used measurements of the Milky Way dwarf spheroidal galaxies (dSphs) to constrain the mass of fermionic DM to be $m_{\text{fermion}} \gtrsim \mathcal{O}(\text{keV})$ [66, 67]. Therefore, DM candidates with mass below $\sim \text{keV}$ must be bosonic,
and are collectively referred to as “ultralight scalar (bosonic) dark matter.” The lower mass bound on the ultralight candidates is \( \sim 10^{-22} \text{ eV} \). At such low masses, the DM behaves as a coherent field—this can be seen by estimating the typical number of particles within a spherical volume defined by the de Broglie wavelength:

\[
\lambda_{\text{dB}} = \frac{h}{m_{\text{DM}} v} \approx 1.2 \text{kpc} \left( \frac{10^{-22} \text{ eV}}{m_{\text{DM}}} \right) \left( \frac{100 \text{ km/s}}{v} \right),
\]

\[
N = \frac{4\pi}{3} \frac{\rho_{\text{DM}}}{m_{\text{DM}}} \lambda_{\text{dB}}^3 \approx 10^{96} \left( \frac{\rho_{\text{DM}}}{0.4 \text{ GeV/cm}^3} \right) \left( \frac{10^{-22} \text{ eV}}{m_{\text{DM}}} \right)^4 \left( \frac{100 \text{ km/s}}{v} \right)^3,
\]

where \( h \) is Planck’s constant, \( m_{\text{DM}} \) and \( v \) are the mass and typical velocity of the DM particle, respectively, and \( \rho_{\text{DM}} = 0.4 \text{ GeV/cm}^3 \) is the local DM density in the solar neighborhood \([68, 69]\). With such a large number of particles within a de Broglie wavelength, we can indeed safely treat the collection of DM particles as a coherent field. A limit on the lower mass of ultralight scalar DM can be placed using the uncertainty principle to argue that halos smaller than the de Broglie wavelength (more accurately, the Jeans scale \([70]\)) would not be stable; the Milky Way dwarf galaxies place this limit around \( m_{\text{scalar}} \gtrsim 10^{-22} \text{ eV} \) \([71, 72]\). DM particles with masses around this lightest bound are known as “fuzzy dark matter” \([73, 74]\) (discussed more below).

On the other end of the mass scale, the (rough) upper limit on the mass of thermally produced DM is set by the partial wave unitarity bound, i.e. the annihilation cross-section cannot be arbitrarily large—this can be translated into an upper bound on the mass, \( m_{\text{DM}} \lesssim 100 \text{ TeV} \) \([75]\).

Roughly in order of increasing mass, some of the important non-WIMP classes of DM candidates are:\(^6\)

\(^6\) As with the rest of this introduction, this is not meant to be an exhaustive list; rather, I hope to demonstrate the diversity of dark matter models by highlighting some of the classes of candidates that are particularly historically significant.
Figure 1.8: The broad range of masses that a DM candidate can have, with fuzzy dark matter (FDM) at the lightest end and primordial black holes (PBHs) at the heaviest end. The masses quoted here and in the text are not exact, but rather serve as rough guidelines; they are derived under specific assumptions, and should be thought of as approximately accurate to order-of-magnitude levels. The colored bars and text indicate broad classes of DM candidates, while the gray text indicates specific models which fall under certain broad classes.

- **Fuzzy dark matter (FDM)** As mentioned previously, candidates with masses around $m_{\text{DM}} \sim 10^{-22}\,\text{eV}$ are known as fuzzy dark matter [73, 74]. FDM was initially proposed as a solution to the small-scale problems in cold dark matter [13, 76]. FDM leads to some unique astrophysical phenomenology, such as the formation of a soliton core in the central regions of DM halos. Additionally, as the growth of perturbations smaller than the Jeans scale is suppressed, there is a suppression of structure on small scales, and measurements of the matter power spectrum can be used to constrain the FDM mass. Constraints from Lyman-$\alpha$ forest measurements are roughly comparable with the $m_{\text{DM}} \gtrsim 10^{-22}\,\text{eV}$ derived from dwarf galaxies [77].

- **Axions and axion-like particles (ALPs)** The QCD axion, typically in the $10^{-4}$–$10^{-6}\,\text{eV}$ mass range, is one of the simplest solutions to the Strong CP problem [78, 79, 80, 81]. The effective field theory of the QCD axion involves the addition of one new particle, the axion $a$, and its coupling $f_a$:

$$\mathcal{L} \supset \left( \frac{a}{f_a} + \theta \right) \frac{1}{32\pi^2} G G.$$  

(1.6)
A nice quality of the QCD axion is that it can also be dark matter [82], thus solving two problems of the Standard Model at once. Axion-like particles are DM candidates that are inspired by the QCD axion, but do not solve the Strong CP problem. In recent years, there has been a lot of effort and excitement towards detecting axions and ALPs via novel experimental techniques [83, 84], making axion/ALP dark matter searches a very active area of ongoing research.

- **Sterile neutrinos** Sterile neutrinos in the $\sim$ keV mass range are an important example of a warm dark matter (WDM) candidate. These are right-handed neutrinos which mix with the Standard Model neutrinos, but otherwise do not have interactions with the Standard Model (hence “sterile”) [85]. Sterile neutrinos, like the QCD axion, solve two problems at once—they can explain the lightness of the SM neutrino masses while also serving as the dark matter particle in the Universe. A potential detection signature of sterile neutrino DM is the production of a photon with energy $m_{DM}/2$ through radiative decay of the sterile neutrino. This would lead to an X-ray line signal in DM-dense regions of the sky. In 2014, two separate groups reported the detection of excess emission with energy $\sim 3.5$ keV in some galaxies and galaxy clusters [86, 87]. This received a lot of attention from the DM community, and many papers were written on interpreting the signal as originating from DM decay; however, there are potential systematics that could be causing the signal as well. Section 3.2 of Ref. [88] reviews the status of the 3.5 keV line.

- **Dark sectors** In the intermediate mass range $m_{DM} \lesssim 1$ GeV, a thermal DM particle with weak-scale couplings to the Standard Model would be overabundant. This is related to the Lee-Weinberg bound on the mass of thermal DM with weak-scale interactions [89]. This motivates the introduction of new mediators, which must be sufficiently dark to have not been detected thus far. Such
a “dark sector” with more than one dark particle (one or more of which constitutes the DM) can be as minimal as introducing one additional particle or as complex as adding many dark copies of the SM [90, 91, 92, 93]. One example of an actively studied new mediator is the dark photon, a new $U(1)$ gauge boson (typically taken to be massive) which kinetically mixes with the SM photon. Dark photons—and dark sectors in general—produce rich phenomenology and motivate a wide range of experimental searches [94, 95].

- **Superheavy and composite dark matter** Up to around the Planck mass $M_{\text{pl}}$, there are still DM candidates that are fundamental particles. These are known as “WIMPzillas” [96, 97]. Above $M_{\text{pl}}$, DM candidates that are composite objects include bound states of dark particles such as dark quark nuggets [98]. In general, at such high masses, DM models become much more sparse.

- **Massive compact halo objects (MACHOs)** At the very high end of the mass scale, the DM candidates are themselves massive astrophysical objects. Early proposals of MACHO candidates included stellar objects such as brown dwarfs and Jupiters [99, 100]. Recently, much more focus has been on primordial black holes (PBH) as DM candidates. This is also an active area of ongoing research, with many different probes seeking to detect or constrain PBH DM (see Refs. [101, 102] for more details); in general, there are stringent constraints that make it very difficult for PBHs to account for all of the DM.

1.3 Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 covers indirect detection, introducing two of the key methods used in indirect detection studies, summarizing some of the important classes of search targets in the indirect detection
literature, and describing the effects that background mismodeling can have on interpreting the results from indirect detection analyses. Chapter 3 presents an indirect detection study of the Milky Way dark matter halo at high Galactic latitudes, which resulted in the strongest limits on dark matter with mass below $\sim 70$ GeV annihilating into $b$ quarks at the time of publication. Chapter 4 presents a systematic study on simulated data in the Inner Galaxy, which demonstrated that the fundamental degeneracy in photon statistics between emission from ultrafaint point sources and diffuse emission, combined with the effect of background mismodeling, could lead to biased interpretations of the contents of the data. Chapter 5 presents a systematic study seeking to understand the limitations on reconstructing the DM distribution in dwarf galaxies, and to assess which approaches to future observations would be most impactful for the purposes of better constraining the DM density profiles in these systems. Finally, Chapter 6 presents ongoing work that seeks to develop methods for systematically and robustly identifying stellar substructure in the Milky Way, with the aid of machine learning methods.
Chapter 2

Dark Matter in the Sky: Indirect Detection

This chapter will focus on indirect detection of annihilating thermal WIMP dark matter, laying the foundation for the two specific indirect detection studies presented in Chapters 3–4.

2.1 Photons from Dark Matter Annihilation

Indirect detection relies on detecting Standard Model particles that result from dark matter annihilation.\(^1\) Processes that produce signals can include annihilation directly into photon pairs, which gives rise to spectral line signals; annihilation into other SM states that produce photons through secondary mechanisms, which results in a broad spectrum signal; or annihilation into charged particles which are subsequently detected as cosmic rays. Within the context of this thesis, indirect detection

\(^1\) The focus here will be annihilation, but the same principles can be applied to searching for signatures of dark matter decay, with slight modifications.
Figure 2.1: An all-sky gamma-ray map measured by Fermi-LAT [103], in the energy range $E_\gamma = 2\text{–}20\text{ GeV}$, visualized using a Mollweide projection. The Galactic Center is at the center of the map, and the Galactic disk runs horizontally through the middle of the map. These regions emit high intensities of gamma rays, due to astrophysical processes and potentially due to dark matter annihilation.

will be focused on the second case: DM annihilating into SM particles, which subsequently produce photons. For WIMP-scale masses of $\sim 10\text{ GeV–}10\text{ TeV}$, the resulting photons are in energy range that is probed by the Fermi Large Area Telescope (Fermi-LAT) [103], a gamma-ray satellite telescope. On the higher end of this mass range ($m_{\text{DM}} \gtrsim 100\text{ GeV}$), ground-based gamma-ray telescopes such as HAWC [104], H.E.S.S [105], MAGIC [106], VERITAS [107], and CTA [108] typically provide higher sensitivity due to their larger effective areas, complementing the parameter space covered by Fermi-LAT. Figure 2.1 shows an example all-sky gamma-ray map measured by Fermi-LAT.

In searching for a gamma-ray signal from DM annihilation, it is important to know what the expected signal strength would be for a given point in the DM parameter space. In the WIMP scenario, we assume that there is a single species of DM particles, all with the same mass $m_\chi$ and interaction strength. The expected photon flux from these WIMPs annihilating into SM quarks, leptons, and gauge bosons can be
calculated as follows: suppose there is a small volume in the sky \(dV\) in which the DM density is constant, \(\rho_{\text{DM}}(dV) = \rho_{\text{DM}}\), and the corresponding number density of DM particles is \(n_{\text{DM}} = \rho_{\text{DM}}/m_\chi\). We can view a single DM particle within this volume as a “target” with effective area \(\sigma\). This target interacts with an incoming flux of DM particles, which we can initially take to have the same relative velocity \(v\) to the target particle. The incoming flux (number/(time\-area)) of DM particles would therefore be \(n_{\text{DM}}v\), and the corresponding incident rate per unit time would be \(n_{\text{DM}}\langle \sigma v \rangle\). The angular brackets denote averaging over the velocity distribution of DM particles, because in practice, the incoming particles do not have the same velocity, and the cross-sectional area \(\sigma\) can also be dependent on the relative velocity.

Accounting for all pairs of DM particles gives us the annihilation rate within the volume,

\[
\frac{dN_{\text{ann}}}{dt} = \frac{N - 1}{2} n_{\text{DM}} \langle \sigma v \rangle \approx \frac{N}{2} n_{\text{DM}} \langle \sigma v \rangle,
\]

where \(N\) is the total number of DM particles within the volume and the approximation holds for large \(N\). Lastly, because we had arbitrarily chosen the small volume \(dV\), we ultimately want to take the rate per unit volume and integrate it over some relevant region of space, for example over the extent of the Milky Way DM halo. The rate per unit volume is given by

\[
\frac{dN_{\text{ann}}}{dV dt} = \frac{\langle \sigma v \rangle}{2} \frac{N}{dV} n_{\text{DM}} = \frac{\langle \sigma v \rangle}{2} n_{\text{DM}}^2 = \frac{\langle \sigma v \rangle}{2 m_\chi^2} \rho_{\text{DM}}^2.
\]

Thus far, I have not discussed the particle physics specifics, which inform the rate at which signal photons are produced through this annihilation process. For simplicity, we can consider a single annihilation channel; the results can be easily generalized to multiple channels by summing over the relevant channels. One common benchmark scenario in the indirect detection literature is annihilation into \(b\) quarks,
\(\chi\chi \rightarrow b\bar{b}\). The quarks hadronize, and the resulting hadrons decay to produce copious neutral pions, which produce photons through \(\pi^0 \rightarrow \gamma\gamma\) with a \(\sim 99\%\) branching ratio. If we denote the spectrum of photons produced through each annihilation by \(dN_\gamma/dE_\gamma\), \(^{2}\) then the rate per volume of producing photons with energy \(E_\gamma\), \(^{3}\) becomes

\[
\left(\frac{dN_\gamma}{dE_\gamma dV dt}\right)_{\text{prod.}} = \frac{\langle \sigma v \rangle dN_\gamma}{2m_\chi^2 dE_\gamma} \rho_{\text{DM}}^2. \tag{2.3}
\]

In practice, when searching in data for signatures of photons produced from DM annihilation, we are interested in the expected rate of detecting the photons, not the rate at which they are produced. After the photons are produced in a certain region of space (more likely to occur in regions with high DM density, e.g. in the central regions of galaxies), they free stream to the detector. If the source of the photons is a distance \(\ell\) away from the detector, the signal is spread out isotropically over the spherical surface area \(4\pi\ell^2\). If the effective differential area of the detector is given by \(dA\), \(^{4}\) then only \(dA/4\pi\ell^2\) of the produced photons would be detected. Thus, the rate per volume of detecting photons is given by

\[
\left(\frac{dN_\gamma}{dE_\gamma dV dt}\right)_{\text{det.}} = \left(\frac{dN_\gamma}{dE_\gamma dV dt}\right)_{\text{prod.}} \times \frac{dA}{4\pi\ell^2} = \frac{\langle \sigma v \rangle dN_\gamma}{8\pi m_\chi^2 \ell^2 dE_\gamma} \rho_{\text{DM}}^2 dA. \tag{2.4}
\]

\(^{2}\) These spectra can be calculated with the help of parton shower Monte Carlo event generators, but have conveniently been calculated and tabulated in \text{PPPC4DMID} \([109, 110]\).

\(^{3}\) More precisely, photons with energy in the range \([E_\gamma, E_\gamma + dE_\gamma]\).

\(^{4}\) The effective area accounts for not only the physical area of the detector, but also the efficiency of the detector. For example, some information about the effective area of \text{Fermi-LAT} can be found at \url{https://fermi.gsfc.nasa.gov/ssc/data/p7rep/analysis/documentation/Cicerone/Cicerone_LAT_IRFs/IRF_EA.html}.
If we gather the terms that do not involve the particle physics or astrophysics and define the incident differential photon flux on the detector as

\[ d\Phi_\gamma \equiv \frac{dN_\gamma}{dA \, dt}, \tag{2.5} \]

then we arrive at the differential flux at the detector per volume (from this point forward we will only be discussing the expected detected signals, so I will omit the "det." subscript):\(^5\)

\[ \frac{d\Phi_\gamma}{dE_\gamma \, dV} = \frac{\langle \sigma v \rangle \, dN_\gamma}{8\pi m_\chi^2 \ell^2 \, dE_\gamma \, \rho_{\text{DM}}^2}. \tag{2.6} \]

Finally, we want to integrate over the observed volume to get the total expected signal strength. This involves integrating over the solid angle \(\Delta \Omega\) and the line-of-sight (l.o.s.) distance \(\ell\). It is often convenient to choose a spherical coordinate system centered on the Earth, such that the volume element is parametrized as \(dV = \ell^2 d\ell d\Omega\). The only position-dependent term when integrating Eq. 2.6 over the full volume is \(\rho_{\text{DM}} = \rho_{\text{DM}}(\ell, \Omega)\). The total observed gamma-ray flux in the specified energy range is thus given by

\[ \frac{d\Phi_\gamma}{dE_\gamma} = \frac{\langle \sigma v \rangle \, dN_\gamma}{8\pi m_\chi^2 \, dE_\gamma} \int_{\text{l.o.s.}} \int_{\Delta \Omega} d\ell \, d\Omega \, \rho_{\text{DM}}^2(\ell, \Omega). \tag{2.7} \]

Eq. 2.7 neatly factorizes into two different terms: the terms outside of the integrals are dependent only on the particle physics, so we can define the "particle physics factor" as

\[ \frac{d\Phi_{\gamma \text{PP}}}{dE_\gamma} \equiv \frac{\langle \sigma v \rangle \, dN_\gamma}{8\pi m_\chi^2 \, dE_\gamma}. \tag{2.8} \]

\(^5\) It is important to note that this derivation has been for a Majorana DM candidate. If the DM candidate were a Dirac fermion, the right-hand side Eq. 2.6 would need to be multiplied by an additional factor of \(\frac{1}{2}\) to account for the fact that on average, only half of the particles in the volume are the DM antiparticles.
The remaining term is dependent only on the astrophysical distribution of dark matter, and is known as the astrophysical “J-factor”:

$$J \equiv \int_{\text{l.o.s.}} dl \int_{\Delta \Omega} d\Omega \rho_{\text{DM}}^2(\ell, \Omega).$$  \hspace{1cm} (2.9)

The $J$-factor quantifies the amount of dark matter in a certain region in the sky. With the particle physics and detector specifics held fixed, the magnitude of the $J$-factor determines the expected signal strength from different search targets in the sky. The brightest source of expected gamma-ray signals due to DM annihilation is the central region of the Milky Way DM halo. The reason for this is two-fold: firstly, within the CDM paradigm that we are seeking to probe, the density of DM is highest at the center of the halo; secondly, the Milky Way halo is the closest signal source to us, so more of the emitted photons can be detected within the instrumental area.
compared to the case of a source that is far away. A $J$-factor map for the Milky Way halo is shown in Figure 2.2.

While the $J$-factor may be used to identify promising search targets in the sky based on expected signal brightness, the strength of the signal is not the only factor to consider—ultimately, the signal-to-noise ratio determines the sensitivity of a search, so it is also important to consider the brightness of background emission in evaluating which objects to study. Section 2.4 discusses some of the important gamma-ray search targets in the indirect detection literature.

2.2 Template Fitting

The goal of an indirect detection search is to find a signal coming from DM annihilation or, in the absence of such a signal, to set a constraint on the DM parameter space (for a given DM mass $m_\chi$, the absence of a signal can be used to set an upper bound on the annihilation cross-section $\langle \sigma v \rangle$). A standard procedure by which to do this is template fitting. The basic idea is that the observed data can be described as a sum of spatial maps, referred to as templates, each describing the expected emission from a different source. In a typical Fermi template fitting analysis, there are several templates describing background (astrophysical) emission: (i) Galactic diffuse emission, which comes from cosmic rays interacting with the interstellar medium as well as inverse-Compton scattering of low-energy photons off of free electrons and positrons, (ii) isotropic emission, which primarily comes from extragalactic sources, (iii) resolved Fermi point sources (such as the 3FGL catalog [113], or more recently, the 4FGL catalog [114]), and (iv) the Fermi bubbles [115], two extended emission regions whose precise physical origin remains unknown. Examples of these background templates are shown in Fig. 2.3. In addition, when conducting a search for a DM
Figure 2.3: Astrophysical templates typically included in Fermi dark matter analyses: (a) Galactic diffuse emission following the Fermi gll_iem_v02_P6_V11_DIFFUSE (p6v11) model [116], (b) isotropic emission, which does not look flat due to the spatially-dependent exposure map of the LAT instrument, (c) emission from 3FGL resolved point sources [113], and (d) emission from the regions of the Fermi bubbles [115] (taken to be spatially flat, then convolved with the exposure map). All templates shown here except for the 3FGL template are normalized to their respective best-fit values in the $|b| > 2^\circ$, $r < 30^\circ$ region with a $0.8^\circ$ mask on 3FGL sources; the normalization of the 3FGL template shown here is arbitrarily chosen.

signal, one would include a signal template, which would be dependent on the particle physics specifics as well as the $J$-factor (see Eq. 2.7 and Fig. 2.2).

The data is binned spatially and in energy, and each template is binned in the same way as the data and associated with a normalization that is treated as a free parameter in the fitting procedure. We can then denote the number of observed photon counts in pixel $p$ and energy bin $i$ by $n^p_i$ and the corresponding expected
number of counts predicted by the model by $\mu_i^p(\theta_i)$. More specifically,

$$
\mu_i^p(\theta_i) = \sum_{\ell} A_{\ell,i} T_{\ell,i}^p, \quad (2.10)
$$

where $\ell$ is the index over templates, the model parameters $\theta_i$ are simply the normalizations (in energy bin $i$) of each of the templates $A_{\ell,i}$, and $T_{\ell,i}^p$ is the value of template $\ell$ in the pixel and energy bin. For a given model, the corresponding observation would be a Poisson realization of the sum of modeled components. Thus, the likelihood for observing the data given the model is given by the product of Poisson probabilities

$$
L_i(d_i|\theta_i) = \prod_p \frac{\mu_i^p(\theta_i)^{n_p^i} e^{-\mu_i^p(\theta_i)}}{n_p^i!}, \quad (2.11)
$$

where $d_i$ is the observed data in energy bin $i$. One can then apply Bayesian inference or frequentist techniques to statistically determine the relative contributions of the different components to the observed data. The frequentist approach is more commonly used in the DM indirect detection literature, to either claim detection of a signal, or to place confidence limits on the dark matter parameters $m_\chi$ and $\langle \sigma v \rangle$ in the absence of a signal. This approach is employed in Chapter 3.

### 2.3 Non-Poissonian Template Fitting

The procedure described in Sec. 2.2 only applies to the cases where the sources of emission are either diffuse in nature or are resolved point sources. For a resolved point source in pixel $p$ that emits $\mu_p$ photons in expectation, the probability of observing $n_p$ photons in that pixel due to the point source is indeed given by the Poisson probability centered at $\mu_p$—this is because the location of the point source is known. In the case of an unresolved point source, the story changes. Suppose there is an unresolved
source that also emits $\mu_p$ photons in expectation, but the probability that it is in pixel $p$ is now less than unity. The likelihood in this case of observing $n_p$ photons in pixel $p$ due to the point source involves the Poisson probability that the source will be in that pixel as well as the Poisson probability of observing $n_p$ photons from it—the resulting likelihood is non-Poissonian. The photon statistics of the signal from a population of unresolved point sources would therefore be very different from that of a smooth signal with the same spatial morphology. In particular, the signal from a population of unresolved sources is more “clumpy,” on average contributing more pixels that are particularly bright or particularly dim. This idea is the basis for a statistical procedure called Non-Poissonian Template Fitting (NPTF) [117, 118] (built on the formalism introduced in Ref. [119]).

The difference in “clumpiness” is shown in Figure 2.4, while Figure 2.5 demonstrates the corresponding difference in photon statistics. In one case, a smooth DM
Figure 2.5: The histogram of photons per pixel across the sky, in the region of interest used in the analysis presented in Chapter 4, $|b| > 2^\circ$, $r < 30^\circ$. The gray shaded histogram corresponds to emission from a smooth DM component following a generalized NFW distribution with an inner slope of $\gamma = 1.2$, similar to the left panel of Fig. 2.4. The red outline histogram corresponds to emission from a population of unresolved point sources (PSs) which follow the same spatial distribution, similar to the right panel of Fig. 2.4. The total photon flux is the same in the two cases. The map of unresolved PSs contains more bright and dim (not apparent here due to the scale of the $y$-axis) pixels than the map of smooth emission.

The map is constructed following a generalized NFW distribution:

$$\rho_{\text{NFW}}(r) = \frac{\rho_0}{(r/r_s)^\gamma[1 + (r/r_s)]^{3-\gamma}},$$  \hspace{1cm} (2.12)

with $\gamma = 1.2$, which is motivated by the best-fit values found in dark matter interpretations of the Fermi Galactic Center Excess (GCE) [120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137]. The map is normalized such that the total flux is equal to the flux of the GCE in the chosen region-of-interest (ROI), $|b| > 2^\circ$, $r < 30^\circ$ (this is the same ROI that was used in the analysis presented in Chapter 4). A sky map of this case is shown in the left panel of Fig. 2.4 (shown in a slightly different ROI for visualization purposes), whereas the corresponding his-
Figure 2.6: The same as Figure 2.4, but with the addition of the best-fit Galactic diffuse emission, modeled by the \textit{Fermi p6v11} model. The color bars here are in units of $\log_{10}(\text{photon counts per pixel})$.

togram of the photon counts per pixel is shown in shaded gray in Fig. 2.5. By contrast, the right panel of Fig. 2.4 and the red outlined histogram in Fig. 2.5 correspond to emission from a population of unresolved point sources (PSs), which follow the same spatial distribution\footnote{More precisely, the line-of-sight integrated NFW squared map is treated as the number density distribution of the PSs, from which the positions of simulated sources are drawn. This way, the emission from the collection of sources follows the same spatial distribution as the smooth NFW DM emission.} and contribute the same total flux. The two flux distributions in Fig. 2.5 are clearly distinct even by eye, especially for the brighter pixels (in fact, the PS map has more pixels in the dimmest bin as well, although it is not apparent due to the $y$-axis scale).

In practice, the presence of bright astrophysical backgrounds makes it less simple to distinguish the two cases—Figure 2.7 is the same as Figure 2.5, with the addition of the histogram of counts per pixel for the best-fit Galactic diffuse emission (using the \textit{Fermi p6v11} template) in the ROI; Figure 2.6 shows the corresponding sky maps.

The Galactic diffuse foreground is by far the dominant astrophysical background in this ROI, and it is also significantly brighter than either of the simulated signals.
Figure 2.7: The same as Fig. 2.5, but with the diffuse foreground emission—which is the dominant component of the astrophysical backgrounds—overlaid, shown by the green hatch-filled histogram. The diffuse model shown here is the \textit{Fermi p6v11} model, which was shown in Fig. 2.3a. The emission from the diffuse foreground is much brighter than DM/PS signals, which are normalized to the total flux of the GCE.

Therefore, one must utilize more sophisticated techniques to \textit{statistically} distinguish the smooth component from the unresolved PS emission. More technical details on the NPTF method and its implementation in the code package \texttt{NPTFit} can be found in Chapter 4 of this thesis as well as Refs. [138, 139]. The unresolved point sources themselves which are picked up by the NPTF method may be dark matter substructure [140], astrophysical objects such as millisecond pulsars [118], or even unphysical sources such as residual “hot spots” due to imperfect modeling of astrophysical backgrounds (see, for example, Sections 2.5 and 4.6).

### 2.4 Search Targets in the Gamma-Ray Sky

The previous sections in this chapter introduced some of the statistical methods used in searching for signatures of DM annihilation in gamma-ray data, while this
FIG. 5. Full-sky map, in Galactic coordinates, of the number of photons (above 3 GeV) produced by DM annihilation (benchmark A). The left (right) panel shows the predicted flux in the Aquarius (Via Lactea II) setup.

FIG. 6. Same as Fig. 5, but with the two simulation setups rescaled to the same local density, same total mass and same fraction of mass in substructures.

An additional source of discrepancy is the fact that the total mass of the MW in the Via Lactea II simulation is smaller than in Aquarius, as reported in Tab. I. However, as shown in Fig. 6, the two predictions can be brought in agreement by requiring that (i) both Via Lactea II and Aquarius have the same local density (we have taken the recent estimate $\rho = 0.385$ GeV/cm$^3$ from [74, 75]), (ii) the same subhalo mass fraction ($f_{\text{tot}} = 0.18$) is adopted and (iii) the same mass profile is assumed.

### A. Experimental detectability

In order to assess the detectability of the $\gamma$-ray annihilation flux with the Fermi-LAT satellite, we have to specify what the signal, background or noise are. If we are interested in finding a signal above the astrophysical backgrounds, the signal is contributed by the sum of all the aforementioned components of the annihilation flux (MW smooth mass distribution + galactic subhalos + extragalactic halos and subhalos). We focus on photons with energies larger than 3 GeV and we assume an exposure time of 1 year, which corresponds to about 5 years of data taking with Fermi, and we assume an effective detection area of $10^4$ cm$^2$. We don't consider here any dependence on the photon energy nor on the incidence angle. The background or noise is contributed by the diffuse Galactic foreground and the unresolved extragalactic background. As mentioned in Sect. I, to model such contributions we have rescaled the EGRET data at $E > 3$ GeV by 50%. We remind that this reduction reflects the fact that the Fermi data do not confirm the so-called galactic excess measured by EGRET. The expected sensitivity is simply given by

$$E_{\text{cut}} = 3 \text{ GeV}.$$
the center of the halo and falls off steeply as a function of radial distance. Additionally, there are smaller, point-like clumps of bright emission coming from subhalos scattered throughout the main halo. The more massive subhalos may host galaxies such as the Milky Way dwarf galaxies, while the less massive subhalos would be completely devoid of luminous matter. In the real data, an additional source of point-like emission could be other galaxies, which are similar to the Milky Way in mass but are much farther away than substructure within the Milky Way halo.

While the expected signal brightness is helpful in determining where to search for a signal, a more informative heuristic for evaluating the potential sensitivity of a particular search target is the signal-to-noise ratio. The following are search targets of particular interest in the literature, each of which comes with its own unique set of advantages and challenges in the interplay between signal and background:

- **Milky Way halo—Inner Galaxy** Because the theory expectation is that the signal is highly concentrated in the inner region of the Milky Way, a natural place to target is the Inner Galaxy.\(^9\) For reference, the total $J$-factor in the region $|b| > 2^\circ$, $r < 30^\circ$ for an NFW halo with scale radius $r_s = 17$ kpc and local density $\rho(r_\odot) = 0.4$ GeV cm$^{-3}$ [68, 69] at the Solar position $r_\odot = 8$ kpc [112] is $J_{IG} \sim 6.8 \times 10^{22}$ GeV$^2$cm$^{-5}$. Additionally, the Inner Galaxy has been of particular interest in the literature because excess emission in the $\sim$ GeV energy range with roughly spherically symmetric morphology was found in the *Fermi* data. This excess was first published in [120], and subsequently received a lot more attention because the spectrum and spatial morphology of the excess emission could be interpreted as the annihilation signal of thermal WIMP dark matter with mass $m_\chi \sim 30–50$ GeV [126, 129, 130, 132, 133, 137]. This is referred to

\(^9\) A note on notation: throughout this thesis, capitalized terms such as “Galaxy”, “Galactic halo”, “Galactic Center”, and “Inner Galaxy” will be used to refer specifically to the case of Milky Way, while lowercase terms will refer to the more general case.
as the Galactic Center Excess (GCE). However, there are many challenges to performing a robust dark matter analysis in the Inner Galaxy, because there is also a high concentration of gas, dust, and stars in this region, and consequently complicated astrophysics to contend with. Section 4.1 is dedicated to providing more details on the recent body of work surrounding caveats to the DM interpretation of the GCE.

- **Milky Way halo—high latitudes** Due to the large spatial extension of the Milky Way dark matter halo, the expected signal can still be relatively bright even as one moves away from the Galactic Center towards higher Galactic latitudes. For example, in the $|b| > 20^\circ$, $r < 50^\circ$ region used in the analysis presented in Chapter 3, the total $J$-factor (for the same assumptions about the profile as above) is $J_{\text{high-lat}} \sim 2.2 \times 10^{22} \text{GeV}^2 \text{cm}^{-5}$. While the signal strength is comparable to that in the Inner Galaxy, the astrophysical backgrounds are substantially less dominant in this region. However, the analysis at high latitudes involves a delicate balancing act between choosing a large enough region that the signal is sufficiently large and there is enough photon statistics, while not choosing so large of a region that the analysis is biased due to fitting imperfect background models over a large region of the sky. (Such biases from mismodeling backgrounds are discussed more in Section 2.5.) This was mitigated in the analysis presented in Chapter 3 by subdividing the region-of-interest into eight radial slices over which the norm of the template for Galactic diffuse emission—the dominant source of astrophysical background emission—was allowed to float independently; for annihilation into $b$ quarks, the analysis excluded DM with mass $m_\chi \lesssim 70 \text{ GeV}$ at the 95% confidence level [143]. Section 4.1 discusses the importance of this result within the context of explaining the GCE.
• **Dwarf spheroidal galaxies (dSphs)** The satellites of the Milky Way, in particular the dwarf spheroidal galaxies (dSphs), are considered to be some of the cleanest systems in which to search for dark matter annihilation signals. This is because the systems are expected to be dark matter-dominated, with little to no gas and therefore very little diffuse gamma-ray emission [144, 145], and negligible emission from other astrophysical sources such as millisecond pulsars as well [146]. Therefore, although the $J$-factors are considerably lower than that of the Milky Way halo (due to the dSphs being considerably farther away), with typical values quoted in the literature being around $J_{\text{dSphs}} \sim 10^{17}$–$10^{19}$ GeV$^2$cm$^{-5}$ [147], results from dwarf galaxies are generally considered to be far more robust. The most recent stacked dwarf galaxy analyses from members of the Fermi-LAT collaboration exclude or are in tension with the DM interpretation of the GCE at the 95% confidence level for annihilation into $b$ quarks [148, 149]. However, there are important systematic uncertainties to consider in calculating the expected DM signal from the dSphs (this is discussed in more detail in Chapter 5), so while astrophysical gamma-ray emission is sub-dominant in these systems, there are also important caveats to the resulting dark matter constraints.

• **Dark subhalos** The bottom-up nature of structure formation in CDM, in which galaxy-sized DM halos are formed through the the merging of smaller halos, predicts that within galactic halos there must be abundant DM subhalos. Furthermore, the prediction is that smaller dark matter subhalos should be far more numerous than more massive ones; in particular, $N$-body simulations of Milky Way-like halos find that the subhalo mass function is well-described by a power law $dN/dm \propto m^{-\alpha}$ with $\alpha \approx 1.9$ [150, 151]. While the more massive subhalos host the dwarf galaxies, the less massive subhalos are not
expected to host star formation (the cutoff virial mass is roughly $M_{200} \lesssim 10^8$–$10^9 M_{\odot}$ [152, 153, 154, 155, 156, 157]), so they would be nearly entirely dark. One cannot rely on measuring the luminous matter to pinpoint the location of one of these subhalos in order to search for a DM signal. Instead, the main method in the literature for searching in Fermi data for dark subhalos is to assume that the emission from unassociated sources in the Fermi point source catalogs comes from DM annihilation in candidate subhalos, compare the observed spectra and number of candidate subhalos to predictions from simulations, and thereby set constraints on the DM parameter space. This method utilizes the brighter of the dark subhalos—ones that are bright enough to be individually resolved by Fermi-LAT. A complementary approach is to use the NPTF method introduced in Section 2.3 to search for the signal from the entire population of subhalos, most of which are dim and unresolved. In a proof-of-principle study on simulated data, the NPTF approach yields a comparable limit to that derived from the resolved subhalo approach [140].

2.5 Background Mismodeling

One of the major challenges in performing gamma-ray dark matter analyses is that the astrophysical background emission can be highly complex and difficult to accurately model. In particular, Galactic diffuse emission is by far the dominant background component and is notoriously difficult to model accurately [133]. While constructing more accurate models for Galactic diffuse emission is an active area of ongoing research within the astroparticle physics community, the general consensus is that even state-of-the-art models are likely far-from-perfect descriptors of the actual data. It is therefore crucial when performing DM analyses to find ways to mitigate potential
effects of mismodeling the backgrounds, and also to compare any results derived from the data to the expectation of what the result should be when the backgrounds are perfectly modeled.

Mismodeling the background emission can bias results in two main ways: excesses, leading to false signals or erroneously weak limits; or “oversubtraction” of the backgrounds, leading to constraints that are erroneously strong. The following is a demonstration of how these can occur, in a very simple scenario to help build intuition. Suppose we want to set an upper flux limit on a spherically symmetric signal, depicted in Fig. 2.9. Assume the data consists purely of background emission, described by the true background template shown in the left panel of Fig. 2.10, i.e. the data is a Poisson realization of the background template multiplied by some true normalization $A_{\text{true}}$, shown in the middle panel for $A_{\text{true}} = 5$. We can then perform a Poissonian template fit to this mock data, using the frequentist profile likelihood method to constrain the normalization of the signal template.\footnote{This is very similar to what we do in practice to constrain DM parameters, because for fixed astrophysics assumptions, pairs of values of $(m_\chi, \langle \sigma v \rangle)$ can be mapped 1-to-1 onto values of the signal luminosity, and therefore the normalization of the signal template.} We can consider three different scenarios.
Figure 2.10: The left panel shows the true background template, from which our mock data is generated; the middle template shows the mock data, which is simply a Poisson realization of the background template multiplied by some true normalization $A_{\text{true}}$ (in this particular example $A_{\text{true}} = 5$); and the right panel shows the mismodeled background template, whose morphology does not accurately describe the underlying flux distribution of the mock data. The color bars here do not have physical units, but could correspond to, e.g., photon counts per pixel.

Figure 2.11: Residual maps from performing a fit with only the background template. The left panel shows the case when the true background template is used; the residuals are consistent with Poisson noise. The middle panel shows the case when a mismodeled background template is used (corresponding to the right panel of Fig. 2.10). In this case, there is oversubtraction of the background, and there are large negative residuals. The right panel uses the same mismodeled background, but the fit is performed over a smaller region, and the degree of oversubtraction is less severe.

**Perfectly Modeled Background**

When the true background template (left panel of Fig. 2.10) is used in the analysis, the best-fit normalization of the background template is very close to the true
normalization, $A_{\text{best-fit}} \sim A_{\text{true}}$. A map of the residuals from fitting the data with background template is shown in the left panel of Fig. 2.11; the residuals are consistent with Poisson noise, as is expected to be the case when the backgrounds are perfectly modeled. We can build up a likelihood profile by sampling discrete values of the signal normalization $A_{\text{signal}}$ and maximizing the likelihood at each point with respect to the background normalization. We can then define the test statistic (TS) to be $TS = 2 \times (\ln \mathcal{L} - \ln \mathcal{L}_{\text{max}})$, and use this TS to set confidence limits. The resulting TS profile from this case is shown by the blue line in Fig. 2.12. In this case, the 95% confidence limit on the signal normalization, set at $TS = -2.71$ (denoted by the dashed gray line), is $A_{\text{signal}} \lesssim 0.8$, i.e. a signal normalization greater than 0.8 is excluded at the 95% confidence level.
Oversubtraction Due to Mismodeling

Instead of the true background template, suppose we now use the template shown in the right panel of Fig. 2.10 to model the background emission. This template mischaracterizes the morphology of the true flux distribution, as well as the relative brightness of the two peaks. In this case, the best-fit background normalization $A_{\text{best-fit}} > A_{\text{true}}$, and there are large negative residuals after fitting the background model to the data (middle panel of Fig. 2.11). As a result, there is less flux that can be attributed to the signal component—oversubtraction of the backgrounds leads to too stringent of an upper bound on the signal strength. The TS profiles shown in Fig. 2.12 explicitly demonstrate this. The red line shows the TS profile when using the wrong background template—the resulting limit on $A_{\text{signal}}$ is much stronger than it should be. The green line shows what happens when the same mismodeled template is used, but the fit is performed over a smaller region of the sky (the same region as shown in the right panel of Fig. 2.11). In this case, the degree of oversubtraction is somewhat mitigated.

Excess Due to Mismodeling

While oversubtraction is a serious concern in the context of trying to set a robust upper limit on the DM annihilation cross-section, background mismodeling does not always lead to oversubtraction. In this final scenario, consider a slightly different true background template and mismodeled background template than before (shown in the left and right panels of Fig. 2.13, respectively). In this case, when using the mismodeled template in the analysis, the best-fit background normalization $A_{\text{best-fit}} < A_{\text{true}}$, and there are large positive residuals (shown in the left panel of Fig. 2.14). Once again, this effect is somewhat mitigated by performing the fit over a smaller region (right panel of Fig. 2.14). The TS profile now has a strong peak at a large value.
Figure 2.13: A different set of true (left panel) and mismodeled (right panel) background templates. The mock data is again a Poisson realization of the true background template multiplied by a true normalization $A_{\text{true}}$, shown in the middle panel for a true normalization of $A_{\text{true}} = 5$. In this case, fitting with the mismodeled template results in an excess.

Figure 2.14: Residual maps from performing a fit with only the background template. The left panel shows the case when a mismodeled background template is used (corresponding to the left panel of Fig. 2.13). In this case, there are large positive residuals. The right panel uses the same mismodeled background, but the fit is performed over a smaller region, and the degree of the excess is less severe.

of $A_{\text{signal}}$, shown by the red line in Fig. 2.15—this would correspond to detecting a spurious signal with best-fit normalization $A_{\text{signal}} \sim 7$ (recall that the data contains no signal in it, so the true signal normalization is 0). A somewhat smaller signal is detected when performing the fit over a smaller region, shown by the green line in Fig. 2.15. In this toy example, there happens to be a very small excess even when
there is no mismodeling, which is seriously exacerbated by the presence of background mismodeling.

The toy examples presented here were intended to give some intuition for the effects of background mismodeling on the search for a hypothetical signal. In practice, background mismodeling is much more complicated in gamma-ray dark matter analyses than what has been demonstrated here—especially in the presence of additional degeneracies between different emission components—and it is important to seek to understand and mitigate its effect, whether by choosing a smaller region-of-interest, using methods to give the background components more freedom, or using additional information such as spectral information.
Chapter 3

Search for Dark Matter

Annihilation in the Milky Way Halo

This chapter is based on work done in collaboration with Siddharth Mishra-Sharma and Mariangela Lisanti, published as *Physical Review D* **98**, 123004 (2018) (arXiv:1804.04132). Section 3.1 provides a overview of gamma-ray searches for signatures of dark matter annihilation and motivates choosing the Milky Way dark matter halo as a search target. Section 3.2 describes the analysis pipeline and statistical procedure. Section 3.3 contains the main results of the study, including a discussion on the effects of Galactic foreground mismodeling and steps taken to mitigate their impact, as well as the implications of the results on the dark matter interpretation of the Galactic Center Excess. Section 3.4 summarizes the main conclusions, while Sections 3.5–3.7 contain extended results. This work has been presented at the following conferences and seminars: *TeV Particle Astrophysics (TeVPA) 2018* in Berlin, Ger-

3.1 Introduction

The Fermi Large Area Telescope [103] provides an unprecedented view of the gamma-ray sky. The all-sky maps that are available can harbor clues about the nature of dark matter, which can annihilate to visible states that produce showers of high-energy photons. A variety of such searches have been performed, focusing on regions where the relative DM density is expected to be significant. Some of the most sensitive bounds come from looking at ultrafaint dwarf galaxies [158, 159, 160, 148] and galaxy groups [161, 162]. In this chapter, we explore emission due to annihilating DM from the Galactic halo, and demonstrate that it can be used to set robust constraints on the DM annihilation cross section. At the time of publication, these constraints were the strongest on DM with mass less than $\sim 70$ GeV, for the $b\bar{b}$ annihilation benchmark.

The halo surrounding our Galaxy provides the brightest source of DM emission on the sky. In general, the DM flux is proportional to the so-called $J$-factor, which is the integral over the line-of-sight, $s$, and solid angle, $\Omega$, of the squared DM density profile:

$$J = \int ds \, d\Omega \, \rho^2(s, \Omega).$$

The $J$-factor provides a useful metric for comparing the strength of an annihilation signal expected from different targets. For example, the $J$-factors from some of the brightest ultrafaint dwarf galaxies are $\sim 10^{19}$ GeV$^2$ cm$^{-5}$ sr [148], comparable to those
of the brightest galaxy groups [161]. In contrast, the center of our own Galaxy has a $J$-factor several orders of magnitude larger, with $J \sim 10^{23} \text{ GeV}^2 \text{ cm}^{-5} \text{ sr}$ in the inner $40^\circ \times 40^\circ$ region. Even if one were to avoid the central part of the Galaxy and only consider an annulus of $r < 50^\circ$ and latitudes greater than $|b| > 20^\circ$, the $J$-factor is still as large as $\sim 10^{22} \text{ GeV}^2 \text{ cm}^{-5} \text{ sr}$.

Despite the strength of the smooth Galactic DM signal, many other factors complicate a potential search. The primary challenge is posed by the bright diffuse emission from cosmic rays propagating in the Galaxy. These contributions arise from $\pi^0$ decay, Bremsstrahlung from the interaction of cosmic-ray electrons with interstellar gas, and inverse-Compton (IC) scattering of photons off of high-energy electrons. This diffuse foreground contributes the vast majority of the high-energy photons we see on the sky, accounting for $\sim 50$–$90\%$ of the observed photons depending on the energy range considered [163], and is challenging to model accurately. Any search for Galactic DM must mitigate these uncertainties and quantify the effects of varying over assumptions in the foreground models.

Searches for Galactic DM can be divided into two broad categories. The first set focuses on the Inner Galaxy, within $r \lesssim 20^\circ$ [120, 122, 164, 123, 124, 125, 126, 127, 128, 129, 130, 131, 133, 134, 135, 137]. These analyses have conclusively found an excess of $\sim \text{GeV}$ photons whose energy distribution and spatial morphology can be consistent with the expectation due to DM annihilation. However, recent studies have shown that the distribution of photons in the Inner Galaxy is more consistent with a population of unresolved point sources, disfavoring the DM interpretation [118, 165]. Additionally, other studies suggest that the spatial morphology of the excess may better trace the stellar bulge [166, 167]. Complementary studies of Milky Way dwarfs [148] and galaxy groups [161, 162] are starting to put in tension the DM interpretation of the excess emission. However, the tension can be alleviated depending on the spe...
cific assumptions made about, e.g., the dwarf halo profiles [168, 159, 169]; the stellar membership criteria used to infer the dwarf halo properties [170, 147]; the shape of the Milky Way halo [171]; or the nature of substructure boost in galaxy groups [162].

Even though the Galactic Center is the brightest DM source on the sky, it is also one of the most complicated due to the large astrophysical foregrounds. A complementary approach to looking at the Inner Galaxy relies on looking at the Galactic halo at higher latitudes where the DM density is still large, but the foreground levels are much smaller [172, 173, 174, 175, 176, 177, 178]. This is the approach that we take in this work. Focusing on a region defined by $|b| > 20^\circ$ and $r < 50^\circ$, we search for signals of DM annihilation from the smooth Milky Way halo. The limits obtained provide the strongest constraints on low-mass DM annihilation signals and tightly constrain the DM interpretation of the GeV Excess. We verify the robustness of these results in the presence of a potential DM signal and discuss how they are affected by variations in the Galactic foreground models.

3.2 Analysis Procedure

We make use of 413 weeks of Fermi-LAT Pass 8 data collected between August 4, 2008 and July 7, 2016. We analyze the subset of photons in the ULTRACLEANVETO event class, restricting to the top quarter of photons by quality of point-spread function (PSF) reconstruction (corresponding to PSF3 event type). The data is binned in 18 logarithmically-spaced energy bins between $\sim 0.8$–50 GeV. The recommended quality cuts are applied, corresponding to zenith angle less than $90^\circ$, LAT_CONFIG = 1, and DATA_QUAL $> 0$.\footnote{https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Data_Exploration/Data_preparation.html} Each energy bin is spatially binned into individual pixels us-
ing HEALPix [179] with \texttt{nside} = 128; the dataset is thus reduced to an array of integers that describes the number of photons in the energy bin, \(i\), and pixel, \(p\).

Template fitting is a standard astrophysical procedure where the data is described by a set of spatial maps (referred to as templates) that are binned in the same way as the data, which describe the separate components that contribute to the total photon count. Each template is associated with a normalization that is treated as a free model parameter in the fit. The likelihood for a given energy bin is then a product of the Poisson probabilities associated with the observed counts \(n^p_i\) in each pixel of the region-of-interest:

\[
L_i(d_i|\theta_i) = \prod_p \frac{\mu^p_i(\theta_i)^{n^p_i}e^{-\mu^p_i(\theta_i)}}{n^p_i!},
\]

where \(d_i\) denotes the data in energy bin \(i\), \(\theta_i\) represents the set of model parameters and \(\mu^p_i(\theta_i)\) is the number of expected counts in a given pixel and energy bin. The total likelihood is simply the product over the individual \(L_i\) for each energy bin.

The region-of-interest (ROI) for this study is chosen to maximize the strength of the DM signal while minimizing the effects of foreground mis-modeling. Specifically, we take \(|b| > 20^\circ\) to avoid the Galactic plane, where cosmic-ray emission is particularly bright and there are more unresolved point sources. In addition, we take \(r < 50^\circ\) to reduce the possibility of over-subtraction and/or spurious excesses obtained from fitting the foreground model over large sky areas. Better modeling of the Galactic diffuse emission is an ongoing effort [180, 181, 182, 183, 184], and the use of image reconstruction and parametric modeling techniques such as SKYFACT [185] and D³PO [186, 178] could improve modeling of the Galactic diffuse emission over significantly larger regions of the sky. Increasing the radial cut in \(r\) beyond that used here could potentially improve sensitivities to DM by \(\sim 20-30\%\) or more, depending on the density profile—see Sec. 3.5 for a discussion of Asimov projections.
The expected photon count, $\mu_i^p(\theta_i)$, in each pixel of the ROI depends on contributions from standard astrophysical sources as well as DM, if present. We account for four astrophysical components that trace: (i) the Galactic diffuse emission, as described by the Fermi gll_iem_v02_P6_V11_DIFFUSE (p6v11) model,\(^2\) (ii) the Fermi bubbles \(^1\)\(^1\), (iii) isotropic emission, and (iv) Fermi 3FGL point sources \(^1\)\(^3\). The smooth Galactic DM template is modeled using a generalized Navarro-Frenk-White (NFW) profile \(^1\)\(^1\):

$$\rho_{\text{NFW}}(r) = \frac{\rho_0}{(r/r_s)^\gamma[1+(r/r_s)]^{3-\gamma}}$$

(3.3)

with inner-slope $\gamma = 1$, scale radius $r_s = 17$ kpc, and local density $\rho(r_\odot) = 0.4$ GeV cm$^{-3}$ \(^6\)\(^8\) at the Solar position $r_\odot = 8$ kpc \(^9\)\(^1\)\(^2\). All templates are smoothed with the energy-dependent PSF of the LAT instrument, modeled as a King function.\(^3\)

In our fiducial study, we use the p6v11 Galactic diffuse emission model, which is designed to capture changes in the cosmic-ray emission on the full sky as a function of Galactocentric radius. The model includes contributions from $\pi^0$-decay and Bremsstrahlung emission, as traced by maps of gas column-densities, as well as inverse-Compton emission, as predicted using GALPROP \(^1\)\(^8\); the relative normalizations of these separate components are fixed. The fact that the p6v11 model should be used with caution for energies above $\sim 50$ GeV sets our upper energy cut-off. To give the p6v11 template more freedom, we divide it into eight radial slices of equal area within our ROI. The normalization of each slice is then varied separately in the fitting procedure. Each slice is roughly $\sim 440$ deg$^2$ in area, comparable in size to the regions used in dwarf and galaxy group studies ($\sim 100$ and 316 deg$^2$, respectively) \(^1\)\(^4\)\(^8\)\(^9\)\(^2\) and smaller than the typical regions used in Inner Galaxy analyses ($\sim 1600$ deg$^2$) \(^1\)\(^3\)\(^0\)\(^3\)\(^0\). The additional freedom given to the Galactic diffuse tem-

\(^2\)https://fermi.gsfc.nasa.gov/ssc/data/access/lat/ring_for_FSSC_final4.pdf
\(^3\)https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_LAT_IRFs/IRF_PSF.html
plate allows the fit to better account for localized excesses or mis-modeled features in the emission. Note that we do not use the gll_iem_v06 (p8R2) model [188], which is recommended for the Pass 8 dataset. p8R2 includes large-scale residuals obtained from a fit to the Fermi data that have been added back into the model, and is therefore not appropriate when searching for extended DM signals. We also do not use the p7v6 diffuse model, which contains large-scale structures including Loop I and the Fermi bubbles with a fixed normalization.

The p6v11 model does not include known large-scale structures that overlap with the ROI, such as the Fermi bubbles. We account for the bubbles by adding two templates that model the Northern and Southern lobes. The shape of the lobes is inferred directly from Fermi data [115], and the intensity of the emission is taken to be flat. We let the normalization of the Northern and Southern lobes float independently in the fit. The ROI also overlaps with Loop I, a large radio lobe in the Northern hemisphere [189, 190]. While features corresponding to the radio observations have been observed in the Fermi data [191, 115, 192, 193], significant uncertainties remain in the modeling of the spatial and intensity profile of Loop I in gamma rays. As a result, we conservatively do not include a template that traces Loop I in our fiducial study. We have performed variants of the fiducial study to assess the impact of this choice. We find that the inclusion of an additional isotropic template in the Northern hemisphere as a proxy for Loop I emission strengthens the limit by a factor of $\lesssim 1.2$.

The isotropic template is intended to primarily capture extragalactic gamma-ray emission from unresolved sources such as blazars and star-forming galaxies, as well as more exotic contributions from extragalactic DM annihilation. The inclusion of the point-source template accounts for emission from resolved (Galactic and extragalactic) sources. The normalizations of all the sources are floated together in the template after fixing their individual fluxes to the values predicted by the 3FGL catalog. We
note that all 3FGL sources are conservatively masked to 95% containment in PSF for the corresponding energy bins. Therefore, the primary purpose of the point-source template is to account for any potential mis-modeling in the tails of the emission.

To summarize, there are twelve free parameters associated with the astrophysical components—eight for the Galactic diffuse slices, two for the Fermi bubbles, and one each for the isotropic and point-source templates. As we are ultimately interested in the intensity of the DM signal, we treat these as nuisance parameters and remove them using the profile likelihood method [194]. Specifically, we build a likelihood profile for the intensity associated with DM annihilation in the smooth Galactic halo, fixing the normalization of this template at various values while profiling over the astrophysical components. The resulting likelihood only depends on the DM intensity in each energy bin, which is related to the annihilation cross section, $\langle \sigma v \rangle$, and mass, $m_\chi$, through the expression for the differential gamma-ray flux:

$$\frac{d\Phi}{dE_\gamma} = J \times \frac{\langle \sigma v \rangle}{8\pi m_\chi^2} \sum_j Br_j \frac{dN_j}{dE_\gamma},$$

(3.4)

where $E_\gamma$ is the gamma-ray energy and $Br_j$ is the branching fraction to the $j^{th}$ annihilation channel. The energy spectrum for each channel is described by the function $dN_j/dE_\gamma$, which is modeled using PPPC4DMID [109]. Note that we do not account for DM substructure in the Milky Way halo in this study, which would increase the strength of the annihilation signal. Given the theoretical uncertainties associated with modeling the spatial distribution and properties of DM subhalos, such a search deserves its own dedicated study.

The test statistic (TS) profile for $\langle \sigma v \rangle$ is defined as

$$TS \equiv 2 \left[ \log \mathcal{L}(d|\mathcal{M}, \langle \sigma v \rangle, m_\chi) - \log \mathcal{L}(d|\mathcal{M}, \hat{\langle \sigma v \rangle}, m_\chi) \right],$$

(3.5)
where $\langle \sigma v \rangle$ is the cross section that maximizes the likelihood for a specified DM model, $\mathcal{M}$, of given annihilation channel and mass. The TS is nonpositive by definition and can be used to set a threshold for limits on the cross section. In particular, the 95% upper limit on the annihilation cross section is given by the value of $\langle \sigma v \rangle$ associated with $\text{TS} = -2.71$. We implement template fitting with the package NPTFit [138] and use the L-BFGS-B [195] minimization algorithm implemented through SciPy [196].

We have performed numerous tests to ensure that the statistical procedure outlined above can recover a potential signal in the data. Such tests are crucial in verifying the robustness of these methods, especially given the potentially large degeneracies between the signal and foreground components, which are both diffuse in

Figure 3.1: The solid black line shows the 95% confidence limit on dark matter of mass, $m_\chi$, annihilating with cross section, $\langle \sigma v \rangle$, in the smooth Galactic halo, within $|b| > 20^\circ$ and $r < 50^\circ$, obtained using the p6v11 foreground model. The green(yellow) band denotes the 68(95)% containment for the expected sensitivity, as derived from Monte Carlo simulations. For the Galactic halo, we assume a generalized NFW profile with inner slope of $\gamma = 1$ and local density $\rho(r_\odot) = 0.4 \text{ GeV cm}^{-3}$. We also show the corresponding limits obtained from dwarf galaxies [148] and galaxy groups [162] (grey dashed and dot-dashed lines, respectively). The expected annihilation cross section for a generic weakly interacting massive particle is indicated by the solid grey line [197]. The inset depicts the eight radially sliced regions within the fiducial ROI over which the p6v11 template is allowed to float.
nature. Additionally, the freedom given to the foreground emission by separately fitting its normalization in the radial slices can lead to challenges in regimes of low photon statistics. We have performed tests on both data and Monte Carlo and verified that our analysis procedure would not exclude a DM signal if one were present in the data. A detailed description of these tests is provided in Sec. 3.6.

3.3 Results and Discussion

3.3.1 Dark Matter Annihilation Limit

Figure 3.1 shows the 95% confidence limit on the DM annihilation cross section into the $b\bar{b}$ final state (solid black). For comparison, the published limits from the most recent dwarf [148] and galaxy group [162] studies appear as the grey dashed and dot-dashed lines, respectively. The $b\bar{b}$ limits from the smooth Galactic halo are the strongest at the time of publication for DM masses below $\sim 70$ GeV.
The green(yellow) band in Fig. 3.1 shows the 68(95)% expected sensitivity obtained from Monte Carlo simulations. To make the simulations, we Poisson fluctuate the sum of best-fit templates on data within the ROI, letting the normalizations for the different foreground slices and bubble lobes float independently. The sensitivity projection is derived from 100 Monte Carlo variations. A data-driven foreground expectation obtained by looking at a large number of blank fields, as is standard for dwarf and galaxy group studies, is not feasible for Galactic DM searches because the overall size of the ROI is a substantial fraction of the full sky. The Monte Carlo bands do, however, provide an important comparison benchmark. For example, if the Galactic foregrounds are over-subtracted in the fitting procedure, then the data limits will be artificially strengthened and appear stronger than the Monte Carlo expectation.

While the morphology of the signal template suggests that one should minimize the latitude cut ($|b| > b_{\text{cut}}$) and maximize the radial cut ($r < r_{\text{cut}}$) for optimal sensitivity to DM (see Sec. 3.5 for more details), a full-sky analysis is not viable in actuality due to the large uncertainties associated with modeling the Galactic foregrounds. As a result, we conservatively choose $b_{\text{cut}} = 20^\circ$ to avoid the Galactic plane, where the foregrounds are particularly bright and there is increased contamination from unresolved point sources. In addition, we choose $r_{\text{cut}} = 50^\circ$ because fitting over larger sky regions can lead to over-subtraction and/or spurious excesses in the data analysis. While the definition of the fiducial ROI is intended to mitigate the large systematic uncertainties associated with the foregrounds, we also give the p6v11 template additional freedom by fitting its normalization separately in eight radial slices of equal area, as discussed in Sec. 3.2. Figure 3.2 demonstrates the need for these additional steps. The left panel shows the data limit and corresponding Monte Carlo expectation obtained when the p6v11 template is not divided into eighths, for our fiducial ROI. The right panel shows the case corresponding to a larger radial cut $r_{\text{cut}} = 100^\circ$. Every other
aspect of the analysis is kept the same as in the fiducial study in these cases, except that the Northern and Southern lobes of the Fermi bubbles are floated together.\textsuperscript{4}

The projected sensitivities obtained from Monte Carlo simulations are essentially equivalent between the fiducial study and these two examples. The data limits, on the other hand, are starkly different. A large excess in the data limit compared to the Monte Carlo expectation is apparent when the larger ROI is used. When the fiducial ROI is used but the foreground template is not broken into radial slices, over-subtraction leads to artificially strong bounds. We therefore conclude that performing the fit over smaller sky regions and varying the p6$v$11 template over additional degrees of freedom stabilizes the analysis in the designated ROI.

### 3.3.2 Galactic Foreground Modeling

Uncertainties due to modeling of the Galactic diffuse emission are inherent in searches for large-scale gamma-ray structures. We have made an effort to minimize the effects of these uncertainties by giving more degrees of freedom to the p6$v$11 template. However, inherent assumptions that go into the construction of the template can still have a potentially large effect on the final result. Here, we present results for three additional foreground models that are designed to span several well-motivated possibilities. Our approach is to understand how each set of assumptions regarding the cosmic-ray modeling impacts the DM sensitivity for the ROI considered in this work.

We repeat the analysis using Models A, B, and C, which were developed by the Fermi-LAT Collaboration specifically for their study of the isotropic gamma-ray background at higher latitudes [163]. These models make distinct but well-motivated choices for the cosmic-ray source distribution, diffusion coefficients, and

\textsuperscript{4}Doing the same for the fiducial study does not change the result.
re-acceleration strengths that span a wide range of possibilities. Separate templates for $\pi^0$ decay, Bremsstrahlung, and inverse-Compton (IC) emission are provided, so their normalizations can be varied independently in the fitting procedure. In these analyses, we use a single combination of the Bremsstrahlung and $\pi^0$-decay templates as obtained from a fit to data using eight separate equal-area slices. Both these components trace the diffuse gas and dust structures in the Galaxy, so giving them separate degrees of freedom is expected to have a negligible effect on the results.

We highlight the fact that the IC and $\pi^0$+Bremsstrahlung templates are allowed to vary separately in the Model A, B, and C fits. As a result, the foreground templates in these tests are given considerable freedom in the fitting procedure, as they are associated with sixteen free parameters (rather than just eight, as in the p6v11 case). This is a very important cross-check of the fiducial results, because the relative normalizations of foreground components are fixed in p6v11, with the ratio set by a previous fit to the data. However, because that fit did not include a DM template, one might worry that a potential signal—if present—would be absorbed by the foreground components (particularly the IC component) in the initial fitting procedure. If this were the case, using p6v11 for a Galactic DM search could potentially give artificially stringent DM limits.

Fig. 3.3 shows the limits obtained using Models A, B, and C. The differences between the results can be understood in terms of the assumptions going into the separate models, which we now describe in detail:

Model A is based on the class of Galactic diffuse models studied in [198], and is described in detail in [163]. Here, we only highlight the main elements that distinguish it from Models B and C. For Model A, cosmic-ray electrons and nuclei are both sourced by the same population of pulsars, and the cosmic-ray diffusion coefficient and re-acceleration strength are held constant. The left panel of Fig. 3.3 shows the
Monte Carlo expectation and data limit when rerunning the fiducial analysis using the Model A templates. The recovered data limit is weaker than the Monte Carlo expectation, which suggests that there is excess gamma-ray emission in the ROI that is not captured by the Model A templates. It should be noted that the foreground templates are given considerable freedom in the fitting procedure, as the normalizations of the $\pi^0+$Bremsstrahlung and IC templates are allowed to float separately in each radial slice. Despite this freedom, a large amount of DM emission is still needed to improve the quality of the fit. A DM “excess” with a $T_{\text{max}} \sim 28$ is observed, with the best-fit 1σ and 2σ (corresponding to deviations in TS of $-2.30$ and $-6.18$ from the global maximum) containment regions as shown in the figure. The fact that the DM parameter space that is favored is clearly excluded by the dwarf searches strongly suggests that the weakening of the bounds is not due to DM, and is likely of astrophysical origin.
Model B provides an important counterpoint to Model A \cite{163}. It includes an additional source population of electrons at the Galactic Center, which contributes to the IC emission. Unlike Model A, which closely reproduces the local cosmic-ray electron spectrum, Model B under-predicts the distribution below $\sim 20$ GV. However, this disparity can be accounted for by contributions from other more local sources. The middle panel of Fig. 3.3 shows the Monte Carlo expectation and data limit for the Model B study. The limit is comparable to the fiducial case at low masses and is somewhat tighter for masses above $\sim 100$ GeV, although still consistent within the Monte Carlo expectation. The predicted IC spectrum from \texttt{Galprop} that is used in Model B tends to be a better match to the fitted spectrum (compared to Models A and C). The better overall fit of Model B to the data in this case and the fact that the additional emission is absorbed by the IC template means that an astrophysical origin of the excess is statistically preferred to the DM component.

For Model C, the cosmic-ray diffusion coefficient and re-acceleration strength depends on the Galactocentric radius and height \cite{163}. Additionally, while the cosmic-ray electron/nuclei are sourced from the same population, their distribution is more central than that used for Model A. The differences between Model A and C predominantly show up in the outer galaxy, and so the two give largely similar results when used within our ROI. The excess emission observed in the case of Model A is also present using Model C, with a preference for roughly similar DM parameter values. Again, the fact that the preferred parameter space is robustly ruled out by dwarf searches strongly indicates that the excess emission in this case is of astrophysical origin.

To summarize, Model B provides limits very similar to those obtained in the fiducial case, while Models A and C exhibit significant excesses above Monte Carlo expectation. This difference can be attributed to the fact that Model B includes
an additional population of electron-only sources near the Galactic Center that contributes to the IC emission. Omission of this population in Models A and C causes the DM template to absorb more flux, thus weakening the overall bounds. Overall, the fitted IC normalization for Model B is closer to its initial $\text{Galprop}$ prediction (with a value $\sim 1.1$), as compared to that for Models A and C (with a value $\sim 2.4$) \cite{163}. This suggests that, of the three scenarios considered, Model B may best capture the IC emission in the ROI used here.

### 3.3.3 The GeV Excess

The results presented in this chapter have direct implications for the interpretation of the excess of GeV photons observed in the Galactic Center. If the GeV excess arises from DM, then the signal should also contribute a photon flux in the ROI studied here. This is a more direct comparison than using dwarf galaxies or galaxy groups because it removes uncertainties having to do with differences in halo density distribution. In Fig. 3.4, we show the Galactic DM limits obtained for various assumptions of the inner slope, $\gamma$, of the generalized NFW density profile. The steeper the inner slope, the stronger the annihilation limit. Results are shown for annihilation into $b\bar{b}$ (left) and $\tau^+\tau^-$ (right). For comparison, we also show the best-fit regions to the Galactic Center gamma-ray excess from previous work as the data point \cite{129} and solid \cite{126}, cross-hatched \cite{130}, and hatched \cite{133} regions. The DM interpretation of the GeV excess typically prefers a steeper inner slope with $\gamma \gtrsim 1.1$, where the limits from the Galactic halo become quite stringent. These Milky Way limits robustly exclude a DM interpretation of the excess for the $b\bar{b}$ channel, and, for the first time, start probing the $\tau^+\tau^-$ scenario. Explanations in terms of other annihilation channels are also highly constrained, as reviewed in Sec. 3.7.
Variations in the foreground modeling can affect the recovered limits from our analysis and their implications for the Galactic Center Excess. Of the variations explored in Sec. 3.3.2, Model B appears to best capture the IC emission in the ROI, as the fitted normalization of this component is closest to its initial Galprop value. The limits obtained using Model B are only marginally weaker than those using p6v11 at low masses and still robustly disfavor the DM interpretation of the excess in terms of annihilation into the $b\bar{b}$ final state. These results are suggestive, but do not eliminate the systematic uncertainties associated with diffuse emission modeling. To sidestep this issue, we can choose to compare our results to only those Inner Galaxy studies that use the same Galactic foreground model as we do. The cross-hatched region in Fig. 3.4 is derived using the p6v11 diffuse foreground model [130] and therefore provides the most direct comparison to our limit. It is strongly excluded by the limit we recover for the corresponding value of $\gamma$.

3.4 Conclusions

In this chapter, we have presented a comprehensive search for DM annihilation from the smooth Milky Way halo in Fermi gamma-ray data. We do not find significant evidence for an annihilation signal, and obtain strong bounds on the properties of annihilating DM. We exclude thermal dark matter at masses below $\sim70$ GeV for the $b\bar{b}$ annihilation channel when using the Fermi p6v11 diffuse model, representing the strongest limits to date in this mass range. We have carefully considered uncertainties associated with the modeling of the diffuse Galactic foregrounds and are able to understand these variations in terms of the different physical assumptions underlying the foreground models. We have performed rigorous Monte Carlo and injected signal tests to ensure the robustness of our results. This study excludes the $b\bar{b}$ annihilation
Figure 3.4: The 95% confidence limits on dark matter annihilation into $b\bar{b}$ (left) and $\tau^+\tau^-$ (right) for the fiducial analysis, varying over the inner slope, $\gamma$, of the generalized NFW density profile. The limits tighten as $\gamma$ increases; the lines shown correspond to linearly spaced steps from $\gamma = 1$ to 1.5. The best-fit parameters obtained by previous studies of the GeV excess are indicated by the data point [129] and the solid [126], cross-hatched [130], and hatched [133] regions. Each region is indicated by 1σ/2σ contours and colored corresponding to the best-fit $\gamma$ obtained by that study, also specified in the legend. For ease of comparison, we have rescaled the best-fit cross-sections to be consistent with $\rho(r_\odot) = 0.4$ GeV cm$^{-3}$. The corresponding limits obtained from dwarf galaxies [148] and galaxy groups [162] (grey dashed and dot-dashed lines, respectively) are also shown. The expected annihilation cross section for a generic weakly interacting massive particle is indicated by the solid grey line [197].

interpretation of the Galactic Center excess at 95% confidence for the p6v11 diffuse model, and for the first time starts probing the $\tau^+\tau^-$ annihilation interpretation.

The following sections complement the discussion here with extended results. In particular, Sec. 3.5 includes further justification for the choice of ROI and Sec. 3.6 summarizes signal injection and recovery tests. Extended results, including limits for different annihilation channels and DM profiles, as well as other variations of the astrophysical templates, are provided in Sec. 3.7.
Figure 3.5: Sensitivity projections for a 30 GeV dark matter particle annihilating to $b\bar{b}$ for different regions of interest, which are defined by latitude ($|b| > b_{\text{cut}}$) and radial ($r < r_{\text{cut}}$) cuts. The projected limit, $\langle \sigma v \rangle_{\text{lim}}$, is compared to the limit for the fiducial region, $\langle \sigma v \rangle_{\text{fid}}^{\text{lim}}$, which corresponds to $|b| > 20^\circ$ and $r < 50^\circ$. The contours indicate the ratio of these two cross sections. The projections are provided for different dark matter density profiles: (left to right) generalized NFW with inner-slope $\gamma = 1, 1.2$, Einasto, and Burkert with a $r_B = 0.5$ and 10 kpc core.

3.5 The Region-of-Interest

The fiducial analysis presented in this chapter uses a region-of-interest (ROI) defined by the annulus $|b| > 20^\circ$ and $r < 50^\circ$. To motivate this choice, we analyze Asimov datasets [199], which can be used to determine the median asymptotic behavior of the test statistic under the assumption that the foregrounds are perfectly modeled, while varying over different choices of the ROI. The Asimov dataset in this case corresponds to the sum of astrophysical templates best-fit to the data in each ROI. Note that the p6v11 template was not divided into independent radial slices in the Asimov study.

As a concrete example, we consider the case of a 30 GeV DM particle annihilating to $b\bar{b}$, although results for other DM masses are largely unchanged. We vary over latitude ($|b| > b_{\text{cut}}$) and radial ($r < r_{\text{cut}}$) cuts spanning $b_{\text{cut}} = \{15^\circ, 16^\circ, \ldots, 30^\circ\}$ and $r_{\text{cut}} = \{40^\circ, 45^\circ, \ldots, 150^\circ\}$. Figure 3.5 demonstrates how the projected cross section limit, $\langle \sigma v \rangle_{\text{lim}}$, compares to that for the fiducial ROI, $\langle \sigma v \rangle_{\text{fid}}^{\text{lim}}$, as a function of $b_{\text{cut}}$ and $r_{\text{cut}}$. We consider the generalized NFW profile as in Eq. 3.3 with scale
radius \( r_s = 17 \) kpc, local density \( \rho(r_{\odot}) = 0.4 \text{ GeV cm}^{-3} \), and inner slope \( \gamma = 1 \) and 1.2 (first and second panel from left, respectively). In general, we see that the projected sensitivity strengthens for smaller \( b_{\text{cut}} \) and larger \( r_{\text{cut}} \), as expected. This dependence weakens for steeper profiles because the dark matter (DM) density is concentrated towards the Galactic Center. We note that the Asimov projections assume perfect knowledge of the astrophysical components, and as such disregard potential degeneracies between a DM signal and astrophysical templates, which are likely to be important in an analysis on data.

For comparison, we also consider several other DM density profiles, each normalized to \( \rho(r_{\odot}) = 0.4 \text{ GeV cm}^{-3} \). The middle panel of Fig. 3.5 shows the results for the Einasto profile [200]:

\[
\rho_{\text{Einasto}}(r) = \rho_0 \exp \left[ -\frac{2}{\alpha} \left( \left( \frac{r}{r_E} \right)^{\alpha} - 1 \right) \right],
\]

with \( \alpha = 0.17 \) and \( r_E = 15.14 \) kpc [201]. The final two panels in Fig. 3.5 show the results for a cored Burkert profile [202]:

\[
\rho_{\text{Burkert}}(r) = \frac{\rho_0}{(1 + r/r_B)(1 + (r/r_B)^2)},
\]

where \( r_B \) is the analog of the NFW scale radius and sets the size of the core. For illustration, we consider \( r_B = 0.5 \) and 10 kpc, which roughly spans the range of allowed possibilities—see e.g. [203, 204]. While the Einasto contours look very similar to those for NFW with \( \gamma = 1.2 \), the Burkert results are quite different. For the smaller core, there is only very mild dependence on \( r_{\text{cut}} \) and the projected signal strength decreases with larger \( b_{\text{cut}} \). In contrast, the signal is strengthened with decreased latitude and increased radial cuts for the case where \( r_B = 10 \) kpc because the DM distribution is less concentrated towards the Galactic Center.
3.6 Signal Injection and Recovery

A vital consistency check involves ensuring that the limit-setting procedure would not exclude a DM signal if one were present in the data. We perform a variety of tests to confirm that we can set a robust limit while recovering the properties of a DM signal. We perform these checks on both Monte Carlo simulations as well as on the data itself.

**Signal injection on Monte Carlo.** We create Monte Carlo simulations of the gamma-ray sky by summing the astrophysical templates best-fit on data, adding the signal from a DM particle annihilating to $b\bar{b}$ in the smooth Galactic halo, and Poisson fluctuating the final map. We create 50 Monte Carlo realizations of the sky map and pass each through the analysis pipeline. This procedure is repeated for different DM masses and cross sections to study the resulting limit and the test statistic associated with the extracted signal.

Figure 3.6 summarizes the results of the signal injection tests for $m_\chi = 100$ and 1000 GeV in the left and right panel, respectively. In each panel, the gold bands indicate the recovered limit, $\langle \sigma v \rangle_{\text{null, limit}}$, when no signal is injected into the simulated sky map. The green band shows the middle 68% containment of the cross section, $\langle \sigma v \rangle_{\text{inj, limit}}$, that is recovered when $TS = -2.71$ in the presence of an injected signal with cross section $\langle \sigma v \rangle_{\text{inj}}$. If the statistical procedure is robust, the green band should lie above the diagonal line (saying that the limit set would be consistent with an injected signal) and should asymptotically approach the gold band for small signal cross sections, as is indeed the case for both masses included here.

The blue line shows the recovered cross section that is associated with the maximum test statistic, $TS_{\text{max}}$:

$$TS_{\text{max}} \equiv 2 \left[ \log \mathcal{L}(d|M, \langle \sigma v \rangle, m_\chi) - \log \mathcal{L}(d|M, \langle \sigma v \rangle = 0, m_\chi) \right],$$  

(3.8)
where $\langle \sigma v \rangle$ is the cross section that maximizes the likelihood. In the regime where $T_{\text{max}} \max < 1$, this is shown as a dashed line. The blue band corresponds to the range of cross sections above and below $\langle \sigma v \rangle$ associated with $T_{\text{max}} - 1$, spanning the extremal values of the middle 68% containment in each case. We expect that the recovered cross section should be consistent with statistical noise once the limit is reached, as is clearly demonstrated. The inset in each panel of Fig. 3.6 demonstrates how $T_{\text{max}}$ depends on the injected cross section.

While we show the representative cases for DM masses $m_\chi = 100$ and 1000 GeV here, we find that signal injection tests on Monte Carlo are well-behaved for DM masses ranging from 10–1000 GeV. The tests fail when the upper cutoff on the photon energy is $\gtrsim 100$ GeV most likely due to limited photon statistics. For this reason, as well as the fact that the \texttt{p6v11} template should be used with caution at energies $\gtrsim 50$ GeV, we have restricted the photon energies to be below $\sim 50$ GeV.

**Signal injection on data.** We also perform a data-driven version of the signal injection tests, adding a Galactic DM signal for $b\bar{b}$ annihilation on top of the actual data and passing this through the analysis pipeline. We repeat this procedure for 10 sky map realizations. This is a particularly important check at lower energies, where effects of point spread function (PSF) and foreground mis-modeling can lead to artificially strong limits for lower DM masses. Figure 3.7 summarizes the results of the signal injection tests on data for DM masses of 10 and 30 GeV (left and right panel, respectively). In each case, we see that the analysis would not exclude an injected DM signal. We restrict ourselves to energies $E_\gamma \gtrsim 0.8$ GeV to mitigate the effects of a significantly degraded PSF at even lower energies. We caution that while this procedure demonstrates that a signal would not be excluded under the null assumption on the data, it is still possible that mis-modeling effects
Figure 3.6: Signal injection tests on Monte Carlo simulations for a 100 (left) and 1000 (right) GeV DM particle annihilating to $b\bar{b}$. In each panel, the gold line corresponds to the limit $\langle \sigma v \rangle_{\text{null}}$ obtained when no signal is injected into the simulated data. The green line corresponds to the median cross section limit, $\langle \sigma v \rangle_{\text{inj}}$, that is recovered for a given injected cross section $\langle \sigma v \rangle_{\text{inj}}$, when $TS = -2.71$. The green band shows the corresponding 68% containment. The blue line corresponds to the median recovered cross section $\langle \sigma v \rangle$ that is associated with the maximum test statistic $TS_{\text{max}}$ (plotted in the inset), and is shown as dashed in the regime where $TS_{\text{max}} < 1$. The blue band spans extremal values of the 68% containment of cross sections associated with $TS_{\text{max}} - 1$. For each injected signal point, we create 50 realizations of simulated sky maps.

can impact the final result for the lowest masses ($\sim 10$ GeV). This can be seen in the left panel of Fig. 3.7 from the fact that the median recovered cross section associated with $TS_{\text{max}}$ (blue line) falls and becomes consistent with zero slightly above the null limit. However, this small discrepancy occurs in the range where $TS_{\text{max}} \lesssim 1$. This is not an issue for higher masses; for example, for a DM mass of 30 GeV, the median recovered cross section associated with $TS_{\text{max}}$ is consistent with zero only for cross sections below the null limit, as shown in the right panel of Fig. 3.7.

We have also performed signal injection tests using Model B to ensure the validity of the recovered bounds. The left panel of Fig. 3.8 shows the results of signal injection on data, as described in the previous section, for a DM mass of $m_\chi = 10$ GeV where
Figure 3.7: The same as Fig. 3.6, except for signal injected on data. The left(right) panel corresponds to a 10(30) GeV DM mass. In this case, for each injected signal point, we create 10 realizations of simulated sky maps.

Figure 3.8: Signal injection test on data using the Model B foreground template, assuming $m_\chi = 10$ GeV (left) and 70 GeV (right). Format as in Fig. 3.7.

foreground and PSF mis-modeling are likely to have the largest effect. We see that a putative DM signal would not be excluded by the analysis in this case. We also show results for an injected DM mass of $m_\chi = 70$ GeV in the right panel of Fig. 3.8, corresponding to the value most consistent with the excess emission seen in Models A and C. Again, we see that a potential DM signal would not be excluded in this case.
3.7 Extended Results

We consider several additional variations to the fiducial analysis, and summarize the results here:

- Although we presented results for DM annihilating into the $b\bar{b}$ and $\tau^+\tau^-$ final states in Sec. 3.3.1, DM annihilation can proceed into a variety of Standard Model final states. In Fig. 3.9 (left), we reinterpret the main results of the fiducial study in terms of annihilation into additional final states. Broadly, the spectra of hadronic channels ($W^+W^-, ZZ, q\bar{q}, c\bar{c}, t\bar{t}$) are predominantly set by boosted $\pi^0$ decays, resulting in comparable final limits beyond the respective mass thresholds. Gamma-rays for the leptonic ($e^+e^-, \mu^+\mu^-$) channels predominantly arise from radiative decays and final-state radiation, resulting in somewhat weaker overall limits. In each case, we assume 100% branching fraction into the specified channel. Note that we only model prompt gamma-ray emission and do not account for inverse-Compton or synchrotron radiation of the final state [109], which is relevant for the lighter leptonic channels.

- In addition to the $b\bar{b}$ and $\tau^+\tau^-$ cases considered in Sec. 3.3.3, we summarize in Fig. 3.9 (right) constraints on other possible annihilation channels contributing to the GeV excess. We show our results for the $q\bar{q}, c\bar{c}, gg$ and $hh$ final states, spanning the range $\gamma = 1.2$–1.3 for the inner slope of the NFW generalized profile (thick bands), along with the corresponding best-fit contours as found by [132] assuming $\gamma = 1.28$. We see that the $q\bar{q}$ and $hh$ explanations are robustly excluded by this analysis, while the $c\bar{c}$ and $gg$ explanations are put significantly under tension. We do not include annihilation channels that are already excluded at the 95% confidence level by spectral fits to the Fermi GeV excess emission [132].
Figure 3.4 demonstrates how the fiducial limit depends on the inner slope of the NFW profile. We have additionally considered the Einasto and Burkert profiles, defined in Eq. 3.6 and 3.7. The associated limits are shown in Fig. 3.10. The Einasto limit (solid green) is a factor of \( \lesssim 1.6 \) stronger than the fiducial case, while the Burkert limit is a factor of \( \lesssim 24(5) \) stronger/weaker for \( r_B = 0.5(10) \) kpc (dotted and dashed green, respectively).

We assumed a local DM density of \( \rho(r_\odot) = 0.4 \text{ GeV cm}^{-3} \) in the fiducial analysis, consistent with recent measurements \([68, 69]\). Other estimates in the literature, however, point to a value closer to \( \rho(r_\odot) = 0.3 \text{ GeV cm}^{-3} \) (see \([112]\) and references therein). Repeating the analysis using this lower value, we find that the limit is \( \lesssim 1.8 \) times weaker (solid blue line in Fig. 3.10). We emphasize that the assumption made about the local DM density does not impact the conclusions drawn about the viability of the GeV excess, as the best-fit regions are similarly shifted to higher annihilation cross sections by roughly the same amount.

Our fiducial analysis does not account for potential emission from Loop I in the Northern hemisphere. As a proxy for this contribution, we include an additional isotropic template in the Northern hemisphere. Modeling this emission results in a slight improvement in the DM constraint by a factor of \( \lesssim 1.2 \) (dashed purple line in Fig. 3.10), as expected because additional foreground components are accounted for.

In the fiducial study, the Northern and Southern lobes of the Fermi bubbles are floated separately. We have verified that floating the Northern and Southern lobes together leave the limit unchanged. Figure 3.10 shows what happens if the Fermi bubbles are not included at all in the analysis. In this case, the limit worsens by a factor of \( \lesssim 6 \) (solid gold line).
Figure 3.9: (Left) The 95% confidence limit on dark matter of mass, $m_\chi$, annihilating with cross section, $\langle \sigma v \rangle$, in the smooth Galactic halo. The limits are obtained following the fiducial analysis procedure described in Sec. 3.2, but varying over the annihilation channel. (Right) The 95% confidence limits on dark matter annihilation into $b\bar{b}$ (fiducial), $q\bar{q}$, $c\bar{c}$, $gg$, and $hh$, varying over the inner slope, $\gamma$, of the generalized NFW density profile. The bands correspond to $\gamma$ values spanning $1.2 - 1.3$. Note that the bands for $q\bar{q}$, $c\bar{c}$, and $gg$ fall essentially on top of each other. The best-fit parameters for the $q\bar{q}$, $c\bar{c}$, and $gg$ channels, as obtained in [132], are indicated by the pink, teal, and purple 1σ/2σ filled contours, respectively. The best-fit $hh$ value (and associated 1σ range) is indicated by the blue diamond [132].

- In the fiducial study, all point sources were masked to 95% containment in PSF, according to energy bin. To estimate the effect of point-source mis-modeling, we increased the mask size to 99% PSF containment; this results in a factor of $\lesssim 1.5$ weakening of the fiducial limit (solid purple line in Fig. 3.10), likely due to the corresponding reduction in the effective size of the ROI.

- The fiducial analysis takes full advantage of the spatial profiles of the expected DM emission and astrophysical components because we sum up the pixel-wise likelihoods. To quantify the gain from using spatial templates, we instead perform the fit using only the total expected number of counts from the DM signal and backgrounds within our ROI, and profile over the astrophysical nuisance
Figure 3.10: The 95% confidence limits associated with variations to the fiducial analysis, as labeled in the legend and described in the text.

parameters. The resulting limit (dotted gold line in Fig. 3.10) is several orders of magnitude weaker than the fiducial bound.

- We show results obtained using the newer \texttt{p7v6} and \texttt{p8R2} diffuse models in Fig. 3.11 (left and right panel, respectively). As outlined in Sec. 3.2, these models have large-scale residuals added back in to various extents, and as such are unsuitable for use in studying large-scale DM structures such as emission from the Galactic halo. Indeed, in both cases, we observe significant over-subtraction for the fiducial ROI. We emphasize that Fig. 3.11 is included for illustration only and should be treated with caution.

- Because the \textit{Fermi} bubbles are not accounted for when constructing the \texttt{p6v11} foreground model, one potential concern is the overestimation of the IC contribution. This could lead to inadequate modeling of the bubbles and potentially over-subtract a DM contribution, leading to an artificially strong limit. We show in Fig. 3.12 the energy spectra of the Northern (left) and Southern (right)
Figure 3.11: Similar to Fig. 3.1, except using the p7v6 and p8R2 foreground models (left and right panel, respectively) [188]. We only include these results for illustration as both of these foreground models are not appropriate for studies of diffuse DM signals, as discussed in the text.

lobes of the Fermi bubbles as obtained from our analysis pipeline when using the various foreground models presented here. The spectra recovered when using p6v11 are broadly similar to those obtained with Models A, B and C, underscoring the fact that the bubbles are adequately modeled in all four cases. We also show the bubbles spectra from [193], obtained for a slightly different ROI (|b| > 10° as opposed to |b| > 20°), which are again similar to those derived in our analysis.

• Given the importance of diffuse foreground modeling in the present study and the potential issues associated with a spectrally hard IC component in the p6v11 model [133], in Fig. 3.13 we show the total energy spectra obtained for the p6v11 model as well as those for Models A, B, and C in the eight radial slices considered in our study. We see that the spectra associated with p6v11 (black line) are roughly consistent with the total spectra associated with Models A, B and C (red, blue and purple lines respectively).
Figure 3.12: Recovered spectra, normalized to the corresponding bubbles region shown, for the Northern (left) and Southern (right) lobes of the Fermi bubbles when analyzed with diffuse model p6v11 as well as Models A, B and C. Our fiducial configuration was used to extract these spectra. The bubbles spectra obtained in [193] are shown for comparison. Note that a slightly different ROI (|b| > 10° as opposed to |b| > 20°) was used in that case. The energy $E_\gamma$ corresponds to the geometric mean of the energy bin edges.

- Figure 3.14 demonstrates the likelihood profiles for the fiducial analysis. In general, there is very good agreement between the observed profile (black line) and the Monte Carlo expectation (blue band), in each energy bin.
Figure 3.13: Energy spectra obtained for the p6v11 model (black) as well as those for Models A, B, and C (red, blue and purple respectively) in the eight radial slices (shown as insets) considered in our study.
Figure 3.14: Likelihood profiles for the fiducial analysis (black lines), presented for each energy bin. The darker(lighter) blue bands denote the 68(95)% containment for the profiles, as determined from 100 Monte Carlo simulations.
Chapter 4

Characterizing the Nature of the Unresolved Point Sources in the Galactic Center: An Assessment of Systematic Uncertainties

This chapter is based on work done in collaboration with Siddharth Mishra-Sharma, Mariangela Lisanti, Malte Buschmann, Nick Rodd, and Ben Safdi, published as Physical Review D 101, 023014 (2020) (arXiv:1908.10874). Section 4.1 provides background on the status of the Galactic Center Excess, summarizing works in the literature that motivated this study. Section 4.2 further motivates the philosophy behind this simulation-based study. Section 4.3 reviews the basics of the Non-Poissonian Template Fitting (NPTF) procedure, focusing specifically on the importance of the source-count function for the point sources (PSs) and its interpretation. We emphasize where the source-count function becomes potentially degenerate with Poissonian
emission, a crucial point for any NPTF study claiming to set constraints on the flux contribution of DM and PSs at the Galactic Center. Sections 4.4–4.6 present the results of NPTF tests on simulated data. Throughout, we emphasize the significance for these results in interpreting the results of signal injection tests on the Fermi data. Section 4.7 summarizes the main conclusions of the study while Sections 4.8–4.10 cover extended results: Section 4.8 includes additional figures that further illustrate the points of the main text; Section 4.9 discusses the residuals from diffuse mismodeling; and finally, Section 4.10 explores the effects of forcing the source-count function to zero below the flux near which the DM and PS contributions become degenerate. This work has been presented at the University of Utah High Energy and Astrophysics Seminar in Salt Lake City, UT (November 2019).

4.1 Foreword: A Closer Look at the Inner Galaxy

When Goodenough and Hooper published the first results on the Galactic Center Excess (GCE) in 2009 [120], they concluded that while the signal appeared tantalizingly similar to that expected from WIMP annihilation, they could not exclude the possibility that the excess had an astrophysical origin. In the ensuing years, a number of papers were written on the GCE [120, 122, 164, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137]. Some works focused on the DM interpretation of the excess, finding it to be consistent with a thermal DM particle in the ∼30–50 GeV mass range annihilating into $b$ quarks [126, 132, 133, 130, 129, 137]. Other works sought to explain the excess with astrophysical sources such as a population of unresolved millisecond pulsars [205, 206, 207, 208, 209, 210, 211, 129, 212, 213, 214, 215, 216]. However, within the literature there was a lack of strong statistical evidence in favor of either the dark matter or astrophysical point source hypotheses.
In 2015, two separate analyses found statistical evidence for unresolved PSs at the Galactic Center. One study applied the NPTF method to the Inner Galaxy, finding strong statistical preference for unresolved PSs over smooth emission, and suggesting that a novel population of millisecond pulsars could account for the GCE [118]. The other study utilized a wavelet decomposition method, and found evidence of small-scale structure that would be consistent with unresolved PSs as opposed to smooth emission. Since then, other studies have found strong statistical preference for the excess being correlated with the shape of the Galactic bulge and nuclear stellar bulge rather than the spherically symmetric morphology of a DM signal [166, 217, 167]. Additionally, there have been important consistency checks from complementary search targets which disfavor the DM interpretation of the GCE. For example, the stacked dwarf galaxy analyses from members of the Fermi-LAT collaboration exclude or are in tension with the DM interpretation of the GCE at the 95% confidence level for annihilation into $b$ quarks [148, 149]. Moreover, in the study presented in Chapter 3, my collaborators and I performed an analysis on the Milky Way halo at high latitudes, which lent itself to an apples-to-apples comparison with the GCE because unlike in the case of the dwarf galaxies, we were analyzing the same physical target. We found that the high latitude analysis similarly excluded the DM interpretation of the GCE at the 95% confidence level for annihilation into $b$ quarks.

Thus, until 2019, the status of the GCE could be summarized as follows: a statistically significant excess was found in the Fermi-LAT data in the Inner Galaxy. Within certain analysis frameworks, the morphology and spectrum of the excess is similar to that expected from thermal WIMP dark matter of mass $m_X \sim 30-50$ GeV annihilating into $b$ quarks, but a number of analyses found strong statistical preference for non-DM origins of the excess emission; additionally, analyses using complementary search targets excluded or were in tension with the regions of DM parameter space.
consistent with the GCE. It is important to emphasize that performing a robust analysis in the Inner Galaxy is challenging due to very bright Galactic diffuse emission in this region; therefore, any mismodeling of the diffuse emission can potentially have significant effects on the inference of dark matter or other astrophysical signals.

In 2019, a paper with the title “Dark Matter Strikes Back at the Galactic Center” [218] called into question the robustness of the original Inner Galaxy NPTF analysis [118], and rekindled excitement within the community for the DM explanation of the GCE. There were two main results presented in Ref. [218]:

1. The authors simulated data that consisted of two different signals: a smooth DM signal as well as a population of PSs tracing the Fermi bubbles. In addition, there was flux in the simulated data from diffuse emission, isotropic emission, emission from disk-correlated PSs, and smooth emission from the Fermi bubbles. In the analysis, they included (in addition to the relevant background templates) a smooth NFW template, which would ideally recover the simulated dark matter signal, as well as an NFW point source template. They did not include a point source template tracing the Fermi bubbles.

The goal of this exercise was to simulate a scenario in which there is a population of PSs in the data whose morphology is not accurately modeled. The authors found that in their setup, the smooth NFW template failed to recover the full DM flux that was present in the simulated data; the recovery was improved as they increased the strength of the DM signal in the simulated data, but was never perfect. While this example illustrates an interesting and important point—that signal inference can be biased by the residuals due to PS components not being accurately modeled—it is not necessarily specific to the NPTF method and more importantly is not likely to be a direct illustration of what is truly happening in the data analysis—indeed, the authors acknowledge
that this is only a proof-of-principle example, and in practice there is no evidence in the actual data that there are unmodeled point sources tracing the Fermi bubbles.

2. The authors then proceeded to artificially inject a DM annihilation signal onto the actual Fermi data, and found that the smooth NFW template failed to recover the full DM flux that was injected. Again, the recovery was improved as they increased the strength of the simulated DM signal injected onto the data, but was never perfect in the cases they tested. In addition, there was a large Bayes factor in preference for including an NFW point source template, indicating a statistical preference for NFW-distributed point sources, similar to the results presented in Ref. [118]. The authors thus argued, “[consequently], we conclude that dark matter may provide a dominant contribution to the GCE after all” [218].

One indication that diffuse mismodeling had a role to play in the results presented in Ref. [218] was that on the data itself (in the absence of an injected signal), if the normalization of the smooth NFW template was allowed to scan over negative values, then a negative normalization was statistically preferred. This is an indicator of the presence of oversubtraction due to diffuse mismodeling (this is similar to the situation in the toy example of oversubtraction presented in Sec. 2.5, where the best-fit normalization is smaller than the true normalization—in this case, even going negative). Furthermore, one might expect additional subtleties to arise when artificially injecting a signal onto data which already contains an excess, particularly in the presence of diffuse mismodeling. Therefore, when attempting to make a robust statement about the analysis results on the data, it is important to compare to the baseline expectation from analyzing Monte Carlo simulated datasets. While Ref. [218] did show the results from a single Monte Carlo trial corresponding to the signal injection test on
real data, demonstrating in that instance that the injected DM flux was fully recovered, my collaborators and I were interested in further understanding the systematic biases that might be present in the analysis and how to attempt to mitigate their effect.

In the first of two papers, we focused on simulated data and analyzed a number of simplified scenarios, varying over the fraction of the GCE flux which is comprised of emission from unresolved NFW-distributed point sources versus smooth dark matter emission, as well as varying the luminosity function of the NFW point sources [219]. One of the key takeaways from our study was that the fundamental degeneracy between ultrafaint point sources and smooth emission following the same spatial morphology, combined with the effect of diffuse mismodeling, can bias the results of an NPTF analysis such that emission from dark matter is reconstructed as emission from unresolved point sources or vice versa. Importantly, we found that there were large variations in the results across different Monte Carlo trials, making it difficult to make a conclusive statement about results on the real data without significant improvements to the analysis procedure. We also confirmed that in certain cases, the presence of residuals due to diffuse mismodeling could lead to strong statistical preference for point sources—this is because the method is agnostic to whether bright clumps of emission come from true astrophysical point sources or bright residuals. This paper (with slight modifications for presentation purposes) is presented in full in the following sections of this chapter, and a more complete list of conclusions of the study can be found in Section 4.7.

In the second paper, we set out to improve how we modeled the diffuse emission in our analysis framework [220]. The improvement to the diffuse modeling was two-fold: firstly, we made use of a new Galactic diffuse model which gave a significantly better fit to the data than any of the previously considered models (referred to as “Model O”
in our terminology); and secondly, we performed a spherical harmonic decomposition of the diffuse model and marginalized over the coefficients of the low-\(\ell\) modes in order to mitigate large-scale mismodeling effects. With these two improvements, after verifying the robustness of the procedure on Monte Carlo simulations, we found that the analysis consistently and stably recovered evidence for point sources.

Thus, after implementing drastic improvements to the analysis procedure since the original NPTF analysis from Ref. [118], there is still strong statistical evidence that there is point source emission in the Inner Galaxy. This emission could be due in part to astrophysical unresolved point sources, or be in part arising from more subtle biasing effects that are not accounted for in the current framework. At present, there is no strong statistical evidence for emission from dark matter annihilation in the region, although it is important to note that there is a fundamental limitation to how well one can distinguish smooth emission versus the signal coming from a population of ultrafaint point sources following the same spatial morphology.\(^1\)

### 4.2 Introduction

The observed excess of GeV gamma-rays at the center of the Milky Way has withstood many tests over the course of the last decade [120, 122, 164, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137]. Referred to as the Galactic Center Excess (GCE), the energy spectrum and morphology of the excess as observed by the *Fermi* Large Area Telescope [103] are consistent with a signal of dark matter (DM) annihilation [126, 132, 133, 130, 137]. However, astrophysical sources—such

\(^1\) Another recent paper, Ref. [221], used the wavelet decomposition method in conjunction with an updated resolved point source mask coming from the new *Fermi* 4FGL catalog [114]. They found decreased evidence for small-scale structure compared to Ref. [165]. However, it is unclear at this time whether this decrease in statistical evidence is due to the significantly decreased size of the region-of-interest in going to the new resolved point source mask, an effect which could be understood and characterized by analyzing Monte Carlo datasets; Monte Carlo tests were absent from the paper, so the robustness of the analysis is yet to be determined.
as a population of unresolved millisecond pulsars—may also explain the signal [205, 206, 207, 208, 209, 210, 211, 129, 212, 213, 214, 215, 216]. Characterizing the nature of these potential sources, either through direct discovery or indirect statistical tests, is of paramount importance in establishing the viability of the DM hypothesis.

Two separate but complementary studies have argued for evidence of unresolved point sources (PSs) at the Galactic Center. The first method, referred to as Non-Poissonian Template Fitting (NPTF), used the statistics of fluctuations in photon counts to demonstrate evidence for an unresolved PS population in the Inner Galaxy [118]. The second study performed a wavelet decomposition of the gamma-ray sky and found evidence of small-scale structure consistent with a population of unresolved PSs rather than smooth emission from DM [165]. Since then, the case for unresolved PSs has continued to be strengthened by studies suggesting that the shape of the excess is correlated with the stellar overdensity in the Galactic bulge and the nuclear stellar bulge, a scenario strongly preferred over spherically-symmetric emission from DM [166, 217, 167].

The exact nature of these unresolved PSs continues to remain a mystery, however. Both the NPTF and wavelet methods are only sensitive to the spatial distribution of PSs and, in the case of the NPTF, their luminosity function, but are otherwise model-independent. These sources can be actual astrophysical PSs, such as millisecond pulsars [222, 223, 213, 214], residual structure due to mismodeling of cosmic-ray background emission [224, 225], or even DM substructure [226]. The statistical analyses are themselves agnostic to these possibilities. In lieu of a direct discovery of these sources [227], we can only hope to infer their properties indirectly.

In this chapter, we study the ability of the NPTF method to characterize the flux contribution of PSs to the GCE as well as their source-count distribution. A source-count distribution describes the number of sources of a given flux, and there-
fore encodes information about the relative number of bright and faint sources in a population. When applied to spatially binned (pixelated) photon count data, the NPTF procedure distinguishes PSs from smooth Poissonian emission based on the number of photons that each PS produces, distributed across a number of pixels due to the finite spatial resolution of the telescope. A population of PSs will typically yield more “hot” and “cold” pixels relative to a smooth Poissonian component.

Using simulated data, we consider scenarios where the GCE is comprised entirely of DM or PSs, as well as cases where the GCE flux is divided between the two. In each case, we study how reliably the NPTF recovers the correct composition of the GCE. Our work here builds on previous studies of the NPTF on simulated data [118], which only considered PSs with a source-count distribution matching that recovered on data. This empirically-motivated distribution described a population of reasonably bright unresolved sources. We now present a more systematic study on simulated data that carefully considers different source-count functions, focusing on cases where more ultra-faint sources are present, as is expected for a population of millisecond pulsars (MSPs) [206, 210, 209, 215, 214, 213, 216]. This enables us to characterize any potential biases of the statistical method that can shift the recovered source-count function away from its true distribution.

When the backgrounds are perfectly modeled, we find that the NPTF always accurately identifies the origin of the signal in the case where the GCE consists entirely of DM. If, instead, 100% of the GCE consists of PSs, then some fraction of the PS flux can be misattributed to DM. When the GCE is comprised of contributions from both DM and PSs, the NPTF can misidentify the flux as belonging entirely to DM or entirely to PSs. The challenge associated with reconstructing the correct fraction

\footnote{Note that, throughout this work, when we refer to the NPTF we also implicitly refer to the standard parametrization of the source-count distribution and the associated priors, which are described in Sec. 4.3. It is possible that different parametrizations of these distributions within the framework of the NPTF would give different results to those presented here.}
of DM when PSs are also present in the data stems from the basic fact that in the ultra-faint limit, PSs are exactly degenerate with smooth Poissonian emission. This degeneracy can lead to biases when inferring the proportion of PSs and DM preferred by the data, which we explore in detail. These biases are tempered for a population of (relatively) bright unresolved sources, as they are easier to distinguish from smooth Poissonian emission with the same spatial distribution. This is the scenario that had been studied previously in Ref. [118].

When the Galactic diffuse backgrounds are mismodeled, as expected in any analysis on actual Fermi data, additional challenges arise. We mock up this scenario by creating simulated data with one diffuse model and analyzing it with a different diffuse model, using this setup to show that bright residuals from mismodeling can be absorbed as point sources in the analysis. This may explain why the source-count distribution recovered on data in Ref. [118] is different from that expected for MSPs. When 100% of the GCE flux is in DM, the statistical preference for PSs is significantly reduced relative to the strong preference recorded when the GCE flux is entirely in PSs. However, we identify some instances where the DM can be misidentified as PSs with reasonable statistical confidence. For the particular pair of diffuse models we use, we find that the significance for PSs varies strongly depending on whether the “correct” or “incorrect” model is assumed in the NPTF analysis, a strong indication that the NPTF is picking up residuals from diffuse mismodeling as PSs. These findings motivate a detailed study of ways to mitigate the effects of diffuse mismodeling in NPTF analyses on data. We present these results separately in a companion paper [220].

Lastly, our work enables us to comment on a recent study that draws doubt on the PS interpretation of the GCE [218]. In reaching this conclusion, the authors inject an artificial DM signal on the Fermi data, pass it through the NPTF pipeline, and find
that the injected DM signal is misattributed to PS flux. In general, signal recovery tests on data can be quite valuable—as we have shown in separate studies on Fermi data [162, 143]. However, interpreting the results must be done with great care, especially in the case when a true signal (either DM or PSs) is present in the data. We demonstrate using simulated data that, within the current NPTF framework, misattribution of injected DM flux to PSs can be a natural consequence of the fact that smooth Poissonian emission is exactly degenerate with a population of ultra-faint PSs; from the perspective of the NPTF, the artificially injected DM signal can either be its own separate Poissonian contribution, or it can be flux that “fills in” the ultra-faint end of the source-count distribution for PSs with the same spatial distribution as the DM.\(^3\) It is thus unsurprising that the NPTF does not recover the injected DM signal if PSs are already present in the data. We also show that the recovery of the injected DM signal can be significantly worsened (\(i.e.,\) an appreciable fraction of the DM signal is misattributed in more realizations) in the presence of diffuse mismodeling, an irreducible effect on real Fermi data. Additionally, on the real data, the injected signal recovery could further be compounded by complications from diffuse mismodeling that are not captured by the simulations studied here, or by other populations of unmodeled PSs. We show that such misattribution of an injected DM signal does not by itself point to issues with an NPTF analysis of the underlying map that does not contain an injected signal, and thus conclude that the signal injection tests performed on data in Ref. [218] are not by themselves indicative of an issue with the original NPTF analysis. We take a pedagogical approach in this work in order to help the reader build intuition for interpreting the output of an NPTF analysis.

\(^3\)In principle, one can construct an analysis framework that is indiscriminate between the DM and PS hypotheses—rather than attribute the flux to one component or the other—in the degenerate regime. However, this is beyond the scope of the work presented here.
4.3 Statistical Methodology

This work uses simulated data to better characterize the ability of the NPTF procedure to recover the properties of the unresolved GCE PSs. We will start from simple maps that only contain PSs, and build up to include diffuse emission, DM, and other non-PS components. In this way, we will clearly see how the recovery of the PS and DM fractions is affected as the simulated maps become increasingly more realistic. This section reviews the NPTF procedure and describes how the maps are made.

4.3.1 NPTF Procedure

We assume that the data can be described by a set of different gamma-ray components, each with its own specified spatial distribution. Each component is modeled by a “template” that traces its spatial morphology. Some of the templates in the study model smooth Poissonian emission, while others trace populations of unresolved PSs that are described by non-Poissonian statistics. Consider a spatially binned data map \( d \) that consists of \( n_p \) photon counts in pixel \( p \). For a given model \( \mathcal{M} \) with free parameters \( \theta \), the likelihood function is defined as

\[
p(d | \theta, \mathcal{M}) = \prod_p p_{n_p}^{(p)}(\theta),
\]

where \( p_{n_p}^{(p)}(\theta) \) is the probability of observing \( n_p \) photons in pixel \( p \) for the assumed model. In the Poissonian case, the templates—which are spatially binned in the same way as the data—predict the mean expected number of counts \( \mu_p(\theta) \) in pixel \( p \):

\[
\mu_p(\theta) = \sum_l \mu_{p,l}(\theta),
\]
where \( l \) is the index over Poissonian templates. \( \mu_p(\theta) \) is fully specified by the overall normalizations of the templates. In this case, \( p^{(p)}_{np} (\theta) \) in Eq. 5.14 is simply given by the Poisson probability of observing \( n_p \) photons given the expected number of counts:

\[
p^{(p)}_{np} (\theta) = \frac{\mu_p(\theta)^{n_p}}{n_p!} e^{-\mu_p(\theta)}.
\]  \( (4.3) \)

When modeling unresolved PSs, however, \( p^{(p)}_{np} (\theta) \) is non-Poissonian. The reason for this is that one must first ask what the probability is that a PS is in pixel \( p \) and then ask what the probability is that it contributes \( n_p \) photons to the data (modulo corrections for a finite point-spread function, which will be discussed below).

When modeling the non-Poissonian templates, an essential input is the flux distribution of the sources. To aid the calculation, this is typically parameterized as a multiply-broken power law. In the first iteration of the NPTF method from Ref. [118], a singly-broken power law was used, but additional breaks can allow for greater flexibility in the recovery of the underlying PS flux distribution. In this work, we use a two-break model to describe how the number of sources \( N \) is distributed with photon count \( S \):

\[
\frac{dN}{dS} = A^{PS} \begin{cases} 
\left( \frac{S}{S_{b,2}} \right)^{-n_3} & S < S_{b,2} \\
\left( \frac{S}{S_{b,2}} \right)^{-n_2} & S_{b,2} \leq S < S_{b,1} \\
\left( \frac{S_{b,1}}{S_{b,2}} \right)^{-n_2} \left( \frac{S}{S_{b,1}} \right)^{-n_1} & S_{b,1} \leq S
\end{cases}
\]  \( (4.4) \)

where \( S_{b,1\ldots2} \) are the breaks, \( n_{1\ldots3} \) denote the power-law indices, and \( A^{PS} \) is the overall normalization. The photon count \( S \) is related to the flux \( F \) through the equation \( S = \langle \epsilon \rangle F \), where \( \langle \epsilon \rangle \sim 6.59 \times 10^{10} \text{ cm}^2 \) is the mean exposure per pixel for the dataset under consideration. Note that, for computational simplicity, we consider a flat exposure map with the value in every pixel equal to the mean Fermi-LAT exposure in the
relevant energy range. The effect of non-uniform exposure can be corrected using the procedure described in Ref. [138], and would not affect the conclusions of our study. In general, such corrections require that Eq. 4.4 be written in terms of flux, with the translation to counts occurring on a pixel-by-pixel basis.

It is important to emphasize that the shape of the flux distribution is a critical assumption of the method, and an inherent systematic uncertainty. The choice of the doubly broken power law is useful as it provides sufficient freedom to capture known features in the distribution. For example, the upper break \( (S_{b,1}) \) corresponds roughly to the threshold of resolved sources, when they are masked. For Fermi, we take this to be the threshold for the third source catalog (3FGL) [113].\(^4\) The lower break \( (S_{b,2}) \) corresponds to the region where the method starts to lose sensitivity.\(^5\) As discussed further below, ultra-faint sources are inherently degenerate with smooth Poissonian emission. In this low-flux regime, we expect that the NPTF analysis will struggle to distinguish PSs from smooth emission.

The NPTF method was discussed in depth in Refs. [117, 118, 138], and we refer the reader to those works for a full review and technical details of algorithms used. Here, we provide basic pertinent information relevant to this study. The NPTF likelihood is most conveniently cast in the language of probability generating functions, also known as moment generating functions, following Ref. [119]. For a discrete probability distribution \( p_k \), with \( k = 0, 1, 2, \ldots \), the generating function is defined as

\[
P(t) \equiv \sum_{k=0}^{\infty} p_k t^k,
\]

\(^4\)Although the fourth source catalog has recently become available [114], we use the 3FGL in our study to motivate comparison with previous work. Using the updated catalog would not affect the conclusions presented here since we restrict ourselves to studying simulated data.

\(^5\)When this is not the case, an additional break is often useful as it allows the model to capture features in the source-count distribution at intermediate fluxes, as was done in Ref. [228].
from which the probabilities can be recovered by taking successive derivatives:

\[ p_k = \frac{1}{k!} \left. \frac{d^k P(t)}{dt^k} \right|_{t=0}. \] (4.6)

The generating function for a Poissonian template is given by

\[ P_p(t; \theta) = \prod_p \exp \left[ \mu_p(\theta)(t-1) \right]. \] (4.7)

The generating function for a PS template takes the more complicated form

\[ P_{NP}(t; \theta) = \prod_p \exp \left[ \sum_{m=1}^{\infty} x_{p,m}(\theta)(t^m - 1) \right], \] (4.8)

where

\[ x_{p,m}(\theta) = \int_0^{\infty} dS \frac{dN_p}{dS}(S; \theta) \int_0^1 df \rho(f) \frac{(fS)^m}{m!} e^{-fS}. \] (4.9)

The \( x_{p,m} \) can be interpreted as the average number of PSs contributing \( m \) photons in expectation within the pixel \( p \), given the pixel-dependent source-count distribution \( dN_p(S; \theta)/dS \). When \( m = 1 \), the functional form of Eq. 4.8 reduces to that of Eq. 4.7 (see Ref. [138] for more details). This demonstrates that a Poissonian component (such as DM or diffuse emission) can be thought of as a population of single-photon sources with the same spatial distribution. Using the property that the generating function of a sum of several random variables is the product of the individual corresponding generating functions, we can write the total generating function in our case as the product of the separate Poissonian and non-Poissonian contributions.

For a “spatially-averaged” source-count distribution, e.g. Eq. 4.4, an overall pixel-dependent prefactor in \( dN_p(S; \theta)/dS \) modulates the expected number of PSs in each
pixel $p$ following the assumed spatial distribution of PSs specified by the template. Additionally, $\rho(f)$ is a function that describes the distribution of flux fractions among pixels, accounting for photon "leakage" due to the finite point-spread function (PSF) of the instrument. By definition $\int_0^1 df \rho(f)$ equals the number of pixels that, on average, contain between $f$ and $f + df$ of the flux from a PS; the distribution is normalized such that $\int_0^1 df \rho(f) = 1$. In the absence of this effect, the PSF is a $\delta$-function and $\rho(f) = \delta(f - 1)$. For a given PSF model, $\rho(f)$ is obtained through a Monte Carlo procedure. For more details on Eqs. 4.7–4.9 and the NPTF algorithm generally, as well as details of instrument PSF and exposure (i.e., scanning strategy) correction, see Ref. [138].

We use the public code NPTFit [138] to implement the NPTF procedure. This is interfaced with MultiNest [229, 230] (with $n_{\text{live}}=500$), which implements the nested sampling algorithm [231, 232, 233], to efficiently scan the (potentially multi-modal) posterior parameter space associated with the Poissonian normalizations and non-Poissonian source-count parameters in a Bayesian framework. This necessitates a specific choice of prior probabilities on each parameter, summarized in Table 4.1.

In all of the results presented and described in this chapter, we have discarded scans for which the upper break $S_{b,1}$ of the NFW PS template is peaked at the lower edge of its prior range. This is because in such cases, the recovered source-count function peaks in precisely the flux regime where the PSs are degenerate with smooth Poissonian emission. Any such source-count function recovered on data would be suspicious as it would suggest a population of ultra-faint sources that is concentrated

\[\text{In practice, we quantify this convergence test by histogramming the NFW PS } S_{b,1} \text{ posterior for a single run into 50 log-spaced bins from } 10^{0.5} \text{ to } 10^2, \text{ and require that either the counts in the first bin do not exceed 0.4 times the maximum counts or the counts in the last bin are no lower than 0.2 times the maximum counts. The former criterion ensures that the peak of the distribution is not pushed against the lower prior edge, and the latter allows for a posterior distribution that is unconverged over the prior range, as is expected for e.g. the case of no PS contribution. Example triangle plots of scans that pass these criteria can be found in Figs. 4.15 and 4.16; a failing example can be found in Fig. 4.17.}\]
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Parameter</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{\text{diff}}$</td>
<td>[0, 20]</td>
<td>$\log_{10} S_{b,1}$</td>
<td>[0.5, 2.0]</td>
</tr>
<tr>
<td>$A_{\text{iso}}$</td>
<td>[0, 2]</td>
<td>$\log_{10} S_{b,2}$</td>
<td>[-2.0, 0.5]</td>
</tr>
<tr>
<td>$A_{\text{bub}}$</td>
<td>[0, 2]</td>
<td>$n_{1}$</td>
<td>[2.05, 15]</td>
</tr>
<tr>
<td>$A_{3\text{FGL}}$</td>
<td>[0, 2]</td>
<td>$n_{2}$</td>
<td>[-3.95, 2.95]</td>
</tr>
<tr>
<td>$\log_{10} A_{\text{NFW}}$</td>
<td>[-5, 2]</td>
<td>$n_{3}$</td>
<td>[-10, 0.95]</td>
</tr>
<tr>
<td>$\log_{10} A_{\text{NFW}}^{\text{PS}}$</td>
<td>[-10, 5]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Fiducial priors assumed for the NPTF analyses in this work. The left column lists the priors for the template normalizations. From top to bottom: diffuse, isotropic, $Fermi$ bubbles, $Fermi$ 3FGL sources, NFW dark matter, and NFW point sources. The right column lists the priors on the source-count parameters, as in Eq. 4.4. A flat prior distribution between the specified parameter ranges is assumed.

in the regime where the NPTF method loses sensitivity. Fig. 4.17 shows the triangle plot from an example of a discarded scan. In Section 4.10, we consider what happens when the source-count function is forced to zero below the flux near which the method loses sensitivity for the aforementioned reason. We find that imposing such a flux cutoff reduces the number of cases that would be removed using this procedure.

### 4.3.2 Simulated Data Maps

The region of interest (ROI) in our analysis is $|b| \geq 2^\circ$, $r < 30^\circ$, and we mask the resolved $Fermi$ 3FGL PSs [113] at a 0.8\textdegree radius. This is essentially the same setup as was used in Refs. [118, 218]. We use the datasets and templates from Ref. [139] (packaged with Ref. [138]) to create our simulated maps. The data corresponds to 413 weeks of $Fermi$-LAT (Large Area Telescope) Pass 8 data collected between August 4, 2008 and July 7, 2016. The top quarter of photons in the energy range 2–20 GeV by quality of PSF reconstruction (corresponding to PSF3 event type) in the event class ULTRACLEANVETO are used. The recommended quality cuts are applied, corresponding
to zenith angle less than 90°, LAT_CONFIG = 1, and DATA_QUAL > 0.1.\textsuperscript{7} The maps are spatially binned using HEALPix \cite{2019arXiv190504090C} with nside = 128.

We build up the simulated maps as a combination of Poissonian and PS contributions, as necessary. A PS population is completely specified by a spatial distribution and a source-count distribution, in our case parameterized with Eq. 4.4. We draw the fluxes of simulated PSs from the source-count function, using it as a probability density. The total number of simulated PSs is determined by the desired total flux contribution from the PS population. We then put the simulated PSs down on a higher-resolution HEALPix map with nside = 2048, using the template describing the spatial distribution of PSs as a probability density informing the locations of PSs. The PS map is smoothed with the Fermi PSF at 2 GeV, modeled as a linear combination of King functions,\textsuperscript{8} and downgraded to the baseline nside = 128. A Poisson realization of this downgraded map then represents a single Monte Carlo simulated map.

In this work, we model the known astrophysical emission as a sum of Poissonian templates, which include, \textit{e.g.}, \textit{(i)} the Galactic diffuse emission, described by the Fermi gll\_iem\_v02\_P6\_V11\_DIFFUSE (p6v11) model\textsuperscript{9} \textit{(ii)} the Fermi bubbles \cite{2015ApJ...804..130A}, \textit{(iii)} isotropic emission, and \textit{(iv)} resolved Fermi 3FGL PSs \cite{2015ApJS..218...17O}. For each template, the normalization is the best fit obtained with a traditional Poissonian template fit on Fermi data in the ROI. The final set of maps is obtained by creating a linear combination of a (sub)set of these templates as a baseline, and subsequently Poisson fluctuating to obtain multiple Monte Carlo realizations. In addition, whenever applicable, we model the DM(PS) GCE emission as a Poissonian(non-Poissonian)

\textsuperscript{7}https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Data_Exploration/Data_preparation.html
\textsuperscript{8}https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_LAT_IRFs/IRF_PSF.html
\textsuperscript{9}https://fermi.gsfc.nasa.gov/ssc/data/access/lat/ring_for_FSSC_final4.pdf
template following the line-of-sight integrated square of an NFW distribution \cite{234}. We perform a separate Poissonian template fit including the astrophysical emission templates (i)–(iv), with the addition of the Poissonian NFW template, to determine the best-fit GCE flux. In Sec. 4.4, we will start with just a simulated PS contribution accounting for the entirety of the GCE flux and build up to increasingly more realistic scenarios by adding astrophysical background templates. We then consider more complicated scenarios where the GCE might consist of flux contributions from PSs and DM. Approaching our study from a pedagogical angle, we will also initially consider the Galactic diffuse emission as the sole tracer of astrophysical emission (and use a single Poissonian template, the Fermi p6v11 model, to describe it), then incorporate the effect of additional Poissonian templates later.

4.4 Anatomy of a Source-Count Function

When studying the ability of the NPTF to distinguish PSs from smooth emission at the Galactic Center, one must know something about the properties of those sources—both their spatial and flux distribution—as well as the average photon count per pixel that is expected for other gamma-ray sources that could be degenerate with the PS signal. In this work, we will only consider PSs whose emission traces the square of an NFW distribution.\footnote{In practice, we treat the line-of-sight integrated NFW squared map as the number density distribution of the PSs, from which we draw the positions of simulated sources.} This is intended to match the spatial distribution of the GCE flux as characterized in Ref. \cite{130, 133}.\footnote{The study could also be repeated assuming a bulge-shaped template (as in Ref. \cite{166, 217, 167}) for the DM and PSs. As the results of this work are mostly driven by the fact that the DM and PSs share the same spatial distribution, we do not expect that the overall conclusions would be significantly altered in this case. However, the finer details of the recovered fit parameters would likely change.}

We consider two benchmarks for the PS flux distribution: a hard source-count distribution where the distribution of PSs is peaked towards high fluxes, and a softer
source-count distribution with a larger number of faint sources. The hard source-count function is generated with a singly-broken power law, with parameters: $S_{b,1} \approx 15$, $n_1 \approx 9.5$, $n_2 \approx -1$. This benchmark is motivated by previous NPTF studies of the Galactic Center and roughly matches the function recovered in the data [118]. The soft source-count distribution is generated with a two-break power law, with parameters: $S_{b,1} \approx 22$, $S_{b,2} \approx 0.2$, $n_1 = 10$, $n_2 = 1.9$, $n_3 = -0.8$ and is motivated by the luminosity function for disk MSPs obtained in Ref. [216] and assuming the energy spectrum found in Ref. [215], normalized to account for 100% of the GCE flux (gray dash-dotted line). The soft source-count distribution is a reasonable approximation to the MSP source-count distribution, while the hard source-count distribution significantly underpredicts the number of dim PSs.

\footnote{When using a two-break power law in the fit to data, an essentially equivalent function is returned.}
nosity function of MSPs estimated in the literature [216, 209, 215, 214, 206, 210, 213], which tend to be softer than that inferred in Ref. [118]. The black points in Fig. 4.1 illustrate these two cases for simulated maps that only include PSs (and no other contributions), for the hard (left) and soft (right) source-count functions. The spatially averaged 3FGL flux threshold is approximately $4-5 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$—i.e., we assume that all sources above this flux value would be resolved by Fermi.$^{13}$ We also show the line corresponding to $\sim 1 \text{ ph}$. Because we can think of Poissonian emission as a combination of single-photon sources, this represents the approximate flux boundary below which Poissonian and non-Poissonian emission are fundamentally degenerate. In the case of the hard (soft) source-count function, about 0.02 (34)% of the emission falls below the $\sim 1 \text{ ph}$ line.

Running the NPTF pipeline on these simulated maps, we can test how well the analysis recovers the properties of the simulated sources in the very simple case where the map consists only of NFW PSs. We include three templates in the model: (i) NFW PSs, (ii) NFW DM, and (iii) Galactic diffuse emission. This will allow us to verify that the PS emission is predominantly picked up by the appropriate non-Poissonian template, and characterize any possible degeneracies with the other two templates. Figure 4.1 provides the best-fit source-count function (solid red line) that is recovered by the NPTF analysis for the map with hard (left) and soft (right) sources; the red bands span the 68 and 95% containment. Above the $\sim 1 \text{ ph}$ threshold, the source-count function is recovered exactly for both benchmark scenarios. Below this threshold, the uncertainty on the recovered source-count function increases for the soft sources, as this is the regime where the sources are so ultra-faint that their photon counts are effectively Poissonian.

$^{13}$This threshold was conservatively estimated as the flux corresponding to the peak of the source-count distribution $F^2 dN/dF$ associated with the sources in the 3FGL catalog in the region of interest.
As a point of reference, we also show in Fig. 4.1 the source-count distribution derived using the median (log-normally parameterized) disk MSP luminosity function obtained in Ref. [216], assuming the MSP energy spectrum found in Ref. [215] and an NFW-squared spatial distribution for the MSP population. We see that our simple model of soft PSs is a reasonable approximation to the MSP scenario, while the hard source-count distribution motivated by previous NPTF studies significantly underpredicts the number of dim sources.

Next, we test the source-count recovery over several realizations of the simulated data maps. In particular, we re-run the analysis for 100 Monte Carlo iterations of each map, and take the best-fit source-count function from each run. The solid yellow line in the top-left panel of Fig. 4.2 shows the median best-fit over these separate iterations for the soft source-count function. The yellow bands show the median 68 and 95% containment regions. Importantly, these uncertainty bands are fundamentally different than those shown in Fig. 4.1, as they summarize the average uncertainty over multiple realizations of the simulated map. Comparing Fig. 4.1 (right) with the top-left panel of Fig. 4.2, we see that the specific case shown in Fig. 4.1 is not an outlier over the the distribution of many realizations.

The bottom-left panel of Fig. 4.2 shows the cumulative PS flux above a given threshold, compared to the total injected PS flux, as a function of decreasing flux threshold. The black points represent the true distribution, and the yellow bands show the results recovered by the NPTF. The cumulative flux distribution is useful for inferring the fraction of the injected PS flux that is correctly recovered by the NPTF procedure. Ideally, the fractional cumulative flux distribution should track the true distribution, asymptoting to unity at low fluxes. Deviations from unity indicate that the PS template has over/under-absorbed flux relative to the truth.
GCE = 0% NFW DM + 100% NFW PS

Figure 4.2: Differential source-count distributions (top panels) and cumulative flux distributions, integrated above a given threshold flux (bottom panels), shown for simulations with PSs following a soft source-count distribution, accounting for 100% of the GCE (no DM contribution). Black points indicate the true simulated distributions while solid lines indicate the median best-fit distributions recovered over 100 Monte Carlo realizations of the simulated data maps. The bands show the median 68 and 95% confidence bands over these 100 iterations. In order, the results are shown without including PSF effects in the simulation or NPTF analysis and without Galactic diffuse emission in the simulated data (left column); without PSF effects but including Galactic diffuse emission (middle column); and finally, accounting for PSF effects while also including Galactic diffuse emission (right column). In the left column, the dotted lines designate the photon counts associated with a given flux; in the middle and right columns, the lines designate the approximate significance of a source with flux $F$. Note that the 3FGL threshold corresponds roughly to $\sim 5\sigma$ sources. The individual flux posteriors for a random subset of 50 out of the 100 realizations summarized in the right panel are provided in Fig. 4.11. All subsequent figures in this chapter include Galactic diffuse emission in the simulated data maps, as well as PSF effects in both the simulated data and the NPTF analysis.

We now add Galactic diffuse emission to the simulated maps. Specifically, we include Poissonian emission from the p6v11 model in addition to the NFW PSs and repeat the NPTF studies described above. The median recovered source-count distribution is provided in the middle column of Fig. 4.2, in blue. The vertical dotted lines
denote the approximate significance ($\sigma$) of the PSs at a given flux $F$. To estimate the significance, we calculate $\sigma = S/\sqrt{B}$, where $S$ and $B$ are respectively the average photon count per pixel for the signal and the diffuse background in the ROI (in our case the pixel size roughly matches the extent of the PSF). We provide the lines for $\{1, 3, 5\} \sigma$. The 3FGL threshold corresponds roughly to a significance of $5\sigma$ for each source.

The presence of diffuse emission results in a greater spread of the recovered source-count distribution relative to the scenario with no diffuse emission. In this case, the median source-count distribution recovered by the NPTF analysis resembles a harder population of sources. As already discussed, a population of ultra-faint sources becomes indistinguishable from Poisson emission, so we expect the uncertainties to grow at low fluxes. This occurs for sources with fluxes $\lesssim 4-5 \times 10^{-11}$ counts cm$^{-2}$ s$^{-1}$. While the recovered distribution is consistent with truth at the $\sim 95\%$ level at these low fluxes, the median best-fit is systematically lower than the true source-count distribution, which indicates that the analysis is biased against PS recovery (we observe this even in the absence of diffuse emission). While a comprehensive study of the source of this bias is beyond the scope of this work, we comment that it may be related to the parameterization of the source-count function as a multiply-broken power law and/or the choice of priors.

Lastly, we repeat the same tests, but now smear the simulated map with the Fermi PSF function at 2 GeV—modeled as a double King function—and perform the NPTF with the correction described by Eq. 4.9. The results are summarized in the right column of Fig. 4.2, in green. We see that the inference with the PSF is further biased against recovery of PSs at the faint end of the source-count distribution. This is intuitively expected, as the inclusion of the PSF exacerbates the challenges with recovering the faintest sources. When the PS flux is underestimated, the flux is picked
up by the DM template or the diffuse template. The amount of flux that is erroneously attributed to these other templates can vary considerably between different Monte Carlo iterations of the map, as demonstrated in Fig. 4.11.

The degeneracy between ultra-faint sources and Galactic diffuse emission is primarily due to the fact that the total flux of the diffuse emission is $\gtrsim 30$ times greater than the GCE flux in this ROI. As a means of testing this hypothesis, we ran an analysis scaling down the diffuse emission flux by an order of magnitude. In this case, the recovery of the PS source-count function is significantly improved relative to the fiducial scenario presented in the right panel of Fig. 4.2. As a separate test, we replaced the diffuse emission with isotropic emission (of equivalent flux magnitude) in the map. In this latter case, the recovered source-count distributions look very similar to the right panel of Fig. 4.2. We therefore conclude that the degeneracy between ultra-faint unresolved sources and the diffuse emission is primarily driven by the brightness, rather than the morphology, of the diffuse emission. This degeneracy is fundamental to the analysis method in this region of the sky. We further note that the diffuse emission is spatially structured towards the Galactic Center region due to gas-correlated emission, which, if not modeled properly, can potentially lead to residual clusters of “hot” or “cold” pixels, as one would also get from a population of PSs. We discuss the effects of diffuse mismodeling in more detail in Sec. 4.6.

We only show the results of these tests for the soft source-count distribution. In general, we find that the hard source-count distribution is unaffected by the types of degeneracies we see here, primarily because there are more unresolved sources with high photon counts that are more easily distinguishable from diffuse background emission.
Figure 4.3: Comparison of the flux posterior (relative to the true injected flux) for the Galactic diffuse, NFW DM, and NFW PS components. The last column shows the Bayes factors (BFs), characterizing the statistical preference of a model of the data that includes NFW PSs over a model that does not include them, for each realization shown. These results pertain to maps where 100% of the GCE is accounted for by DM and there are no PSs present in the simulated data. Each row in the figure represents a different Monte Carlo iteration of the map. In cases where the flux distributions are recovered exactly, the posteriors are centered at zero. Importantly, we see that when there are no unresolved PSs in the map, the NPTF analysis does not erroneously attribute the DM to PSs. For this example, the simulated map is made using the p6v11 model, and the same template is used in the analysis—this represents the case where there is no diffuse mismodeling. As expected, the ln (BF) is negative for the majority of realizations, pointing to a preference for a model without PSs. These are a random subset of 50 out of the 100 iterations shown in the left half of Fig. 4.7.

4.5 Dark Matter and the GCE

Having introduced the NPTF procedure and demonstrated how well it works in recovering soft and hard source-counts functions in simulated data, we now begin to test the method on more complex simulated maps. Specifically, this section will explore what happens as the relative amount of DM and PS flux contributing to the GCE...
varies, and the ability of the method to accurately recover these flux combinations. All examples in this section include modeling of the diffuse background emission and PSF effects, both in the construction of the simulated maps and in the analysis. For now, we assume that there is no diffuse mismodeling.

### 4.5.1 Dark Matter-Only GCE

We begin by considering the case where the entirety of the GCE is comprised of DM; the simulated maps consist of a DM signal accounting for 100% of the GCE flux, as well as \texttt{p6v11} diffuse background emission. We generate 100 different Monte Carlo realizations of this map and run the NPTF on each realization using the following three templates: NFW DM, NFW PS, and Galactic diffuse emission. Figure 4.3 shows (for a random subset of 50 out of the 100 scans) the flux posteriors for the three separate components, centered around the true injected value. The last column shows the Bayes factors (BFs) for each realization, which quantifies the statistical preference of a model of the simulated data that includes NFW PSs over a model that does not include them. Each row in the figure corresponds to a different Monte Carlo iteration. The posterior distribution for the NFW PS template is sharply peaked at zero in every case, meaning that the analysis recovers no PSs. The recovered DM flux is always close to the injected amount, albeit with some spread due to degeneracy with the diffuse emission. Importantly, however, a non-zero DM flux is recovered in all cases (the vertical line at “−True NFW DM” in the second panel denotes where the flux posteriors would lie if zero DM flux were recovered), and there is no statistical preference for an NFW PS population based on the Bayes factor, as should be the case.
Figure 4.4: Comparison of the flux posterior (relative to the true injected flux) for the Galactic diffuse, NFW DM, and NFW PS components, as well as the total flux, varying the relative fraction of the GCE accounted for by DM and PS. These results pertain specifically to the soft source-count function. We consider the four cases where the GCE is 100% PS (first column), 25% DM and 75% PS (second column), 50% DM and 50% PS (third column), and 75% DM and 25% PS (fourth column). In each panel, the solid line represents the median and the shaded region spans the minimum and maximum value in a given flux bin, across 100 Monte Carlo iterations. In each case where there are contributions from both DM and PSs, there is a probability that the DM signal is absorbed by the PS template, as evidenced by the peaks at “−True NFW DM” in the second row of each of the right three panels; the probability of the PS signal being absorbed by the NFW DM template increases with increased DM fraction, as evidenced by the increasingly large peak at “−True NFW PS” in the third row of each of the right three panels. The corresponding plot for the hard source-count distribution is shown in Fig. 4.18.
4.5.2 Fractional Dark Matter Recovery

We now consider the case where the simulated maps include Galactic diffuse emission, NFW DM, and NFW PSs—but vary the relative fraction of the total GCE flux that comes from DM and PSs. We specifically consider 25/75%, 50/50%, and 75/25% splits as examples. In each case, we run the NPTF analysis to test the recovered flux of each component and the details of the best-fit source-count function. We use the standard three templates: (i) NFW PSs, (ii) NFW DM, and (iii) Galactic diffuse emission.

Figure 4.4 summarizes the range of possibilities that occur when varying the DM/PS contributions. As a point of comparison, the left-most column shows the results for the case where the GCE is comprised of 100% soft PSs (and no DM). For each of the 100 Monte Carlo realizations of the map, we obtain the posterior distributions for the Galactic diffuse, NFW DM, and NFW PS components, similar to what is shown in Fig. 4.3. The median posterior distribution for each component is indicated by the solid line in each panel of Fig. 4.4; the shaded bands denote the maximum and minimum value obtained in any given flux bin over the separate map realizations. In this case, the median PS flux posterior is peaked at its true injected value, but there is a tail towards lower fluxes where some of the emission is instead absorbed by the DM template. We see this explicitly as the tail in the DM posterior extending towards the ‘+True NFW PS’ line. This speaks to the inherent degeneracy between the ultra-faint PSs and truly Poissonian emission.

It is notable that the effects of this degeneracy are evident in this case, but not when the GCE consists of 100% DM (Fig. 4.3). When there is only DM present, there are not enough bright pixels to give the PS template any statistical advantage in the fit. However, in the opposite scenario where the GCE is 100% PS, the PS template can still pull the statistical weight of fitting the comparatively brighter
unresolved sources, while the fainter sources are either attributed to the PS or DM template. Fig. 4.18 shows the corresponding figure for the hard source-count function. As expected, when there are fewer ultra-faint sources present (in the 100% PS GCE case), the probability that the DM template picks up the PS flux is reduced.

The second column of Fig. 4.4 illustrates the case where 25% of the GCE flux originates from DM and the remaining 75% comes from PSs. In this case, the median DM posterior is no longer peaked at the true injected value. The peak at ‘−True NFW DM’ indicates that the entirety of the DM flux is not recovered most of the time—the flux is instead absorbed by the PS template, whose posterior distribution now has a tail extending to larger fluxes and encompasses the ‘+True NFW DM’ line. However, a wide range of possibilities remains viable, depending on the Monte Carlo realization of the map; this includes the possibility that the PS flux is underestimated and is instead absorbed by the DM template.

As the relative amount of DM to PS flux in the GCE increases, this trend reverses. In particular, it becomes increasingly more likely that the PS flux is underestimated and incorrectly absorbed by the DM template. This can be seen in Fig. 4.4: going from left to right, the median PS flux posterior becomes increasingly peaked towards ‘−True NFW PS,’ while the median DM flux posterior becomes increasingly peaked towards ‘+True NFW PS.’ Although the average 75/25% DM/PS map clearly follows this behavior, there are still cases where the entirety of the DM flux is absorbed by the PS template. These cases are more unlikely, but still viable.

When the unresolved PSs have a hard flux distribution, as illustrated in Fig. 4.18, these trends continue to hold, but are less pronounced. In particular, the PS flux is never fully absorbed by the DM template, even in scenarios where the DM constitutes the majority of the GCE. Such behavior makes sense as it is more difficult for a collection of bright sources to fake a Poissonian DM signal.
To summarize, we find that, in the absence of diffuse mismodeling:

- The NPTF analysis never misattributes DM as PSs in the case where the GCE is 100% DM.

- When the GCE is 100% PSs, some fraction of the PS flux can be misattributed to DM. This depends sensitively on the flux distribution of the PSs, and is exacerbated for cases where there are more ultra-faint sources.

- When the GCE flux is split between DM and PSs, it is possible that the emission is misattributed between the two templates. In particular, when the DM accounts for a minority of the emission, it can be entirely absorbed by the PS template. As the DM fraction increases, this behavior reverses, with the DM template preferentially absorbing the PS contribution. This effect is again exacerbated for the soft source-count function.

- In all instances where the GCE flux is split between DM and PSs, we find cases—regardless of whether the relative DM/PS flux contribution is 25/75, 50/50, or 75/25%—where the entirety of the DM flux is misattributed to PSs. This becomes increasingly rare as the relative DM contribution to the GCE increases, but can still occur. We emphasize, however, that this never occurs when the GCE consists entirely of DM.

These results pertain specifically to the case where there is no mismodeling of the Galactic diffuse emission. We will consider the implications of diffuse mismodeling in the following section.

### 4.5.3 Signal Injection Tests on Monte Carlo

The study performed by Ref. [218] showed that injecting an artificial DM signal on data results in the signal being misattributed to PSs in the NPTF analysis. When in-
interpreting results from such signal injection tests, it is important to consider potential subtleties that may arise from injecting an artificial DM signal into data that itself contains a signal—the GCE. We explore on simulated datasets the NPTF recovery of a GCE-strength DM signal injected on top of an existing GCE signal, in the cases where the GCE is entirely accounted for by DM or by NFW PSs.

In the case where the GCE is 100% DM, we choose 10 of the Monte Carlo realizations from Fig. 4.3 as the base “data” maps onto which we inject an additional GCE-strength DM signal. The base maps are chosen to bracket a range of the possibilities depicted in Fig. 4.3. We inject 10 different Monte Carlo realizations of the additional DM signal onto each of the base maps, resulting in a total of 100 composite maps. We show the results across the 100 maps in the first column of Fig. 4.5. In this case, the DM signal is never absorbed by the PS template. We note that analyzing 10 Monte Carlo iterations for each base map is sufficient, because the variations in the recovered fluxes are dominated by differences in the base maps themselves rather than differences in Poisson realizations of the injected signal.

The results are quite different in the case where the GCE is 100% soft PSs. As we have already seen, a population of soft PSs can be more challenging to distinguish from DM or Galactic diffuse emission in Inner Galaxy. This confusion can be exacerbated as the total Poissonian flux increases and becomes spatially correlated with the PS distribution. Indeed, this is precisely what we see on simulation, as shown in the third column of Fig. 4.5. Similarly to the previous case, we choose 10 base maps spanning a range of possibilities, and inject 10 separate Monte Carlo realizations of an additional DM signal onto each of the base maps. There is considerable spread in the posterior distributions—in some cases, no DM is recovered and all the flux is entirely attributed to PSs; in other cases, a large fraction of the PS flux is attributed to DM. On average, there is a higher probability of the latter occurring. The bimodality
of the NFW DM and NFW PS flux posteriors is exacerbated compared to the base case with no additional injected signal, shown in the first column of Fig. 4.4. In that case, the recovered DM and PS fluxes are correct the majority of the time. However, for the signal injection test, the recovered DM and PS fluxes are almost always incorrect. Additionally, the spread in signal-on-signal case is larger than that of Fig. 4.4 (left panel)—note that the $x$-axes are different between the two figure panels. The corresponding results for the hard PSs can be found in the left column of Fig. 4.19. In this scenario, the spread in the DM and PS posteriors is smaller—in particular, the PS flux is never absorbed by the DM template, while the DM flux still may be absorbed by the PS template.

The results presented in this section demonstrate that signal injection tests on the GCE can be biased, even when the NPTF works robustly on the actual underlying data with no artificial signals present. This bias can arise from the fact that the injected DM is degenerate with ultra-faint PSs. If a soft PS population is already present in the data, the fit may not be penalized by absorbing injected DM flux into the PS template. Signal injection tests therefore yield less information than they would in the absence of this degeneracy. Indeed, we see that when a DM signal is injected on a map that already contains a population of PSs at the Galactic Center, the NPTF may naturally absorb this injected signal into the PS template. When the GCE consists entirely of DM, on the other hand, the signal injection test is well-behaved.

### 4.6 Diffuse Mismodeling

Thus far, the Poissonian templates included in the NPTF analyses have perfectly described the astrophysical backgrounds in the data (up to Poisson noise). In a
Figure 4.5: Same as Fig. 4.4, but with mock data constructed by injecting an additional GCE-strength DM signal on top of an existing GCE signal. All of the simulated data maps are made using the p6v11 diffuse model; to explore the effects of diffuse mismodeling, the second and fourth columns are analyzed using a Galactic diffuse template that is based on Model F from Ref. [133]. Importantly, when injecting the additional DM signal onto an existing GCE-strength DM signal, the analysis is well-behaved and there is no significant misattribution of flux between the DM and PS components, even in the presence of diffuse mismodeling. When injecting the additional signal onto an existing PS signal, there is confusion between the two components and the flux posteriors for the DM and PS templates become bimodal. Notably, in the absence of diffuse mismodeling, the amount by which the median PS flux posterior is shifted negative matches the true total flux contributed by ultrafaint sources below the $\sim 1\sigma$ significance threshold. In the presence of diffuse mismodeling, the DM signal is typically absorbed by the PS template, as evidenced by the median DM posterior peaking at “$-\text{True NFW DM}$” in the fourth column. The corresponding plot for the hard source-count distribution is provided in Fig. 4.19.
realistic setting however, the spatial morphology of the Galactic diffuse emission is rather poorly constrained. As a result, the templates used to model the diffuse emission describe the actual underlying background with uncertainty far exceeding the level of Poisson noise. This raises the possibility of, e.g., spurious residuals in the data that could mimic a PS signal even when actual astrophysical sources, such as MSPs, are not present. Conversely, it could be possible for a PS signal to be absorbed into the mismodeled diffuse background, resulting in the true flux and source-count distribution not being properly reconstructed.

In this section, we explore both these effects and comment on how recovery of DM and PS signals in the Inner Galaxy could be affected by mismodeling of the Galactic diffuse emission. We mock up the effect of diffuse mismodeling by analyzing the same simulated data that we have used so far (created using the *Fermi* p6v11 diffuse model) with an alternate diffuse model. In particular, we use Model F, which was found to be the best-fit to Inner Galaxy data out of the models considered in Ref. [133]. We explore the impact of diffuse mismodeling on PS and DM recovery in turn.

### 4.6.1 Point Source Signal Recovery

We redo the analysis presented in Sec. 4.4, but now use diffuse Model F to analyze the simulated maps. To build the Model F template, we obtain the best-fit normalizations of the gas (Bremsstrahlung and π⁰ decay) and Inverse Compton components on data, and then sum them together. Therefore, the Model F template is associated with a single fit parameter (i.e., its overall normalization), just like the p6v11 template that we used previously. The primary differences between the two lie in the specific assumptions made regarding the gas and IC models, as described in Ref. [133] (see also Ref. [198]).
Figure 4.6: Same as the right-most panel of Fig. 4.2, but analyzed using Model F rather than the p6v11 model, which was used to generate the simulated data maps. Results are shown for the soft PSs (left panels) as well as the hard PSs (right panels). Even at the ∼95% level over 100 Monte Carlo iterations, the recovered source-count function fails to capture the low-flux sources for the soft PSs. The cumulative flux distribution for the soft PSs shows that in the presence of diffuse mismodeling, the recovered flux is consistent with the true injected flux down to ∼$2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$, but is in excess for fluxes between ∼$1–2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$. In the case of the hard source-count distribution, the recovered function reliably captures the input at the ∼95% level. The individual flux posteriors for a random subset of 50 out of the 100 runs shown in the left panels are provided in Fig. 4.12. For comparison, we also provide (in Fig. 4.13) the flux posteriors for the case where the GCE consists of 100% DM and there is diffuse mismodeling.

The results are presented in Fig. 4.6, in analogy to the right-most panel of Fig. 4.2. For the soft PSs (left panels), the lower-flux biasing effects seen in Sec. 4.4 are exacerbated in the presence of diffuse mismodeling, leading in general to a steeper downturn in the source-count function towards lower fluxes. Additionally, an excess in the recovered PS flux is observed in the flux regime of ∼$1–2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$ as a
consequence of the clumpy residuals present due to mismodeling. In the case of the hard PSs (right panels), the PS recovery is unaffected by the presence of the diffuse mismodeling and is accurate, to within 95% confidence, within the entire flux range. We note that in the absence of diffuse mismodeling, the median recovered source-count distribution is accurate over the full flux range. Interestingly, the recovered source-count functions in the left and right panels are remarkably similar. These results demonstrate that diffuse mismodeling can make a genuinely soft population of sources “fake” a harder population as the diffuse residuals can mimic bright PSs. This could provide one explanation for why the best-fit source-count function recovered by the NPTF in the Inner Galaxy [118]—which is similar to the hard source-count function modeled here—differs from e.g. the MSP expectation.

Despite mischaracterizing the soft source-count distribution, the preference for a PS population remains robust in the face of mismodeling. This is quantified in the right half of Fig. 4.7, which shows the distribution of Bayes factors (BFs) in preference of a model including NFW PSs over a model without them, for 100 Monte Carlo realizations. This is illustrated as a heatmap, with the BFs along the horizontal axis corresponding to those obtained when the simulations are analyzed with the “correct” diffuse model (p6v11), and those along the vertical axis corresponding to analyses with the alternative diffuse model (Model F). Projected distributions are shown along both axes. Note that the overall scale of these BFs should not be compared directly to any results on actual Fermi data, as the setup here is very simple and is not intended to accurately represent the real data. The main take-away from these figures is the relative differences in BFs for the tests presented in this section.

Preference for a PS population remains robust in either case, with $\ln(BF) \gtrsim 20$. A stronger preference for PSs is seen in the case of diffuse mismodeling, as additional residuals are also picked up by the NFW PS template. There is a tight correlation
Figure 4.7: Bayes factors (BFs) characterizing the statistical preference for a model of the data that includes NFW PSs over a model that does not include them. The mock data consists of the GCE, accounted for by either 100% NFW DM (left) or by 100% NFW PS following the soft source-count distribution (right), and diffuse emission modeled by p6v11. Results are shown for 100 Monte Carlo iterations. We run the NPTF on these maps using three templates: (i) NFW PS, (ii) NFW DM, and (iii) Galactic diffuse emission. The BFs recovered using the p6v11 model for the diffuse template are shown along the horizontal axes, while the BFs recovered using the Model F template are shown along the vertical axes. (Note the difference in scale for the axes between the left and right halves of the figure.) Even in the presence of diffuse mismodeling, the NPTF robustly picks up evidence for PSs when they constitute 100% of the GCE. When the GCE is 100% DM, the evidence for PSs is always negligible in the p6v11 case; in the Model F case, while there is a strong peak around ln(BF) \sim 0, there are 35 iterations in which the diffuse mismodeling leads to residuals that are picked up as PSs with 5 < ln(BF) \lesssim 13. The significance of these detections is still smaller than the BF range when true PSs are actually present. We emphasize that while the relative values of the BFs are useful in comparing the different tests studied here, their overall scale should not be compared to any results on Fermi data as these maps are not intended to closely model an actual data realization.
between the BFs obtained from the p6v11 and Model F analyses, bolstering the fact
that, in the 100% PS scenario, the preference for PSs comes from the true underlying
PS population rather than as a consequence of the diffuse mismodeling on its own.

The analogous results with the addition of other astrophysical background com-
ponents (resolved 3FGL points sources, isotropic emission and emission from the
Fermi bubbles) and corresponding Poissonian templates are shown in the right half
of Fig. 4.8. The same overall conclusion is seen to hold, with the typical BFs in prefer-
ence for a model with PSs now being somewhat tempered, as expected in the presence
of additional smooth background emission. We emphasize once more that the overall
scale of these BFs should not be compared directly to any results on actual Fermi
data, as we are not accounting for additional PS populations that may be present
in the real data (such as disk-correlated sources), and the degree of mismodeling we
study here could be different than that in an analysis on the real data.

We conclude from these tests that, with the degree of diffuse mismodeling that
we have considered (p6v11 vs Model F), it is unlikely that a true underlying PS
population would be mischaracterized as DM.

To gain a sense of how comparable the degree of mismodeling studied here is to
that from a typical analysis on real data, we compare the residuals from our Monte
Carlo analyses (created with diffuse model p6v11 and analyzed with diffuse Model
F) to the residuals from fitting the p6v11 or Model F diffuse templates to data. In
all cases, we include Poissonian templates to account for emission from the Fermi
bubbles [115], isotropic emission, as well as emission from the resolved 3FGL point
sources [113]. We show residual sky maps and a histogram of residual counts in
Sec. 4.9. In our region of interest, the median and 16th/84th percentile magnitudes of
residuals are (in photon counts per pixel): $2.40^{+3.07}_{-1.70}$ for p6v11 fit to data, $2.38^{+3.00}_{-1.68}$ for
Model F fit to data, and $2.27^{+2.80}_{-1.59}$ (median values over 100 realizations) for Model F
Figure 4.8: Same as Fig. 4.7, but where the mock data consists of, in addition to the DM(PS) signal: p6v11 diffuse emission, emission from the Fermi bubbles [115], isotropic emission, and emission from 3FGL sources [113]. The left(right) panel shows results where the GCE is 100% DM(soft PSs). The NPTF analysis now also includes templates to model the isotropic, 3FGL, and bubbles emission. Compared to what we see in Fig. 4.7, the overall scale of the BFs is lower. Additionally, in the 100% DM case with diffuse mismodeling, the distribution is much more strongly peaked near a low value of ln(BF) \(\sim 1\), and the number of iterations for which residuals are picked up as PSs with \(5 < \ln(BF) \lesssim 13\) is reduced to 7 (compared to 35 cases in Fig. 4.7). We emphasize again that while the relative values of the BFs are useful in comparing the different tests studied here, their overall scale should not be compared to any results on Fermi data as the simple setup is not intended to accurately model the analysis on real data.

fit to p6v11 simulated data, respectively. The degree of mismodeling we simulate is slightly less than but roughly commensurate with that on real data, which suggests that our simulated mismodeling provides a reasonable proxy. However, we emphasize that the magnitude of residuals gives only a rough comparison, and that there will be additional differences in mismodeling on the real data versus on our simulated data.
For example, the spatial distribution of the residuals could be very different, which could have implications on the results of the NPTF analysis.

### 4.6.2 Dark Matter Signal Recovery

The effect of mismodeling on the recovery of a DM signal can be somewhat more subtle. In particular, whether DM recovery is successful or not can be strongly affected by the specifics of a given Poisson realization. This is expected, since the diffuse emission accounts for a large fraction of the flux in the Inner Galaxy, and large Poisson fluctuations combined with the effects of mismodeling can “fake” PS-like statistics.

Similarly to Sec. 4.5.1, we consider the case where the GCE consists of 100% DM. The corresponding BFs in preference of a model with PSs over a model without them are shown in the left halves of Fig. 4.7 (with only diffuse background emission) and Fig. 4.8 (with additional Poissonian background components). When the simulations are analyzed with the “correct” diffuse model, the BFs are always small, peaking around $\ln(BF) \sim 0$ and never showing significant preference for a PS population. When the diffuse emission is mismodeled, the $\ln(BF)$ still peaks near 0, but there is a tail in the distribution extending up to $\ln(BF) \sim 10$. These cases correspond to realizations where mismodeled residuals conspire to mimic a PS-like population in the data. The number of instances where the 100% DM case yields a $\ln(BF) \gtrsim 5$ in the presence of diffuse mismodeling is reduced when additional backgrounds are included in the map. This reduction may be due to the presence of additional Poissonian components in the model that may absorb the residuals.

Figure 4.9 shows the differential source-count distributions for the NFW PS template in simulations where the DM accounts for 100% of the GCE flux (no PS contribution), and there is diffuse mismodeling. Note that, for this example, we use the
Figure 4.9: Differential source-count distributions in the presence of diffuse mismodeling. The simulated data consists of the GCE, which is entirely accounted for by DM, and Galactic diffuse emission (corresponding to the left panel of Fig. 4.7). The simulated maps are generated using the p6v11 model and analyzed using Model F. We separately show the results for iterations with ln(BF) < 5 (left panel) and iterations with ln(BF) > 5 (right panel). In each case, the solid green line is the median best-fit distribution, and the bands denote the median 68/95% confidence intervals, recovered over 100 Monte Carlo realizations. In the cases with ln(BF) < 5, the typical best-fit source-count function is suppressed relative to the examples we have considered thus far, and does not look like the source-count function recovered in the NPTF analysis on Fermi data [118]; in the outlying cases with ln(BF) > 5, which correspond to the tail of the distribution shown in the leftmost panel of Fig. 4.7, the median recovered source-count function resembles the hard PS population in the left panel of Fig. 4.1. Note that when adding in additional Poissonian contributions to the mock data, as in Fig. 4.8, the number of instances where ln(BF) > 5 are significantly reduced.

simulated data corresponding to Fig. 4.7 as opposed to Fig. 4.8 simply because there are more runs with ln (BF) > 5; as we have seen, the inclusion of the additional Poissonian backgrounds decreases the frequency of such anomalously large BFs. In each panel of Fig. 4.9, the solid green line is the median best-fit distribution recovered over 100 Monte Carlo realizations, and the bands denote the median 68/95% confidence intervals. As there are no PSs in the simulated map, the NFW PS template should not pick up significant flux, which is confirmed by the peak near ln (BF) ∼ 1 in Fig. 4.7.
have considered thus far, and certainly does not look like the source-count function recovered in the NPTF analysis on *Fermi* data [118]. The 95% containment band, however, does encompass distributions that resemble the hard PS population shown in the left panel of Fig. 4.1. In the right panel, we show the anomalous cases where the BF in preference for PSs falls in the range $\ln(\text{BF}) \sim 5–13$. In this case, the typical recovered source-count distribution does resemble that of the hard PS population.

As we have shown, there are some instances where a DM signal can be mischaracterized as a PS population when the diffuse emission is mismodeled. While this only happens for a subset of realizations when diffuse Model F is used to analyze a map created using the *p6v11* diffuse model, it is plausible that a differences in mismodeling from what we have considered could lead to more consistent mischaracterization of the DM signal. It is therefore important when performing the NPTF analysis on the *Fermi* data to vary over the template(s) for the Galactic diffuse emission. In addition, it is crucial to compare the obtained BFs with those expected from corresponding simulations. This is because for different diffuse models (with potentially different degrees of freedom), the interplay between the diffuse model with other components could lead to different expected BFs for the same underlying PS population. If the set of diffuse models span the range of viable possibilities, then it is plausible that the recovered BFs would more consistently favor the PS interpretation across different models if PSs are in fact present in the data, compared to the scenario in which the GCE is truly DM—in that case, there would likely be more variation in the recovered BFs because the results would be more sensitive to the specifics of the residuals from mismodeling. A version of this test was performed in the original NPTF study of the GCE [118], and a preference for PSs was consistently recovered over the different diffuse models studied. An updated analysis focusing on mitigating the effects of
diffuse mismodeling in the NPTF procedure is presented in a companion paper [220], and is summarized more fully in the Conclusions.

### 4.6.3 Signal Injection Tests

Lastly, we revisit the signal injection tests discussed in Sec. 4.5.3, in the presence of diffuse mismodeling. Like before, we consider the cases where we have injected an additional DM signal (with flux equivalent to the GCE) onto simulated data maps in which the GCE is comprised entirely of either DM or soft PSs. The $p6v11$ model was used to generate the Galactic emission in the simulated maps, but we now repeat the NPTF analysis using Model F for the diffuse template instead. The flux posteriors for 100 Monte Carlo iterations are shown in the second and fourth columns of Fig. 4.5.

When the GCE consists of 100% DM, the NPTF almost always recovers the correct DM flux (up to a small offset between the diffuse and DM components), including the injected contribution (second column, Fig. 4.5). Additionally, the analysis finds on average that there are no PSs. This result clearly demonstrates that the additional injected DM signal is recovered when there are no PSs present in the data, even if the diffuse emission is mismodeled to the extent that we consider.

In contrast, when the artificial DM signal is injected onto a map where the GCE consists of soft PSs, there is more variation in the results. In particular, when the diffuse emission is accurately modeled, the NPTF on average recovers all of the injected DM flux and additionally absorbs the total flux contributed by ultrafaint PSs (below the $\sim 1\sigma$ significance threshold) into the DM template (third column, Fig. 4.5). On the other hand, when the diffuse emission is mismodeled, the NPTF consistently absorbs the injected DM flux into the PS template (fourth column, Fig. 4.5). When the PS population is hard, as shown in Fig. 4.19, the injected DM is almost always reconstructed as PSs, especially when analyzed using the Model F template. These
Figure 4.10: Differential source-count distributions recovered when the simulated data consists of the GCE, comprised entirely of soft PSs, with an additional injected GCE-strength DM signal. The p6v11 model is used to generate the Galactic diffuse emission in the simulated data, and the analysis is performed using the “correct” diffuse model (p6v11, left panel) or the “incorrect” diffuse model (Model F, right panel). The left(right) panel corresponds to the third(fourth) column of Fig. 4.5. These results can be compared with the cases without the additional DM injection, shown in the top right panel of Fig. 4.2 for p6v11 and the top left panel of Fig. 4.6 for Model F. In the analysis with diffuse model p6v11, a significant portion of the PS flux is typically absorbed by the DM template (correspondingly, the source-count distribution is suppressed in the $\sim 5 \times 10^{-11} - 2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$ flux range, relative to Fig. 4.2, top right panel). On the other hand, when the data is analyzed with the diffuse Model F, the injected DM is consistently absorbed into the PS model (correspondingly, the source-count distribution captures excess flux in the $\sim 10^{-11} - 10^{-10}$ counts cm$^{-2}$ s$^{-1}$ range, relative to Fig. 4.6, top left panel).

tests demonstrate that the additional DM photons that are injected into the map can conspire with the residuals from diffuse mismodeling to look like a population of PSs. As a result, the signal injection test fails and the injected DM flux is not correctly reconstructed.

This point is further emphasized in Fig. 4.10, which shows the differential source-count distributions corresponding to the third and fourth columns of Fig. 4.5. The left panel shows the result using the “correct” p6v11 diffuse model, and the right panel shows the result using the “incorrect” diffuse Model F. These can be directly compared to the recovered source-count distributions in the absence of the additional
DM injection, shown in Fig. 4.2 (top right panel) for p6v11 and Fig. 4.6 (top left panel) for Model F. It can be seen that, when the diffuse emission is mismodeled using Model F, the injected DM is consistently absorbed by the PS model—evident from the fact that the analysis consistently recovers more PS flux in the $\sim 10^{-11} - 10^{-10}$ counts cm$^{-2}$ s$^{-1}$ range, compared to the case without the injected DM signal. This is similar to the observed misattribution of injected DM flux to the PS model in Ref. [218], and is in contrast to the analysis with the “correct” p6v11 diffuse model, where on average, a significant portion of the PS flux gets absorbed by the DM template—correspondingly, the recovered source-count distribution is suppressed in the $\sim 5 \times 10^{-11} - 2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$ flux range, compared to the case without the injected DM signal.

Similarly to [218], we have also explored the effect of injecting even brighter DM signals (200% and 300% of the GCE flux) onto soft PSs constituting the GCE. These results are presented in Fig. 4.14. In the absence of diffuse mismodeling (first and third columns), the injection of brighter DM signals mitigates the bimodality of the posterior flux distributions, but the typical results are largely unchanged from the case where the injected DM is 100% of the GCE flux (third column, Fig. 4.5). When diffuse mismodeling is present (second and fourth columns), the injection of increasingly bright DM signals reduces the probability that the injected DM signal is absorbed by the PS template in the NPTF analysis. The effects of diffuse mismodeling on signal injection tests are likely to be further exacerbated when analyzing the real data. Understanding how to mitigate these effects on the data warrants a dedicated study, which we present separately in a companion paper [220].

We emphasize that whether or not the signal injection test fails has no bearing on the validity of the NPTF analysis on the original dataset. Indeed, we find that the signal injection test fails most spectacularly when PSs are already present in the
simulated data and—as we see in Figs. 4.7 and 4.8—the NPTF analysis finds strong preference for PSs in these cases (with no injected DM), as it should.

4.7 Conclusions

In this chapter, we performed a systematic study of the Non-Poissonian Template Fitting (NPTF) method on Monte Carlo data, focusing on its ability to distinguish between the DM and PS origins of the Fermi Galactic Center Excess (GCE). Our primary conclusions are as follows:

- When the Galactic diffuse backgrounds are perfectly modeled, the NPTF accurately identifies a GCE that is comprised entirely of DM. If the GCE is 100% PSs, then some of the PS flux may be misattributed to DM. When the GCE is split between DM and PSs, then the NPTF can struggle to identify the correct fluxes of each, especially when the PSs are relatively soft. These challenges arise from the fact that PSs are exactly degenerate with smooth Poissonian emission in the ultra-faint limit.

- Assuming no diffuse mismodeling, we find that, when the GCE is 100% PSs, the source-count distribution recovered by the NPTF in the Inner Galaxy accurately characterizes the underlying flux distribution of PSs down to a per-source significance of $\sim 1\sigma$ while being potentially biased at lower fluxes. This bias is especially true when the PS population is characterized by a soft source-count distribution with a large number of ultra-faint PSs.

- Evidence for a PS population can still be robustly recovered when the Galactic diffuse emission is mismodeled, at least for the one particular (albeit representative) case we considered. However, the residuals from the mismodeling can make the recovered source-count function appear to be brighter than it truly is.
This may suggest that the best-fit source-count function in the Inner Galaxy NPTF analysis of Ref. [118] is not necessarily indicative of the true distribution for the underlying PS population. This may potentially explain why the empirical distribution, which is peaked close to the 3FGL threshold, does not resemble models of the MSP luminosity function.

- In the presence of the diffuse mismodeling we consider, the NPTF almost always correctly identifies a GCE consisting entirely of DM. Correspondingly, the Bayes factors in preference for PSs are typically not significant and the recovered source-count functions for the NFW PS template are suppressed. In a small fraction of cases, however, we do find that the NPTF can erroneously show evidence for PSs. This preference is never as strong as what we find when PSs are truly present, and is particularly sensitive to the choice of diffuse template (and corresponding residuals) used in the study. One way to test that such effects are not driving the preference for PSs on data is to simply rerun the NPTF analysis using a variety of different diffuse templates. In doing so, it is also important to compare the inferred Bayes factors to their expected values from simulation.

Our work also allows us to comment on the results presented in Ref. [218], which found that the NPTF can misattribute an artificial DM signal injected onto the Fermi data to PSs. We have mocked up such signal injection tests on simulated data maps, injecting an additional DM signal (with the same flux as the GCE) onto maps where the GCE is comprised entirely of DM or of hard/soft PSs. We conclude that, at least to the extent we have tested:

- When an additional DM signal is injected onto a map where the GCE is comprised entirely of DM, the NPTF correctly reconstructs the total (original +
injected) DM flux. This remains true in the majority of iterations even when the Galactic diffuse emission is mismodeled.

- When an additional DM signal is injected onto a map where the GCE is comprised entirely of PSs, there is often confusion between the DM and PS components. This can arise from the fact that the DM signal is degenerate with the ultra-faint PSs. In particular, in the presence of diffuse mismodeling, the injected DM signal is consistently reconstructed as PS flux.

- When an additional DM signal, 2–3 times brighter than the GCE, is injected onto a map where the GCE is comprised entirely of PSs, the confusion between the DM and PS components can be somewhat mitigated. In particular, in the presence of diffuse mismodeling, the probability that the injected DM signal gets reconstructed as PS flux is reduced as the brightness of the injected DM signal is increased.

Our results demonstrate that the failure of the NPTF to extract an injected DM signal can be natural in the presence of PSs in the data, particularly in the presence of diffuse mismodeling. Additionally, whether the signal injection tests succeed or fail is not an accurate diagnostic of the NPTF analysis on the original dataset (without the injected signal). Indeed, we find that in cases where the signal injection tests fail, the NPTF accurately recovers the PSs on the original dataset. Our findings demonstrate that great care must be taken when interpreting the results of DM signal injection tests on data.

Throughout this work, we have only considered the case where the DM and PSs both trace the NFW profile, as opposed to different spatial distributions. Ref. [218] considered a scenario where there is a population of unresolved PSs that trace the Fermi bubbles—a proof-of-principle example as there is no evidence for such sources
in data. However, in such cases where the DM and PSs follow different spatial morphologies, there are additional handles that may be used to discriminate the DM and PS hypotheses, such as simply changing the ROI. Such possibilities are addressed in more detail in a companion paper [220].

In conclusion, this chapter provides a systematic assessment of the NPTF method on simulated data. Taking a pedagogical approach, we highlight the cases where the NPTF method works robustly, and cases where the output may be biased. These results provide important context for interpreting the results of the NPTF studies on actual data. In a companion paper [220], we revisit the NPTF analysis of the Inner Galaxy in the Fermi data, exploring how the results vary with the region of study as well as with the choice of diffuse emission model. We will also present a novel method that uses a spherical harmonic decomposition of the diffuse model to help lessen the effects of large-scale mismodeling. Taken together with the discussion presented in this chapter, these tests are designed to reduce the systematic uncertainties and biases associated with the NPTF analysis and strengthen the conclusions drawn from such studies on data.

4.8 Supplementary Figures

In this section, we provide a set of supplementary figures that are referenced and described in the main text. In Figs. 4.11–4.17, PSs (when relevant) correspond to the soft source-count distribution. Figs. 4.18–4.19 show results for the hard source-count distribution.
Figure 4.11: Comparison of the flux posterior (relative to the true injected value) for the Galactic diffuse, NFW DM, and NFW PS components. The last column shows the Bayes factors (BFs), characterizing the statistical preference of a model of the data that includes NFW PSs over a model that does not include them. These results pertain to maps where 100% of the GCE is accounted for by soft PSs (corresponding to the right-most panel of Fig. 4.2). We run the NPTF on these maps using three templates: (i) NFW PS, (ii) NFW DM, and (iii) Galactic diffuse emission (p6v11 model). Each row in the figure represents a different Monte Carlo iteration of the map. These are a random subset of 50 out of the 100 iterations shown in the left half of Fig. 4.7. In cases where the PS flux is underestimated, it is typically picked up by the NFW DM template. Even when the PS flux is underestimated, decisive evidence for a PS population is still seen based on the Bayes factors. We emphasize that the overall scale of the BFs should not be compared to any results on Fermi data as these maps are not intended to closely model an actual data realization.
Figure 4.12: Same as Fig. 4.11, but this time analyzed using Model F as the Galactic diffuse emission model. This corresponds to the case where the diffuse emission is mismodeled. Decisive evidence for a PS population is seen for each realization, even with the diffuse mismodeling. We emphasize that the overall scale of the BFs should not be compared to any results on Fermi data as these maps are not intended to closely model an actual data realization.
Figure 4.13: Same as Fig. 4.3, but this time analyzed using Model F as the Galactic diffuse emission model (note the different scale in the rightmost panel from Fig. 4.3). Misattribution of the DM flux to PSs is significantly exacerbated in the presence of diffuse mismodeling, and evidence for a PS population can be erroneously inferred for a subset of the realizations due to residuals from the mismodeling. However, the misattribution is ameliorated when more DM signal is injected into the mock data, as demonstrated in the second column of Fig. 4.5. We emphasize that while the relative values of the BF s are useful in comparing the different scenarios and realizations studied here, their overall scale should not be compared to any results on Fermi data as these maps are not intended to closely model an actual data realization.
Figure 4.14: Same as the right two columns of Fig. 4.5, but with even brighter DM signals injected on top of an existing GCE-strength soft PS signal. The simulated data maps consist of 10 distinct “base” maps in which the GCE is entirely accounted for by soft PSs, onto which an additional 200% GCE flux (left two columns) or 300% GCE flux (right two columns) DM signal is injected. The injected DM signal is Poisson fluctuated to generate 10 realizations for each base case. In the absence of diffuse mismodeling (first and third columns), the injection of increasingly bright DM signals reduces the bimodality of the DM and PS posterior flux distributions (see third column of Fig. 4.5 for comparison). On average, the injected DM signal is fully recovered, and the DM template additionally picks up the flux contribution from ultrafaint PSs below the 1σ significance threshold. In the presence of diffuse mismodeling, the injection of brighter and brighter DM signals mitigates the absorption of the DM signal by the PS template. Note the different scale of the horizontal axes in this case compared to Fig. 4.5.
Figure 4.15: An example triangle plot for an NPTF scan that passes the convergence criteria on $S_{b,1}$ described in Section 4.3.1: the posterior distribution for $S_{b,1}$ is nicely converged within the prior range. The simulated data in this particular instance consists of 100% soft PSs and diffuse emission. The templates used in this scan are: (i) NFW PS, (ii) NFW DM, and (iii) Galactic diffuse emission (p6v11). Where applicable, the true simulated values are indicated on the 1d posterior distributions by thick solid red lines.
Figure 4.16: Same as Fig. 4.15, but where the simulated data consists of 100% DM and diffuse emission. In this case, the posterior distribution for $S_{b,1}$ is unconstrained over the prior range. Where applicable, the true simulated values are indicated on the 1d posterior distributions by thick solid red lines.
Figure 4.17: An example triangle plot for an NPTF scan that fails the convergence criteria on $S_{b,1}$ described in Section 4.3.1: the posterior distribution for $S_{b,1}$ is pushed against the lower prior edge. The simulated data in this particular instance consists of 100% DM and diffuse emission, and there is diffuse mismodeling present in the scan. The templates used in the scan are: (i) NFW PS, (ii) NFW DM, and (iii) Galactic diffuse emission (Model F). Where applicable, the true simulated values are indicated on the 1d posterior distributions by thick solid red lines—the “true” diffuse model normalization here corresponds to the norm that yields equivalent flux to the true simulated (p6v11) flux. Scans like this are discarded from all results presented in this chapter. Additionally, we note that implementing a lower flux cutoff in the source-count function (detailed in Section 4.10) drastically reduces the number of such scans.
Figure 4.18: Same as Fig. 4.4, except for the case of the hard PS source-count distribution. In this case, the PSs never get fully absorbed by the DM template, as is evidenced by the absence of a strong peak at “−True NFW PS” in the NFW PS posterior flux distributions. However, there is still a probability that the DM signal gets absorbed by the PS template.
Figure 4.19: Same as Fig. 4.5, but for the case of the hard PS source-count distribution. Compared to case of the soft source-count distribution, there is less spread in the NFW DM and NFW PS flux posteriors. In particular, the PS flux is never absorbed by the DM template in these cases, while the DM template still may be absorbed by the PS template. The injected DM flux is consistently misattributed to PSs in the presence of diffuse mismodeling (right column).
Figure 4.20: Residual photon counts in our ROI from fitting the data with a model that includes (i) the Galactic diffuse emission, (ii) the Fermi bubbles [115], (iii) isotropic emission, and (iv) resolved Fermi 3FGL PSs [113]. We show the residual sky maps for the cases where the Galactic diffuse emission is described by the p6v11 model (left panel) or Model F (middle panel). We also show the difference between the best-fit sky maps obtained using each of the two models. For presentation purposes, we have smoothed each map by a Gaussian with $\sigma = 1^\circ$.

### 4.9 Residuals from mismodeling

This section provides some relevant figures to give a sense of the degree of mismodeling we have considered in Sec. 4.6. In each case described within this section, the model is purely Poissonian, and consists of (i) the Galactic diffuse emission, (ii) the Fermi bubbles, (iii) isotropic emission, and (iv) resolved Fermi 3FGL PSs. Figure 4.20 shows the photon count residuals from fitting the Fermi data to such a model, where the diffuse emission template is either p6v11 (left panel) or Model F (middle panel). To illustrate how the p6v11 and Model F fits to data differ, we also show difference of the two best-fits (right panel).

For a more quantitative comparison, Fig. 4.21 shows the histogram of the residual photon counts for the p6v11 case (green dashed) and the Model F case (gray dotted). We have also generated 100 simulated data maps using the best-fit p6v11 diffuse emission (along with the other Poissonian components), and analyzed the simulated maps using Model F to describe the diffuse emission. The median residual histogram
Figure 4.21: Histogram of residual photon counts in our ROI. The green dashed line indicates the case where the Fermi data is analyzed using the p6v11 diffuse model (this corresponds to the left panel of Fig. 4.20). The gray dotted line indicates the case where the Fermi data is analyzed using Model F (this corresponds to the middle panel of Fig. 4.20). For comparison, we generate 100 Monte Carlo (MC) data maps consisting of the following components (best-fit from the Fermi data): (i) p6v11 Galactic diffuse emission, (ii) the Fermi bubbles [115], (iii) isotropic emission, and (iv) resolved Fermi 3FGL PSs [113], which we analyze using Model F and Poissonian components (ii)–(iv). The median result is shown by the solid blue line, while the shaded blue bands span the minimum and maximum value in each bin over the 100 simulated maps. The three cases are roughly comparable, although on average, there tend to be fewer residuals with large magnitudes when fitting the p6v11 simulated data with Model F than when fitting the data with either model.

from the 100 Monte Carlo realizations is shown in solid blue, and the shaded blue band spans the minimum/maximum across realizations. On average, the residuals from the latter scenario are somewhat smaller in magnitude than the residuals obtained on the real data. However, taking into account the variation across Monte Carlo realizations, the three are comparable. We therefore conclude that our method for simulating diffuse mismodeling in Sec. 4.6 is a reasonable proxy for the typical degree of mismodeling on the real data.
4.10 Source-count Function with Flux Cutoff

The fact that emission from unresolved PSs is degenerate with Poissonian DM emission in the ultra-faint limit is a critical point for understanding the output of the NPTF analysis. Here, we consider the scenario where the NPTF only identifies PSs that are bright enough to be distinguished from Poissonian emission, and does not attempt to distinguish fainter sources from DM below some flux cutoff. In practice, this means that we use the source-count function from Eq. 4.4 above some flux cutoff, but set it to zero below. For illustration, we will consider a cutoff value that corresponds to a 1σ point-source significance within our setup.

Figure 4.22 shows the effect of the cutoff on the recovered source-count and cumulative flux distributions for the case where the simulated data consists of soft PSs and Galactic diffuse emission, modeled assuming p6v11. The templates used in the model include: (i) NFW PS, (ii) NFW DM, and (iii) p6v11 Galactic diffuse emission (i.e., no mismodeling). Note that, in this implementation, the NFW DM template should be picking up both the true DM emission, as well as the flux from unresolved sources that fall below the flux cutoff. The left column is a copy of Fig. 4.2 (right panel), and the right column shows the corresponding result when the cutoff is implemented. We see that the presence of the cutoff does not greatly affect the best-fit source-count function, though it does lead to a slight overestimate of the flux above the cutoff. We do however find a significant reduction in the number of iterations where the upper break ($S_{b,1}$) of the NFW PS template is peaked at the lower edge of the prior range—cases that we discarded in the main analyses (see Sec. 4.3 for a discussion). This suggests that the cutoff may help to regulate the anomalous cases where the source-count distribution is peaked in the regime where the NPTF loses sensitivity.
Figure 4.22: Differential source-count distributions (top panels) and cumulative flux distributions, integrated above a given threshold (bottom panels), shown for simulations with PSs with a soft source-count distribution, accounting for 100% of the GCE (no DM contribution). The diffuse emission is not mismodeled in this case. The solid lines indicate the median best-fit distributions recovered over 100 Monte Carlo realizations of the simulated data maps. The bands show the median 68 and 95% confidence bands over these 100 iterations. We show a copy of the right column of Fig. 4.2 (left) and the corresponding result when a flux cutoff is added to the parameterization of the source-count function in the NPTF analysis (right). The addition of the cutoff does not strongly affect the recovered source-count distribution. However, a slight discrepancy (at the 68% level) is introduced between the recovered and true cumulative flux distributions near the cutoff.

Figure 4.23 shows the corresponding result in the presence of diffuse mismodeling. In this case, the addition of the cutoff in the source-count function regulates the excess above $\sim 10^{-10}$ counts cm$^{-2}$ s$^{-1}$.

We have also considered the effect of the flux cutoff when the NPTF is run on simulated data maps where the GCE consists entirely of DM. Figures 4.24 and 4.25 show
Figure 4.23: Same as Fig. 4.22, except for the case where the diffuse emission is mismodeled. We show a copy of the left column of Fig. 4.6 (left) and the corresponding result when a flux cutoff is added to the parameterization of the source-count function in the NPTF analysis (right). Here, we find that the cutoff helps to reduce the peak in the recovered source-count distribution near $\sim 2 \times 10^{-10}$ counts cm$^{-2}$ s$^{-1}$.

the resulting source-count distributions when, respectively, the p6v11 and Model F templates are used in the analysis. Again, the addition of the flux cutoff does not seem to have a strong effect on the source-count distribution. However, for the case of diffuse mismodeling, the addition of the cutoff does reduce the number of instances when the DM is misidentified as PSs with ln (BF) $> 5$, as shown in Fig. 4.26 (right panel). For comparison, we also show in the left panel of Fig. 4.26 that the distribution of BFs in the absence of diffuse mismodeling is shifted to lower values and unchanged in shape with the addition of the cutoff.

Lastly, we have studied the effect of the flux cutoff in the cases where the simulated data consists of the GCE, comprised entirely of soft PSs, with an additional GCE-strength DM signal injected on top. Figure 4.27 shows the resulting posterior flux
distributions. We find that in the absence of diffuse mismodeling (left two columns), the cutoff slightly mitigates the bimodality of the flux posteriors, whereas in the presence of diffuse mismodeling (right two columns), the cutoff slightly reduces the probability for the injected DM signal to be absorbed by the PS template. We note that we have additionally tested the effect of the flux cutoff in the cases where the GCE consists of 25% DM and 75% PS, 50% DM and 50% PS, and 75% DM and 25% PS. In those cases, we have found that the cutoff has a negligible effect on the results, and we therefore omit the corresponding figures.
Figure 4.26: The Bayes factors in preference for PSs when the GCE consists entirely of DM in the absence (left panel) or presence (right panel) of diffuse mismodeling. In each case, solid(hatch)-shaded histogram corresponds to recovered source-count distributions without(with) a flux cutoff. In the p6v11 analyses (left panel), the addition of the flux cutoff shifts the distribution of BFs to lower values while leaving the shape of the distribution unchanged. On the other hand, in the Model F analyses (right panel), the cutoff reduces the number of iterations with $\ln(BF) \gtrsim 5$ while increasing the number of iterations with $1 \lesssim \ln(BF) \lesssim 5$. 
Figure 4.27: Posterior flux distributions for the case where the simulated data is made up of the GCE, comprised by soft PSs, with an additional GCE-strength DM signal injected on top. The first and third columns are duplicates of the third and fourth columns of Fig. 4.5, respectively, and are provided here for comparison. The second and fourth columns show the corresponding results when a flux cutoff is applied to the source-count distribution in the NPTF analysis. In the absence of diffuse mismodeling (first and second column), the cutoff mitigates the bimodality of the posterior flux distributions. In the presence of diffuse mismodeling (third and fourth columns), the cutoff reduces—on average—the probability that the injected DM signal is absorbed by the PS template. This is demonstrated by reduced peak in the median DM posterior at “−True NFW DM” in going from the third column to the fourth column.
Chapter 5

Distressed Jeans: A Systematic Assessment of Limitations in Dynamical Mass Modeling

This chapter is based on work done in collaboration with Lina Necib, posted as arXiv:2009.00613. We systematically examine the spherical Jeans modeling method applied to simulated dwarf galaxy kinematic data. We focus on identifying limitations on the ability of the method to accurately reconstruct the density distribution of dark matter in dwarf galaxies as well as determining which observational advancements would be most impactful towards addressing these limitations. Section 5.2 presents details on the Jeans modeling methods used in this study. Section 5.3 presents the results of the study. In particular, we explore the effects of the number of observed stars (Sec. 5.3.1), the measurement errors of the line-of-sight velocities (Sec. 5.3.2), and the locations of the observed stars (Sec. 5.3.3) on the DM inference. Sec. 5.3.4 explores the impact of degeneracies between model parameters. Sec. 5.4 recasts our results into
the context of $J$-factors for indirect detection, where we emphasize the dependence of the results for ultrafaint dwarf analogues on the priors chosen in the analysis. In Sec. 5.4.4, we discuss the implications of our results on future observations. Our main conclusions are summarized in Sec. 5.5. This work has been presented at the Small Galaxies, Cosmic Questions (CosmicDwarfs) 2019 conference in Durham, England (July 2019).

5.1 Introduction

The standard ΛCDM model, consisting of the cosmological constant $\Lambda$ and cold dark matter (CDM), has had remarkable success at predicting physics on large scales (e.g., the cosmic microwave background [45] and the large-scale distribution of matter in the Universe [235, 236, 237]), but faces several small-scale challenges [13]. Among these challenges is the “core-cusp problem” [238, 239]—$\Lambda$CDM predicts that, in the absence of baryonic physics, dark matter (DM) halos universally follow a Navarro-Frenk-White (NFW) density profile [111], which steeply rises as $\rho \propto r^{-1}$ towards central regions. However, a number of measurements of rotation curves and stellar dynamics have suggested that the DM distribution in the centers of dwarf galaxies may be more consistent with having a constant density core [238, 239, 240, 241, 242, 243, 244, 245].

If the DM halos of dwarf galaxies truly are cored, one potential way to explain this apparent discrepancy is through baryonic physics. In baryonic contraction, the central density of a galaxy increases due to the infall of dissipative baryons, deepening the potential well and dragging DM into the central region, leading to the formation of a DM core—this happens primarily in Milky Way-sized galaxies [246]. On smaller scales, stellar feedback can lead to core formation due to the ejection of baryons [247, 248, 249, 250]. While there is qualitative agreement in the simulation
literature surrounding the formation of cores in dwarf galaxy-sized DM halos, there is considerable scatter in the quantitative results from various works.

However, while there is qualitative agreement in the simulation literature surrounding the formation of cores in dwarf galaxy-sized DM halos, there is considerable scatter in the quantitative results from various works. Recent studies of hydrodynamic simulations have shown that lower mass dwarfs ($M_* \lesssim 10^6 M_\odot$) have cuspy DM halos, while efficient core formation from stellar feedback turns on around $M_* \sim 10^9 M_\odot$; for galaxies slightly more massive than the Milky Way, the DM halo reverts back to a cuspy distribution [204, 251, 252]. Ref. [253] correlated the presence of cores to an active stellar formation history in isolated simulated dwarf galaxies. Similarly, simulations with a lower density threshold for star formation, for example Auriga [254] and APOSTLE [255], find that cores do not form at dwarf galaxy sizes [256]. Core formation thus depends on the baryonic feedback model, and while present observations are inconsistent with low star formation thresholds [257, 258], reliable observational evidence for cusps or cusps in dwarf galaxies has important implications for understanding stellar feedback and galaxy formation.

A different approach to resolving the core-cusp problem is to modify the particle model of DM itself—for example, models of self-interacting dark matter (SIDM) notably predict the formation of central cores in the DM density profiles of low-mass galaxies [259]. There has been extensive work in the literature studying halo formation in SIDM [260, 261, 262, 263]. Ref. [264] additionally found that the shape of the stellar distribution is correlated with the SIDM core at sub-kpc scales for high enough self-scattering cross sections. SIDM has also been invoked to address the so-called “diversity problem” of dwarf galaxy rotation curves [265], namely the observation that there is a spread in the inner DM density profiles that are preferred by the rotation curves of different galaxies [266, 267, 268], which is inconsistent with
the expectation from standard ΛCDM. In addition to SIDM, other theories of DM can also predict different halo properties from the ΛCDM prediction—for example, theories of dissipative DM have been shown to lead to the formation of halos with inner density profiles that are more steeply cusped than NFW halos [269]. The inner profiles of dwarf galaxy DM halos can therefore encode information about the underlying particle physics that governs the DM.

Whether the Milky Way dwarf galaxies truly all reside in cored or cuspy halos, or there is a large scatter in the inner density profile shapes, there would be important consequences for our understanding of the underlying baryonic and DM physics. By nature of dwarf galaxies being dispersion-supported and low in gas content, any dynamical inference of the DM distribution in dwarfs relies on utilizing stellar data. At present, there is a lack of consensus in the dwarf galaxy literature on whether the stellar data favors cuspy or cored DM distributions. One specific example is the case of Sculptor, one of the more extensively analyzed dwarf galaxies in the mass modeling literature. Sculptor has been observed to have two chemo-dynamically distinct subpopulations of stars with different half-light radii, which can be leveraged to constrain the DM density at two different radii. Ref. [270] applied separate Jeans analyses to the two stellar components and found that either a cored halo or an NFW halo were statistically consistent with their data. Ref. [271] applied a mass estimator to the data for the two components and concluded that their analysis ruled out an NFW profile at \( \gtrsim 99\% \) significance. Ref. [272] used a separable distribution function method and found strong statistical preference for a cored DM profile, while Ref. [273] found that with a more flexible distribution function model, the statistical preference went away and the data was consistent with an NFW halo.

Additionally, there are important caveats to consider when interpreting the results from the various mass modeling methods. For example, while many rotation curve
analyses have shown preference for cored DM distributions, studies have shown that systematic effects in rotation curve analyses can erroneously bias the inferred DM distribution towards a centrally cored profile (see, e.g., [274, 275, 276] and references within). Meanwhile, mass estimator methods such as the one introduced in Ref. [271] rely on assuming that the velocity dispersion profile of the stars is isotropic and radially flat, which may not be a reliable assumption in the data; moreover, it has been shown that the results of these methods can depend sensitively on the specific line of sight that is chosen, and can result in predicting a cored profile when the true halo is cuspy [277, 278].

Aside from addressing the cusp-core problem, robustly inferring the DM density distribution in dwarf galaxies is also important in the context of DM indirect detection. Indirect detection is the process in which DM annihilates or decays into Standard Model (SM) particles, and the resulting SM particles are subsequently detected. The probability of detecting such a signal is maximized in regions of the sky with high DM density, such as the centers of dwarf galaxies or the Milky Way Galactic Center (GC). Indeed, an excess of \( \sim \) GeV photons was detected near the GC by the \textit{Fermi} Large Area Telescope [103], which could be interpreted as a signal of DM annihilation [120, 130, 133, 135]. However, DM analyses near the GC are complicated by bright and complex astrophysical backgrounds, and it is important to have complementary search targets, some of which have excluded or placed the DM interpretation of the excess under tension [160, 148, 162, 143, 279, 149, 280].

Some of the complementary targets studied in the indirect detection literature have been the Milky Way halo at high latitudes [143, 177, 176], galaxy groups [162, 161], Andromeda [280], and stacked dwarf galaxies [281, 282, 160, 148, 279, 149]. In particular, dwarf galaxies are generally considered to be the most robust search
targets within the indirect detection literature, because they are expected to have little astrophysical background emission \[144, 145\].

In general, the expected signal flux from DM annihilation is proportional to the so-called astrophysical $J$-factor, which is defined as the integrals over the solid angle $\Omega$ and along the line of sight $s$ of the DM density squared:

$$J = \int ds \int d\Omega \, \rho^2(s, \Omega),$$

(5.1)

where $\rho$ is the DM density. The robustness of any dwarf galaxy-based indirect detection constraint on DM annihilation is dependent on accurately estimating the $J$-factors of the analyzed dwarf galaxies, and therefore dependent on accurately inferring their DM density distributions.

Finally, reliably reconstructing the total DM mass in dwarf galaxies also has important scientific implications. This is related to—but can be separate from—accurately inferring the full DM density distribution, because while the density and enclosed mass distributions are directly related, there can be cases where the total mass is accurately estimated even if the shape of the density distribution is not fully reconstructed, as we demonstrate in this chapter. Obtaining accurate estimates of the total DM mass in dwarf galaxies plays a key role in determining the low-mass end of the stellar-to-halo mass relation (SHMR) (see Ref. [283] for a review of the galaxy-halo connection). Various works have delved into the determination of the expected DM halo mass as a function of the stellar mass using abundance matching, a method that works reliably for massive galaxies \[284, 285, 286\]. On smaller scales, studies on simulations have found that galaxy formation is significantly suppressed in DM halos with virial mass below $\sim 10^8 M_\odot$ \[154, 287\], leading Refs. \[288, 289\] to propose scatter at the lower mass end of the SHMR. A more accurate determination of the DM halo mass in the smallest dwarf galaxies would help empirically anchor the
SHMR for the smallest systems, for which the uncertainty on the relation between galaxies and their DM halos is the largest.

In this chapter, we apply spherical Jeans modeling [290, 291, 292, 293] to simulated dwarf galaxy kinematic datasets, varying over properties of the mock observations such as the total number of observed stars, the measurement error on line-of-sight velocities, as well as the locations of the observed stars (e.g., whether they are primarily in the central region of the dwarf or farther out). We choose to focus on spherical isotropic dwarf galaxies in equilibrium. By studying the limitations of the Jeans analysis method even in this simplified scenario, we are able to identify which observational advancements are more likely to make an impact on our ability to accurately reconstruct the properties of dwarf galaxy DM halos in the near future.

5.2 Methods

5.2.1 Jeans Modeling

We summarize the standard procedure for inferring the velocity dispersion profile of the stars in a dwarf galaxy from measurements of their line-of-sight velocities, following the derivations of Refs. [294, 295]. We start with the collisionless Boltzmann equation,

$$\frac{\partial f}{\partial t} + \vec{v} \frac{\partial f}{\partial \vec{x}} - \frac{\partial \Phi}{\partial \vec{x}} \frac{\partial f}{\partial \vec{v}} = 0,$$

(5.2)

where $f$ is the phase-space density of a stellar tracer population, a function of the position $\vec{x}$ and velocity $\vec{v}$ of each star, and $\Phi$ is the gravitational potential of the dwarf galaxy. Multiplying Eq. 5.2 by velocity component $v_j$ and integrating over all velocities, we have

$$\frac{\partial}{\partial t}(\nu v_j) + \frac{\partial}{\partial x_i}(\nu \vec{v} \dot{v}_j) + \nu \frac{\partial \Phi}{\partial x_j} = 0,$$

(5.3)
where we have defined $\nu = \int d\vec{v}^3 f(\vec{x}, \vec{v})$, the spatial density of the tracer stars. Assuming the system is spherically symmetric, and is in steady state (and therefore the $\partial/\partial t$ term is negligible), we have

$$\frac{\partial}{\partial r} (\nu \sigma_r^2) + \nu \left( \frac{\partial \Phi}{\partial r} + \frac{2\sigma_r^2 - \sigma_\theta^2 - \sigma_\phi^2}{r} \right) = 0, \quad (5.4)$$

where $\sigma_i^2$ is the square of the $i^{th}$ component of the velocity dispersion for $i \in \{r, \theta, \phi\}$, i.e., $\sigma_i^2 = \langle v_i^2 \rangle - \langle v_i \rangle^2$.

We can then define the velocity anisotropy,

$$\beta(r) = 1 - \frac{\sigma_\theta^2 + \sigma_\phi^2}{2\sigma_r^2}, \quad (5.5)$$

and explicitly write the potential as

$$\Phi = -\frac{GM(< r)}{r}, \quad (5.6)$$

where $G$ is the gravitational constant and $M(< r)$ is the enclosed mass within radius $r$. Plugging these quantities back into Eq. 5.4, we end up with the following first-order differential equation for $\nu \sigma_r^2$:

$$\frac{1}{\nu} \left[ \frac{\partial}{\partial r} (\nu \sigma_r^2) + \frac{2\beta(r)}{r} (\nu \sigma_r^2) \right] = -\frac{GM(< r)}{r^2}. \quad (5.7)$$

The generic solution to Eq. 5.7 takes the form

$$\nu(r) \sigma_r^2(r) = \frac{1}{g(r)} \int_r^\infty \frac{GM(< \tilde{r}) \nu(\tilde{r})}{\tilde{r}^2} g(\tilde{r}) d\tilde{r}, \quad (5.8)$$
where the new function $g(r)$ is defined as

$$g(r) = \exp \left( 2 \int \frac{\beta(r)}{r} dr \right). \quad (5.9)$$

The enclosed mass $M(< r)$ in Eq. 5.8 can be related to the overall density distribution by

$$M(< r) = 4\pi \int_0^r \rho(s)s^2 ds, \quad (5.10)$$

where, again, we have assumed spherical symmetry of the system. While both the stars and DM contribute to the mass density distribution, i.e., $\rho = \rho_{\text{DM}} + \rho_{\text{stars}}$, we expect the density of DM to dominate, and therefore make the approximation $\rho \approx \rho_{\text{DM}}$. This is a valid approximation due to the large mass-to-light ratios of dwarf galaxies, $M_{\text{halo}}/M_* \approx 10^2-10^5$ [289].

In practice, typically only projected radii and line-of-sight velocities are measured, and therefore Eq. 5.8 needs to be projected along the line of sight. To do so, we use the Abel transform, defined for a spherically-symmetric function as

$$S(R) = 2 \int_R^{+\infty} \frac{s(r)rdr}{\sqrt{r^2 - R^2}}, \quad (5.11)$$

where $s(r)$ is the function in three-dimensional spherical coordinates, $R$ is the projected radius, and $S(R)$ is the resulting projected function. Projecting Eq. 5.8 along the line of sight leads to the equation [294, 296]

$$\sigma_p^2(R)I(R) = 2 \int_R^{\infty} \left( 1 - \beta(r) \frac{R^2}{r^2} \right) \frac{\nu(r)\sigma_r^2(r)r}{\sqrt{r^2 - R^2}} dr, \quad (5.12)$$
where \( \sigma_p \) is the projected velocity dispersion profile and \( I(R) \) is the projected density distribution of the tracer stars, given by

\[
I(R) = 2 \int_R^\infty \frac{\nu(r)rdr}{\sqrt{r^2 - R^2}}. \tag{5.13}
\]

Throughout the remainder of this chapter, \( I(R) \) is referred to as the surface brightness profile or light profile.

Using Eq. 5.12, we build a likelihood function to fit the observed data and extract information on the dark matter distribution. In the literature, the analysis has been performed in either a binned \([297, 298]\) or unbinned \([299]\) fashion. In this work, we will focus on the unbinned analysis as it can take into account the errors of each star separately. The unbinned Gaussian likelihood function is given by \([299]\)

\[
\mathcal{L} = \prod_{i=1}^{N_{\text{stars}}} \frac{(2\pi)^{-1/2}}{\sqrt{\sigma_p^2(R_i) + \Delta v_i^2}} \exp \left[ -\frac{1}{2} \left( \frac{(v_i - \overline{v})^2}{\sigma_p^2(R_i) + \Delta v_i^2} \right) \right], \tag{5.14}
\]

where \( \overline{v} \) is the mean velocity for the population of tracer stars, and for star \( i \), \( v_i \) is the measured line-of-sight velocity, \( \sigma_p(R_i) \) is the intrinsic velocity dispersion at the projected radius \( R_i \), and \( \Delta v_i \) is the velocity measurement error. In our analysis, we choose closed-form parameterizations for the stellar and dark matter distributions, thereby reducing the number of integrals that need to be performed when calculating the likelihood.

It is important to emphasize the interplay between the intrinsic velocity dispersion and the measurement error in Eq. 5.14—if the measurement errors are subdominant to the intrinsic velocity dispersion of the system, it is not expected that improvements to the line-of-sight velocity measurements would drastically improve the quality of the fit. This will be further discussed in Sec. 5.3.2.
In this work, we focus on isotropic models. Namely, we take the velocity anisotropy, defined in Eq. 5.5, to be

\[ \beta(r) = 0. \] (5.15)

The velocity anisotropy \( \beta(r) \) is a known complication in Jeans modeling, because it is degenerate with the enclosed mass profile \( M(<r) \) [300, 301, 302, 303, 304, 305]. It can be seen from Eq. 5.12 that \( \beta(r) \) and \( \sigma_r^2(r) \) are degenerate with each other, which, combined with Eq. 5.8, implies that \( \beta(r) \) is degenerate with \( M(<r) \). Unfortunately, \( \beta(r) \) can only be measured with full 3D velocity information, which is not yet available for the majority of the stars in dwarf galaxies. It is therefore standard practice in Jeans analyses to assume a parametric model for \( \beta(r) \) and fit for it in conjunction with fitting for \( M(<r) \) [306, 307, 147, 308, 309]. The effect on dynamical mass modeling estimates when the assumed \( \beta(r) \) model does not match the true velocity anisotropy distribution has been studied in Ref. [310]. In this work, we choose to focus entirely on isotropic datasets and models in order to understand the limitations of the Jeans modeling procedure even in the absence of additional complications due to velocity anisotropy.

### 5.2.2 Dark Matter Profile

Using the public code STARSMPLER\(^1\), we generate the tracer stars in a DM potential which follows the Hernquist/Zhao profile [311, 312]

\[
\rho_{\text{DM, Zhao}}(r) = \rho_0 \left( \frac{r}{r_s} \right)^{-\gamma} \left[ 1 + \left( \frac{r}{r_s} \right)^{\alpha} \right]^{(\gamma-\beta)/\alpha},
\] (5.16)

\(^1\text{https://github.com/maoshen1/StarSampler}\)

159
where $\alpha, \beta, \gamma$ are the slopes of the distribution, $\rho_0$ is the overall normalization of the density profile, and $r_s$ is the scale radius—in particular, $\gamma$ sets the asymptotic inner slope of the distribution. This model has five free parameters, which introduces too many degenerate degrees of freedom into the model to effectively constrain the DM distribution (we discuss the role of degeneracies in Sec. 5.3.4). We therefore simplify the DM profile by setting $\alpha = 1$ and $\beta = 3$, which reduces Eq. 5.16 to a generalized Navarro-Frenk-White (gNFW) distribution with inner slope parameter $\gamma$, defined as

$$
\rho_{\text{DM}}^{\text{gNFW}}(r) = \rho_0 \left( \frac{r}{r_s} \right)^{-\gamma} \left[1 + \left( \frac{r}{r_s} \right) \right]^{-(3-\gamma)}.
$$

While we use the gNFW distribution to model the DM profile in our fiducial analysis setup, we additionally consider the special cases where the inner slope $\gamma = 0$ or 1. The case of $\gamma = 1$ corresponds to the standard, cuspy Navarro-Frenk-White (NFW) profile

$$
\rho_{\text{DM}}^{\text{NFW}}(r) = \rho_0 \left( \frac{r}{r_s} \right)^{-1} \left(1 + \frac{r}{r_s} \right)^{-2},
$$

whereas the case of $\gamma = 0$ leads to a constant density central core. We refer to this distribution as the cored NFW (NFWc) distribution, given by

$$
\rho_{\text{DM}}^{\text{NFWc}}(r) = \rho_0 \left(1 + \frac{r}{r_s} \right)^{-3}.
$$

The profiles defined by Eqs. (5.17)–(5.19) give rise to closed-form enclosed mass distributions, which we include below for reference.

$$
M_{\text{DM}}^{\text{gNFW}}(r) = \frac{4\pi}{3 - \gamma} \rho_0 r^3 \left( \frac{r}{r_s} \right)^{-\gamma} _2F_1 \left(3 - \gamma, 3 - \gamma; 4 - \gamma; -\frac{r}{r_s} \right) \quad (5.20)
$$

$$
M_{\text{DM}}^{\text{NFW}}(r) = 4\pi \rho_0 r^3_s \left( \frac{-r}{r + r_s} + \log \left(1 + \frac{r}{r_s} \right) \right) \quad (5.21)
$$

$$
M_{\text{DM}}^{\text{NFWc}}(r) = 4\pi \rho_0 r^3_s \left[ -r(3r + r_s) \frac{2}{2(r + r_s)^2} + \log \left(1 + \frac{r}{r_s} \right) \right] \quad (5.22)
$$

160
5.2.3 Light Profile

Using StarSampler, we model the stellar density distribution also as a Hernquist/Zhao profile

\[
\nu(r) = \rho_* \left( \frac{r}{r_*} \right)^{-\gamma_*} \left[ 1 + \left( \frac{r}{r_*} \right)^{\alpha_*} \right]^{(\gamma_* - \beta_*)/\alpha_*}.
\]

(5.23)

Eq. 5.23 reduces to a Plummer profile when the slope parameters are set to \( \alpha_* = 2 \), \( \beta_* = 5 \), and \( \gamma_* = 0 \). In this chapter, we focus on this specific case. The stellar samples in this analysis are all generated with the same level of “embeddedness” in their respective DM halos by setting the scale radius of the tracers, \( r_* \), to be equal to the scale radius of the DM distribution, \( r_s \).

Correspondingly, in our Jeans analysis, we model the stellar density \( \nu(r) \) as a 3d Plummer profile [313], defined as

\[
\nu(r) = \frac{3L}{4\pi a^3} \left( 1 + \frac{r^2}{a^2} \right)^{-5/2},
\]

(5.24)

where \( L \) is the total luminosity and \( a \) is scale length of the distribution. Eq. 5.24 has the same form as Eq. 5.23, with \( \alpha_* = 2 \), \( \beta_* = 5 \), \( \gamma_* = 0 \), \( r_* = a \), and \( \rho_* = 3M/(4\pi a^3) \).

The surface brightness profile (or light profile), which is the projection of \( \nu(r) \) along the line of sight, is then given by the closed form expression

\[
I(R) = \frac{L}{\pi a^2} \left( 1 + \frac{R^2}{a^2} \right)^{-2}.
\]

(5.25)

Because the contribution of the stellar tracers to the gravitational potential is negligible, changing the value of \( L \) in Eqs. (5.24) and (5.25) does not meaningfully affect the result of the Jeans modeling.

\[\text{In practice, rather than setting } \gamma_* = 0, \text{ we follow the example StarSampler and Ref. [271] and set } \gamma_* = 0.1 \text{ for ease of comparison. We do not expect it to affect the results.}\]
Table 5.1: Prior ranges for the stellar and DM parameters used in our analysis. We implement uniform priors within each of the listed prior ranges. The ranges listed here for $\log_{10}(a)$ and $\log_{10}(L)$ are used in the initial light profile fit; in the full Jeans scan, we set the prior ranges for $\log_{10}(a)$ and $\log_{10}(L)$ to be the middle 95% containment range of the posterior for each parameter from the initial fit (see Section 5.6 for more discussion on the light profile fit).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_{10}(a/kpc)$</td>
<td>$[-3,3]$</td>
</tr>
<tr>
<td>$\log_{10}(L/L_\odot)$</td>
<td>$[-2,5]$</td>
</tr>
<tr>
<td>$\ln(\rho_0/(M_\odot \text{kpc}^{-3}))$</td>
<td>$[5,30]$</td>
</tr>
<tr>
<td>$\ln(r_s/kpc)$</td>
<td>$[-10,10]$</td>
</tr>
<tr>
<td>$\pi/(\text{km s}^{-1})$</td>
<td>$[-100,100]$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$[-1,5]$</td>
</tr>
</tbody>
</table>

5.2.4 Parameters and Priors

We perform our Jeans modeling procedure in two stages. First, we perform a fit to only the positions of the stars. We describe this light profile fitting procedure in Section 5.6. We do so because the light profile is generally much better constrained than the stellar kinematics. We can then use the results from the initial fit to set the prior range on the light profile parameters in our full Jeans fit. We conservatively set the prior ranges on the light profile parameters in the full scan to be the middle 95% containment range of the posterior probability distributions output from the initial fit. In both stages, we use the PYMULTINEST module (introduced in Ref. [230]), which interfaces with the nested sampling Monte Carlo library MULTINEST [229], to sample the relevant likelihood.\(^3\) We summarize the priors for all of the parameters in our model in Table 5.1. We choose a wide prior of $[-1,5]$ for the parameter $\gamma$, which sets the inner slope of the gNFW distribution. We choose the lower edge to be at $-1$ such that there is sufficient range for convergence at $\gamma = 0$ while not allowing for

\(^3\)We use $n_{\text{live}} = 100$ live points in the nested sampling procedure throughout this chapter, but have verified that increasing to $n_{\text{live}} = 1000$ does not affect the results.
Table 5.2: DM halo parameters and properties of the datasets generated in this work. \( \rho_0 \), \( r_s \), and \( \gamma \) are the true values of the normalization, scale length, and inner slope input into Eq. 5.17. \( M_{200} \) is defined as the enclosed mass at \( r_{200} \), the radius within which the average density is equal to 200 times the critical density of the Universe at redshift \( z = 0 \), derived from the true density distribution. We adopt a generalized definition of the concentration \( c_{200} \equiv r_{200}/r_s \) for all of our parameter sets. \( \sigma_p \) is the median line-of-sight velocity dispersion across all the datasets generated for each set of parameters (10 realizations each for sample sizes of 20, 100, 1000, and 10,000 stars, resulting in a total of 40 datasets).

|      | \( \rho_0 \) [\( M_\odot/\text{kpc}^3 \)] | \( r_s \) [kpc] | \( \gamma \) | \( M_{200} \) [\( M_\odot \)] | \( c_{200} \) | \( \sigma_p \) [km/s] |
|------|------------------------------------------|----------------|-----------|-------------------------------|-------------|----------------|}
| Cusp I | \( 6.4 \times 10^7 \) | 1 | 1 | \( 1.9 \times 10^9 \) | 25.8 | 14.6 |
| Cusp II | \( 6.4 \times 10^7 \) | 0.2 | 1 | \( 1.5 \times 10^7 \) | 25.8 | 2.9 |
| Core III | \( 6.4 \times 10^7 \) | 1 | 0 | \( 1.4 \times 10^9 \) | 23.6 | 9.5 |
| Core IV | \( 6.4 \times 10^7 \) | 0.2 | 0 | \( 1.1 \times 10^7 \) | 23.6 | 1.9 |

In our fiducial model, there are a total of six free parameters: two parameters for the light profile, three parameters for the DM density distribution parameterized as a gNFW profile, and one parameter for the mean stellar velocity. In our discussion on characterizing the inner slope of the DM distribution, we additionally perform fits assuming either an NFW or cored NFW distribution, and compare the Bayesian evidence between the two models—in these fits, there are a total of five free parameters.

5.2.5 Mock Data

Using STAR_SAMPLER, we generate datasets with four different sets of DM halo parameters (summarized in Table 5.2). Our parameter choices span different halo masses
and either an inner cusp or inner core for the DM density profile while maintaining approximately the same halo concentration. Due to the large amount of scatter in the theoretical predictions for the subhalo mass-concentration relation, we choose not to focus on a specific mass-concentration model; however, the concentrations of our simulated halos are consistent with theoretical predictions in the literature for the relevant mass range [141, 314, 315].

Parameter sets I and III correspond to $M_{200} \sim 10^9 M_\odot$ halos, while sets II and IV correspond to smaller halos with mass $M_{200} \sim 10^7 M_\odot$. We emphasize that we have chosen to study $M_{200} \sim 10^7 M_\odot$ halos for demonstrative purposes, to study how the effect of the measurement error on the line-of-sight velocities impacts less massive halos differently from more massive ones. We have adopted a generalized definition for the halo concentration, $c_{200} \equiv r_{200}/r_s$, for all of the parameter sets that we generate, where $r_{200}$ is the radius within which the average density is 200 times the critical density of the Universe at redshift $z = 0$. The virial mass $M_{200}$ is subsequently defined as the enclosed mass at $r_{200}$.

For each set of DM parameters, we generate 10 realizations each of datasets with 20, 100, 1000, and 10,000 stars, respectively. The chosen sample sizes are meant to provide comparison with current measurements of ultrafaint dwarfs and classical dwarfs (see Table 5.5 for comparison), as well as projections for how future measurements might improve the quality of the DM inference. For our fiducial analyses, we assume a measurement error of $\Delta v = 2 \text{km/s}$ on the line-of-sight velocity. This is comparable to the typical uncertainty in current measurements (see, e.g., references within Table 5.5).
5.3 Results

We now apply the analysis pipeline described in Secs. 5.2.1–5.2.4 to the simulated stellar samples described in Sec. 5.2.5 and summarized in Table 5.2. Our main figures of merit for evaluating the success or limitations of our analyses are: 

(i) the overall recovered DM density profile, 
(ii) the recovered enclosed DM mass, which we quantify as the recovered virial mass $M_{200}$, and 
(iii) the recovered inner slope of the DM density profile, i.e., the parameter $\gamma$ in Eq. 5.17. Of the figures of merit, (i) has important implications on the inferred astrophysical $J$-factors (Eq. 5.1) which are used in indirect DM searches, (ii) is crucial for empirically probing the SHMR down to low halo masses, while (iii) can shed light on the particle physics properties of the DM as well as baryonic feedback and galaxy formation mechanisms.

We explore how several factors in the analysis influence the accuracy of the inferred DM profiles, focusing primarily on the effects of variations on the specifics of the analyzed datasets. In Sec. 5.3.1, we study how the total number of observed stars influences the inferred DM profile. In Sec. 5.3.2, we study the role of the line-of-sight velocity measurement errors; we explore how the magnitude of the error differently impacts the DM inference in dwarf galaxies with different halo masses. In Sec. 5.3.3, we study the effect of the locations of observed stars on the inferred DM profile. In Sec. 5.3.4, we explore how the presence of degeneracies between the DM profile parameters affects the inference of the inner slope $\gamma$.

5.3.1 Increase in Sample Size

Our first question of interest is how the number of observed stars in a dwarf galaxy affects the DM inference. In Figure 5.1, we show the inferred DM density profiles $\rho(r)$ and corresponding enclosed mass profiles $M(r)$ for parameter set I (which has
\( \gamma = 1, \; r_s = 1 \text{kpc}, \; M_{200} \approx 1.9 \times 10^9 \text{M}_\odot \)

Figure 5.1: Inferred DM density profiles \( \rho(r) \) (left panels) and corresponding enclosed mass profiles \( M(r) \) (right panels) for parameter set I—from lightest to darkest, samples with 20, 100, 1000, and 10,000 stars. We show the full reconstructed distributions in the top panels as well as the fractional (relative to truth) distributions in the bottom panels. For each sample size, the solid line denotes the median (across our 10 independent realizations) of the median recovered profiles, while the shaded band shows the median of the 68% containment regions, plotted from the innermost to outermost star across all 10 datasets for that sample size. For the samples with fewer than 10,000 stars, we additionally extrapolate the median 68% containment regions over the full radial range, shown bracketed by each pair of dashed lines in the color corresponding to the sample size. Across all sample sizes, the typical inferred density profile and enclosed mass profile are consistent within uncertainty with the true distributions. Increasing the observed sample size reduces the uncertainty on the recovered profiles, as expected.

\( \gamma = 1 \), for the four different sample sizes—from lightest to darkest, 20, 100, 1000, and 10,000 stars. Throughout this chapter, we will use \( r \) to denote the 3d galactocentric radius and \( R \) to denote the projected radius. For a given sample size, we run each of our 10 realizations through the analysis pipeline and obtain the resulting posterior density and enclosed mass profiles. Each solid line in Fig. 5.1 shows the median of the median recovered profiles across the 10 realizations, while the shaded band depicts
the median of the 68% containment regions across the realizations. The solid line and shaded band for each sample size are plotted from the innermost to outermost star across the 10 generated datasets for that sample size; outside of the data range for the smaller samples, we extrapolate the results and outline the 68% containment region with dashed lines in the color corresponding to each sample size. The extrapolation down to smaller radii is particularly important in understanding the implications for indirect detection, which we discuss in Section 5.4.

We find that, for all sample sizes in parameter set I ($\gamma = 1$), the typical inferred density profile and enclosed mass profile are consistent within uncertainty with the true distributions over the full range of measured radii. This can be seen from the fact that the dashed black lines in the top panels of Figure 5.1, indicating the true distributions, are contained within the bands for all of the sample sizes, as well as the fact that all the bands in the bottom panels overlap with the horizontal dashed black line. Additionally, we find that increasing the observed sample size reduces the uncertainty on the inferred density and enclosed mass profiles, as is to be expected. For all sample sizes and parameter sets, we list the median across our 10 realizations of the median and $\pm 1\sigma$ values of the inferred virial mass, $M_{200}$, in Table 5.2.

We show the analogous results for parameter set III ($\gamma = 0$) in Figure 5.3. In this case, for sample sizes of 20 stars and 100 stars, the inferred density distribution is typically biased towards a steeper inner profile than the true distribution, which has an inner slope of $\gamma = 0$, while for the datasets with 1000 and 10,000 stars, the typical inferred density profiles are consistent with the true distribution within uncertainty. Importantly, across all of the sample sizes, we obtain an accurate estimate for the total mass of the system, with the uncertainties on the estimate reduced as the sample size is increased (values listed in the fourth column of the corresponding panel in Tab. 5.2).
I. $\gamma = 1$, $r_s = 1 \, \text{kpc}$, $M_{200} \approx 1.9 \times 10^9 \, M_\odot$, $\log_{10} [J(0.5^c)/(\text{GeV}^2 \text{cm}^{-5})] \approx 19.3$

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>$M_{200} [10^9 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^9 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^9 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$3.0^{+22.0}_{-2.1}$</td>
<td>$19.6^{+1.8}_{-1.1}$</td>
<td>$2.1^{+6.7}_{-1.3}$</td>
<td>$19.9^{+1.9}_{-1.1}$</td>
<td>$1.6^{+4.9}_{-1.1}$</td>
<td>$19.9^{+2.1}_{-1.3}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.5^{+1.2}_{-0.5}$</td>
<td>$19.8^{+1.1}_{-0.5}$</td>
<td>$1.3^{+0.6}_{-0.3}$</td>
<td>$19.7^{+1.0}_{-0.5}$</td>
<td>$1.4^{+0.2}_{-0.3}$</td>
<td>$19.8^{+0.6}_{-0.6}$</td>
</tr>
<tr>
<td>1000</td>
<td>$1.9^{+0.5}_{-0.3}$</td>
<td>$19.4^{+0.4}_{-0.3}$</td>
<td>$1.9^{+0.6}_{-0.3}$</td>
<td>$19.4^{+0.4}_{-0.2}$</td>
<td>$1.9^{+0.6}_{-0.3}$</td>
<td>$19.4^{+0.4}_{-0.2}$</td>
</tr>
<tr>
<td>10,000</td>
<td>$1.8^{+0.1}_{-0.1}$</td>
<td>$19.3^{+0.2}_{-0.1}$</td>
<td>$1.9^{+0.1}_{-0.1}$</td>
<td>$19.3^{+0.2}_{-0.1}$</td>
<td>$1.9^{+0.2}_{-0.1}$</td>
<td>$19.3^{+0.2}_{-0.1}$</td>
</tr>
</tbody>
</table>

II. $\gamma = 1$, $r_s = 0.2 \, \text{kpc}$, $M_{200} \approx 1.5 \times 10^7 \, M_\odot$, $\log_{10} [J(0.5^c)/(\text{GeV}^2 \text{cm}^{-5})] \approx 17.3$

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>$M_{200} [10^7 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^7 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^7 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$0.7^{+1.0}_{-0.3}$</td>
<td>$18.6^{+1.3}_{-1.2}$</td>
<td>$0.8^{+0.5}_{-0.6}$</td>
<td>$18.4^{+1.6}_{-1.5}$</td>
<td>0.00001$^{+0.00006}_{-0.00001}$</td>
<td>$11.2^{+2.6}_{-7.4}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.2^{+0.9}_{-0.4}$</td>
<td>$17.6^{+0.7}_{-0.4}$</td>
<td>$1.0^{+1.4}_{-0.5}$</td>
<td>$17.8^{+1.2}_{-0.6}$</td>
<td>0.14$^{+0.3}_{-0.1}$</td>
<td>$15.3^{+3.5}_{-7.6}$</td>
</tr>
<tr>
<td>1000</td>
<td>$1.5^{+0.3}_{-0.2}$</td>
<td>$17.4^{+0.3}_{-0.2}$</td>
<td>$1.3^{+0.4}_{-0.3}$</td>
<td>$17.4^{+0.3}_{-0.2}$</td>
<td>0.9$^{+0.7}_{-0.4}$</td>
<td>$16.0^{+0.8}_{-0.5}$</td>
</tr>
<tr>
<td>10,000</td>
<td>$1.6^{+0.1}_{-0.1}$</td>
<td>$17.4^{+0.2}_{-0.1}$</td>
<td>$1.4^{+0.2}_{-0.1}$</td>
<td>$17.4^{+0.2}_{-0.1}$</td>
<td>1.4$^{+0.8}_{-0.4}$</td>
<td>$17.6^{+0.4}_{-0.2}$</td>
</tr>
</tbody>
</table>

III. $\gamma = 0$, $r_s = 1 \, \text{kpc}$, $M_{200} \approx 1.4 \times 10^9 \, M_\odot$, $\log_{10} [J(0.5^c)/(\text{GeV}^2 \text{cm}^{-5})] \approx 17.9$

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>$M_{200} [10^9 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^9 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$0.8^{+3.6}_{-0.5}$</td>
<td>$19.3^{+2.1}_{-1.1}$</td>
<td>$0.7^{+2.3}_{-0.3}$</td>
<td>$19.9^{+2.2}_{-1.3}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.5^{+1.6}_{-0.6}$</td>
<td>$18.5^{+0.6}_{-0.4}$</td>
<td>$1.6^{+1.8}_{-0.7}$</td>
<td>$18.5^{+0.6}_{-0.4}$</td>
</tr>
<tr>
<td>1000</td>
<td>$1.6^{+0.3}_{-0.2}$</td>
<td>$18.2^{+0.2}_{-0.2}$</td>
<td>$1.5^{+0.5}_{-0.2}$</td>
<td>$18.2^{+0.2}_{-0.2}$</td>
</tr>
<tr>
<td>10,000</td>
<td>$1.4^{+0.1}_{-0.1}$</td>
<td>$18.0^{+0.3}_{-0.1}$</td>
<td>$1.5^{+0.1}_{-0.1}$</td>
<td>$18.0^{+0.3}_{-0.1}$</td>
</tr>
</tbody>
</table>

IV. $\gamma = 0$, $r_s = 0.2 \, \text{kpc}$, $M_{200} \approx 1.1 \times 10^7 \, M_\odot$, $\log_{10} [J(0.5^c)/(\text{GeV}^2 \text{cm}^{-5})] \approx 16.3$

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>$M_{200} [10^7 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
<th>$M_{200} [10^7 , M_\odot]$</th>
<th>$\log_{10} [J(0.5^c)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$0.4^{+1.7}_{-0.3}$</td>
<td>$17.0^{+1.2}_{-0.7}$</td>
<td>0.00004$^{+0.00003}_{-0.00004}$</td>
<td>$11.6^{+1.8}_{-0.7}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.5^{+2.3}_{-0.8}$</td>
<td>$16.5^{+0.3}_{-0.2}$</td>
<td>0.4$^{+1.7}_{-0.3}$</td>
<td>$16.9^{+1.2}_{-0.7}$</td>
</tr>
<tr>
<td>1000</td>
<td>$1.4^{+0.3}_{-0.2}$</td>
<td>$16.4^{+0.2}_{-0.1}$</td>
<td>$1.5^{+0.8}_{-0.4}$</td>
<td>$16.5^{+0.3}_{-0.2}$</td>
</tr>
<tr>
<td>10,000</td>
<td>$1.1^{+0.8}_{-0.6}$</td>
<td>$16.5^{+0.2}_{-0.2}$</td>
<td>$1.2^{+0.2}_{-0.1}$</td>
<td>$16.5^{+0.2}_{-0.2}$</td>
</tr>
</tbody>
</table>

Figure 5.2: Inferred values of the virial mass $M_{200}$ and $J$-factor for the different parameter sets, sample sizes, and values of $\Delta v$. The $J$-factors are in units of $\log_{10} [J(0.5^c)/(\text{GeV}^2 \text{cm}^{-5})]$. Each entry in this table represents the median across 10 realizations of the median and $\pm 1\sigma$ values.
Figure 5.3: Same as Figure 5.1, except for parameter set III. For sample sizes of 20 and 100 stars, the inferred density distribution is typically biased towards a steeper inner profile than the true distribution; however, the inferred virial mass is still consistent with the true virial mass (values listed in Table 5.2). For sample sizes with 1000 and 10,000 stars, both the inferred density distribution and enclosed mass profile are consistent within uncertainty with the true distributions across the measured radial range.

This suggests that while the inferred density distribution may not always accurately represent the true underlying distribution, the virial mass estimate remains fairly robust. Namely, if the inferred density profile is biased high in the inner region of the dwarf, as in the example of parameter set III for 20- and 100-star samples, this is compensated for by the density profile being biased low in the outer region. We note that because the outer slope of the density profile is not a free parameter in the fit, the outer profile is uniquely determined by the scale radius and overall normalization. Additionally, our likelihood (Eq. 5.14) depends directly on the enclosed mass distribution of the system rather than the density distribution, and therefore it is not surprising that the fit is successful at recovering the total mass of the system even when it fails to accurately reproduce the inner density profile.
Figure 5.4: Posterior distributions for the inner slope $\gamma$. The top row corresponds to the scans shown in Fig. 5.1 (parameter set I), while the bottom row corresponds to the scans shown in Fig. 5.3 (parameter set III). The lines (bands) show the median (middle 68%) in each $\gamma$ bin across the 10 realizations. In both cases, the inner slope is generally poorly constrained for the smaller samples, with the median posterior distribution only peaking near the true value of $\gamma$ (vertical dashed line in each panel) for the largest sample size of 10,000 stars. All panels in this figure share the same vertical scale.

Figs. 5.1–5.3 have demonstrated that the inner regions of the inferred density profiles can be biased and/or poorly constrained, especially for the smaller datasets. We can further assess how well the inner density profile is recovered by directly examining the posterior probability distribution of the parameter in our model which sets the asymptotic inner slope, $\gamma$. In the top row of Figure 5.4, we show histograms of the posterior $\gamma$ values corresponding to the scans shown in Fig. 5.1 (parameter set I), i.e., for a true inner slope of $\gamma = 1$. The lines (bands) show the median (middle 68%) in each bin across the 10 realizations. The inner slope is generally poorly constrained for the smaller samples, with the median posterior distribution only weakly peaking around the true value of $\gamma = 1$ for the largest sample size of 10,000 stars—notably, even in this case, there is a non-negligible posterior probability at $\gamma = 0$, so we would not be able to exclude an incorrect inner slope value of 0 at high significance. We also
draw attention to the fact that, for the samples with 100 and 1000 stars, although the posterior distributions are fairly flat and poorly constrained, the posterior probability sharply drops off above $\gamma \sim 2$. This is important because the enclosed mass for a gNFW profile (Eq. 5.20) diverges at finite $r$ for $\gamma \geq 3$. For the samples with 20 stars, the fit is so statistics-limited that even unphysical values of $\gamma \geq 3$ cannot be fully excluded.

In the bottom row of Figure 5.4, we show the results for parameter set III, which has a true inner slope of $\gamma = 0$. The results are qualitatively similar: the posterior distributions of $\gamma$ tend to be poorly constrained for the smaller sample sizes, and we are only able to recover the true value of the inner slope for the 10,000-star samples. In this case, for the largest sample size, we would be able to exclude an incorrect inner slope value of 1 at high significance. However, for datasets with $\lesssim 1000$ stars from both parameter sets—on par with the existing dwarf galaxy measurements—we cannot determine whether the underlying halo has an inner slope of $\gamma = 0$ or $\gamma = 1$ in a statistically significant manner. We further note that, for the smaller sample sizes, the fact that the posterior $\gamma$ distributions are unconstrained implies that the results are highly sensitive to the choice of priors on $\gamma$, and we therefore choose to present the full posterior distributions rather than to quote recovered median values or quantiles.

A separate method for quantifying the ability of this procedure to distinguish whether the underlying DM distribution has an inner cusp or core is to compare the statistical preference for a cuspy DM model over a cored DM model, or vice versa. In particular, we analyze the same datasets as before, this time fixing the value of $\gamma$ in our model to either 1 or 0 in Eq. 5.17. The resulting models respectively correspond to the standard NFW distribution (Eq. 5.18) or the cored NFW distribution (Eq. 5.19). We then calculate the Bayes factor (BF) in preference for a model in which $\gamma$ is fixed.
I. $\gamma = 1$, $r_s = 1\text{kpc}$

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>BF$_{1,0}$</th>
<th>min(BF$_{1,0}$)</th>
<th>max(BF$_{1,0}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$1.27^{+0.70}_{-0.52}$</td>
<td>0.44</td>
<td>2.60</td>
</tr>
<tr>
<td>100</td>
<td>$1.18^{+0.74}_{-0.40}$</td>
<td>0.56</td>
<td>3.05</td>
</tr>
<tr>
<td>1000</td>
<td>$1.77^{+1.57}_{-0.96}$</td>
<td>0.67</td>
<td>6.19</td>
</tr>
<tr>
<td>10,000</td>
<td>$2.02^{+54.45}_{-1.52}$</td>
<td>0.22</td>
<td>225.70</td>
</tr>
</tbody>
</table>

Table 5.3: Values of the Bayes Factor (BF) from fitting parameter set I with a model assuming a cusp ($\gamma = 1$) relative to a model assuming a core ($\gamma = 0$). The second column lists the median and lower/upper 1σ, while the third(fourth) column lists the minimum(maximum) BF value across the 10 datasets for each sample size. Of the 10,000-star samples, two realizations have $10 \leq \text{BF}_{1,0} < 100$, providing strong evidence, and one realization has $\text{BF}_{1,0} \geq 100$, providing decisive evidence in favor of a cusp over a core.

to the true value for the given dataset, relative to a model in which $\gamma$ is fixed to the alternative value, i.e.,

$$BF = \frac{Pr(d|\gamma = \gamma_{\text{true}})}{Pr(d|\gamma = \gamma_{\text{alt.}})}.$$ (5.26)

On the Jeffreys scale, as amended by Ref. [316], $BF < 3.2$ is “not worth more than a bare mention,” $BF \in [3.2, 10)$ provides substantial evidence, $BF \in [10, 100)$ provides strong evidence, and $BF \geq 100$ provides decisive evidence.

In Table 5.3, we list for parameter set I the median and $\pm 1\sigma$ (second column) as well as the minimum (third column) and maximum (fourth column) BF values in preference for the true value of $\gamma = 1$ across the 10 datasets. For the smaller samples, the BF values are generally indeterminate, which is consistent with the relatively unconstrained posterior distributions shown in the top row of Fig. 5.4. For a sample size of 10,000 stars, the median BF is also indeterminate, although we find that there is one realization for which there is decisive evidence, and two realizations for which there is strong evidence, in favor of a model with a cusp. This is consistent
III. $\gamma = 0$, $r_s = 1$ kpc

<table>
<thead>
<tr>
<th>$n_{\text{stars}}$</th>
<th>BF$_{0,1}$</th>
<th>$\min(\text{BF}_{0,1})$</th>
<th>$\max(\text{BF}_{0,1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$0.58^{+0.25}_{-0.06}$</td>
<td>0.50</td>
<td>1.42</td>
</tr>
<tr>
<td>100</td>
<td>$0.58^{+0.19}_{-0.16}$</td>
<td>0.27</td>
<td>1.78</td>
</tr>
<tr>
<td>1000</td>
<td>$0.96^{+2.83}_{-0.60}$</td>
<td>0.17</td>
<td>6.78</td>
</tr>
<tr>
<td>10,000</td>
<td>$256.74^{+837.48}_{-235.60}$</td>
<td>2.98</td>
<td>46971.85</td>
</tr>
</tbody>
</table>

Table 5.4: Same as Table 5.3, but for parameter set III, in this case comparing a model assuming a core ($\gamma = 0$) to a model assuming a cusp ($\gamma = 1$). Of the 10,000-star samples, seven realizations have BF$_{0,1} \geq 100$, providing decisive evidence in favor of a cored distribution over a cuspy one.

with the rightmost panel in the top row of Fig. 5.4, in which the average posterior probability is non-negligible at $\gamma = 0$ and there is significant variation in the height of the peak at $\gamma \sim 1$ across realizations. Although there is significant scatter in the BF values between realizations, we emphasize that the BF in preference for the cored model over the cuspy one is always less than 10—the minimum benchmark for claiming statistically significant preference for a cored DM profile—and therefore, even in cases where we are unable to robustly identify the presence of a cusp, we would not falsely claim the presence of a core.

We list the analogous results for parameter set III in Table 5.4. In this case, the median BF for a sample size of 10,000 stars is decisively in favor of a model with a core. This is also consistent with the posterior distribution shown in the bottom rightmost panel of Fig. 5.4, which is peaked at $\gamma \sim 0$, sharply drops near $\gamma \sim 1$, and has relatively little spread across realizations. Importantly, across all sample sizes and realizations for parameter set I(III), for which the true DM profile is cuspy(cored), the BF in preference for a cored(cuspy) profile over a cuspy(cored) one is always less than 10. This demonstrates that, even when we are unable to recover statistical evidence
for the true inner DM profile, we would not erroneously claim evidence for the wrong inner profile.

Thus far, we have demonstrated that, for datasets with $\lesssim 1000$ measured stars—on par with the current measurements—we can robustly recover the total enclosed DM mass, but it can be challenging to accurately reconstruct the inner profile, and therefore to constrain the inner slope, of the DM density distribution, even within our simplified framework. We have also tested samples with 5000 stars and found that the posterior $\gamma$ distributions were typically flat as well, demonstrating that in order to constrain $\gamma$ in our setup, a sample size of $\sim 10,000$ (or at least $> 5000$) stars is truly needed. In Section 5.3.4, we explore how degeneracies between DM model parameters contribute to the difficulty of recovering $\gamma$. In Section 5.4, we investigate how the limitations on being able to accurately reconstruct the full density profile—which we emphasize is related to, but separate from, the issue of constraining the posterior distribution of $\gamma$—may affect the results of indirect detection analyses.

5.3.2 Velocity Uncertainties

Looking towards future measurements, it is important to understand how increasingly precise measurements of line-of-sight velocities might affect our ability to reconstruct DM halo properties. To address this, we generate simulated datasets assuming different values of measurement error $\Delta v$ (uniform across all generated stars), and repeat our analysis setting $\Delta v_i = \Delta v$ for all stars in Eq. 5.14. We compare our fiducial results, which assume a measurement error of $\Delta v = 2 \text{ km/s}$, to results assuming a more conservative value of $\Delta v = 5 \text{ km/s}$, as well as results in the limit of perfect measurements, $\Delta v = 0 \text{ km/s}$. As is clear from Eq. 5.14, if the intrinsic velocity dispersion of a system is much larger than $\sim 5 \text{ km/s}$, we do not expect such variations on $\Delta v$ to have a significant effect on the analysis results. On the other hand, if the intrinsic velocity
\[ \gamma = 0, \ r_\text{s} = 0.2 \text{kpc}, \ M_{200} \approx 1.1 \times 10^7 M_\odot \]

\[ \rho/\rho_{\text{True}}(r), \ \Delta v = 0 \text{ km/s} \]

\[ \rho/\rho_{\text{True}}(r), \ \Delta v = 2 \text{ km/s} \]

\[ \rho/\rho_{\text{True}}(r), \ \Delta v = 5 \text{ km/s} \]

\[ M/M_{\text{True}}(r), \ \Delta v = 0 \text{ km/s} \]

\[ M/M_{\text{True}}(r), \ \Delta v = 2 \text{ km/s} \]

\[ M/M_{\text{True}}(r), \ \Delta v = 5 \text{ km/s} \]

Figure 5.5: Fractional recovered density profiles (left panels) and enclosed mass profiles (right panels) for parameter set IV, varying over the line-of-sight velocity measurement error \( \Delta v \) as well as the sample size. For each sample size, the solid line denotes the median (across our 10 independent realizations) of the median fractional recovered profiles, while the shaded band shows the median of the 68% containment regions, plotted over the maximal radial range across all 10 datasets for that sample size. For the samples with fewer than 10,000 stars, we additionally extrapolate the median 68% containment regions over the full radial range, shown bracketed by each pair of dashed lines in the color corresponding to the sample size. Varying \( \Delta v \) has a particularly drastic effect on the smaller samples—for a sample size of 20 stars, a measurement error of \( \Delta v = 2 \text{ km/s} \) is insufficient for recovering the DM density and enclosed mass profiles.

dispersion is \( \lesssim 5 \text{ km/s} \), we expect the results to depend sensitively on the value of \( \Delta v \), especially if the sample size is small. For parameter sets I and III discussed in Section 5.3.1, the intrinsic velocity dispersion is \( \sim 10\text{-}15 \text{ km/s} \). Parameter sets II and IV have the same DM inner slope and concentration as parameter sets I and III, respectively, but are approximately 100 times less massive and have an intrinsic velocity dispersion of \( \sim 2\text{-}3 \text{ km/s} \).
In Figure 5.5, we show the inferred fractional DM density and enclosed mass profiles for parameter set IV.\textsuperscript{4} From top to bottom, the rows correspond to $\Delta v = 0, 2, 5\,\text{km/s}$. The results are consistent with our intuition: because the typical intrinsic velocity dispersion for this set of systems is $\sim 2\,\text{km/s}$, a measurement error $\Delta v \gtrsim 2\,\text{km/s}$ has a drastic effect on the inferred results, especially when combined with limited sample size. A value of $\Delta v = 5\,\text{km/s}$ results in an inferred virial mass of $M_{200} \sim 0$ for both the 20- and 100-star samples (see Tab. 5.2). For the 20-star samples, even our fiducial choice of $\Delta v = 2\,\text{km/s}$ results in essentially no DM being recovered. This can be understood as the measurement error accounting for the entirety of the observed velocity dispersion, and therefore there is no need for additional DM to explain the data.

\textsuperscript{4}We choose to present parameter set IV here because it has the smallest intrinsic velocity dispersion out of all of our parameter sets, and therefore is most drastically affected by increasing $\Delta v$. 

Figure 5.6: Same as Figure 5.5, but for parameter set III.
\( \gamma = 1, r_s = 1 \text{kpc}, M_{200} \approx 1.9 \times 10^9 M_\odot \)

\( \Delta v = 0 \text{ km/s} \)

\( \Delta v = 0 \text{ km/s} \)

\( \Delta v = 2 \text{ km/s} \)

\( \Delta v = 2 \text{ km/s} \)

\( \Delta v = 5 \text{ km/s} \)

\( \Delta v = 5 \text{ km/s} \)

\( R_{1/2} \)

\( R_{1/2} \)

Truth

20 stars

100 stars

1000 stars

10000 stars

Figure 5.7: Same as Figure 5.5, but for parameter sets I (top) and II (bottom).
For the larger sample sizes, with 1000 and 10,000 stars, the Jeans analysis is able to recover the correct density profile even as the measurement errors are of the same order as the dispersion of the system. This can be attributed to the fact that with large enough statistics, the analysis can distinguish the radially-dependent velocity dispersion $\sigma_p(R)$ from the radially-independent measurement error.

For parameter set III, which has the same DM inner slope and concentration as parameter set IV but is 100 times more massive, varying the measurement error has negligible effect on the inferred DM halo properties as expected (shown in Fig. 5.6). The results for parameter sets I and II (which have $\gamma = 1$) are qualitatively similar to the results for parameters sets III and IV (which have $\gamma = 0$), although quantitatively different due to slightly larger values of the intrinsic velocity dispersion; we present those results in Fig. 5.7.

Similarly, $\Delta v$ affects the recovery of the inner slope more for the less massive halos than for the more massive ones. Figure 5.8 shows the posterior $\gamma$ distributions corresponding to the scans shown in Fig. 5.5. As $\Delta v$ is increased, $\gamma$ becomes increasingly unconstrained for the smaller sample sizes, whereas for the 10,000-star samples, increasing $\Delta v$ appears to lead to biases in the best-fit value of $\gamma$. For the more massive halo with the same inner slope (parameter set III), the posterior $\gamma$ distributions are mostly insensitive to these variations in the measurement error (shown in Fig. 5.9). The corresponding posterior $\gamma$ distributions for parameter sets I and II are shown in Figs. 5.10–5.11, and are qualitatively similar to the cases of parameter sets III and IV, respectively.

### 5.3.3 Location of stars

In this section, we are interested in understanding how the DM inference depends not only on how many stars are measured, but also on where the observed stars are
Figure 5.8: Posterior distributions for the inner slope $\gamma$, varying over the line-of-sight velocity measurement error $\Delta v$ as well as the sample size, shown for parameter set IV. These posteriors correspond to the results shown in Fig. 5.5. The lines(bands) show the median(middle 68%) in each $\gamma$ bin across the 10 realizations. Varying $\Delta v$ has a drastic effect on the inference of $\gamma$ in this case; in contrast, for the more massive halo of parameter set III, varying the measurement error has negligible effect (Fig. 5.9). All panels in this figure share the same vertical scale.

within the dwarf galaxy. To study this effect, we start with our datasets of initial size $n_{\text{stars}}$ and apply the following selection functions, then repeat our analysis on the resulting datasets:

- Inner stars analysis: keep only the stars in the inner region, with projected radius $R < R_{1/2}$.
- Outer stars analysis: keep only the stars in the outer region, with projected radius $R > R_{1/2}$.
To account for the ∼ 50% change in the number of stars from implementing these selection functions, we compare the results to “benchmark” results on datasets with \( n_{\text{stars}}/2 \) stars which are also generated from the original \( n_{\text{stars}} \)-star datasets, subsampled uniformly to preserve the radial probability distribution of the original dataset. In doing so, we can compare the results for datasets that have approximately equal numbers (∼ \( n_{\text{stars}}/2 \)) but distinct spatial distributions of stars.

As before, we generate 10 independent datasets for each selection function. In Figure 5.12, we show the distributions of the projected radius \( R \) as well as the 3d radius \( r \) for parameter set III with \( n_{\text{stars}} = 100 \) (which is qualitatively representative of the distributions for all the parameter sets and sample sizes), for the three different selection functions. We note that, because we implement the selection function on the projected radius, and \( r \geq R \) for all values of \( R \), the \( R < R_{1/2} \) datasets extend...
Figure 5.10: Same as Figure 5.8, but for parameter set I.

slightly beyond a 3d radius of \( r = R_{1/2} \). We test the effect of selection functions on datasets with initial sizes of \( n_{\text{stars}} = 100, 1000, \) and 10,000 stars for each of the four parameter sets. For the purpose of studying the effects of spatial distributions in the cleanest setup, the studies presented in this section have been performed assuming \( \Delta v = 0 \text{ km/s} \).

In Figure 5.13, we show the recovered DM density and enclosed mass profiles for the three different selection functions, for a particularly demonstrative example. This example is for parameter set IV, with an initial sample size of \( n_{\text{stars}} = 100 \) stars; after applying each of the selection functions, we end up with a selected sample size of \( n_{\text{sel}} \sim 50 \) stars. From lightest to darkest, we show the results for the benchmark, \( R < R_{1/2} \), and \( R > R_{1/2} \) datasets. Like before, the solid lines denote the median of the median recovered profiles across the 10 realizations, while the shaded bands
Figure 5.11: Same as Figure 5.8, but for parameter set II.

depict the median of the 68% containment ranges across the realizations. For ease of presentation, we choose in this case to show the solid line and band for each selection function from the median $r_{\text{min}}$ to the median $r_{\text{max}}$ across the realizations, where $r_{\text{min}}(r_{\text{max}})$ is the galactocentric distance of the innermost(outermost) star in each individual realization. Beyond this range, we extrapolate the inferred profiles, shown by each pair of dashed lines in the color corresponding to the selection function.

As expected, when the measured stars are all in the outer region of the dwarf, the DM profile is poorly constrained at small radii compared to the benchmark scenario. Conversely, when the measured stars are all in the inner region of the dwarf, the DM profile is poorly constrained at larger radii. Interestingly, for the $R < R_{1/2}$ samples in this example, the DM profile is also typically less well-constrained at small radii; additionally, the density profile is biased high at small radii and low at large radii, to
Figure 5.12: Histograms of the projected radius $R$ (left panel) and the 3d radius $r$ (right panel) for parameter set III starting with a sample size of $n_{\text{stars}} = 100$ stars, resulting in selected samples of $n_{\text{sel}} \sim 50$ stars, for the three different spatial selection functions. For each selection function, the line denotes the median counts per bin and the band shows the 68% containment across 10 realizations. The recovered DM density and enclosed mass profiles corresponding to these datasets are shown in Figure 5.13.

the extent that the total enclosed mass is also biased low (the recovered virial mass is $M_{200} \sim 0.2^{+0.7}_{-0.1} \times 10^7 M_\odot$, while the true value is $M_{200} \sim 1.1 \times 10^7 M_\odot$). These biases, as well as the larger uncertainties on the DM profile in both the inner and outer regions, are present in spite of there being approximately twice as many stars within the half-light radius in the $R < R_{1/2}$ datasets as in the benchmark datasets. In this particular example, the posterior $\gamma$ distribution is unconstrained for all three selection functions due to the small size of the dataset, so we do not recover a corresponding bias in $\gamma$.

The specific behavior of the results for the $R < R_{1/2}$ samples noted in this example is not generic to all of the variations we have tested—in particular, for the datasets with larger selected sample size $n_{\text{sel}}$, the bias in the DM density profile is less severe, and in some cases the median 68% containment band on the inner density profile is slightly narrower than in the benchmark case. This can be seen in Figure 5.14, which
Figure 5.13: Inferred DM density profiles $\rho(r)$ (left panels) and corresponding enclosed mass profiles $M(r)$ (right panels) for parameter set IV, starting with a sample size of $n_{\text{stars}} = 100$ stars (resulting in selected samples of $n_{\text{sel}}^{\text{stars}} \sim 50$ stars), with spatial selection functions applied. From lightest to darkest, the results are for the benchmark datasets, the datasets keeping only stars with $R < R_{1/2}$, and the datasets keeping only stars with $R > R_{1/2}$. We show the full reconstructed distributions in the top panels as well as the fractional (relative to truth) distributions in the bottom panels. For each selection function, the solid line denotes the median (across our 10 independent realizations) of the median recovered profiles, while the shaded band shows the median of the 68\% containment regions; these are plotted from the median $r_{\text{min}}$ to the median $r_{\text{max}}$ across the 10 realizations, where $r_{\text{min}}(r_{\text{max}})$ is the galactocentric radius of the innermost(outermost) star. We additionally extrapolate the median 68\% containment regions over the full radial range of the benchmark samples, shown bracketed by each pair of dashed lines in the color corresponding to the selection function. As expected, compared to the benchmark scenario, when the measured stars are all in the outer region of the dwarf, the DM profile is poorly constrained at small radii. Conversely, when the measured stars are all in the inner region of the dwarf, the DM profile is poorly constrained at larger radii. Moreover, the inner density profile is also less well-constrained for the $R < R_{1/2}$ case than for the benchmark scenario, suggesting that in order to constrain the inner DM profile, it is important to have measured stars across the full radial distribution, and not only in the inner region.
\( \gamma = 0, \ r_s = 0.2 \text{kpc}, \ M_{200} \approx 1.1 \times 10^7 M_\odot \)

Figure 5.14: Same as Figure 5.13, but starting with a sample size of \( n_{\text{stars}} = 1000 \) stars, resulting in selected samples of \( n_{\text{sel}} \approx 500 \) stars. In this case, the bias on the density profile for the \( R < R_{1/2} \) case is less severe than in Fig. 5.13, and the uncertainty on the inner density profile is smaller for the \( R < R_{1/2} \) case than for the benchmark case.

is the same as Fig. 5.13, except for an initial sample size of \( n_{\text{stars}} = 1000 \), i.e., for spatially selected datasets of size \( n_{\text{sel}} \approx 500 \).

We can quantitatively compare the performance of the different selection functions, for different sample sizes \( n_{\text{sel}} \), by comparing the recovered virial mass estimates as well as the recovered \( J \)-factors (discussed more in Sec. 5.4), both detailed in Table 5.15. Across our four parameter sets, the results on spatial selection functions are the following:

- Inner stars analysis \((R < R_{1/2})\)

  - For the smallest sample size \( n_{\text{sel}} \approx 50 \), for all parameter sets, the inferred virial mass is systematically underestimated (inconsistent with the true value within 1\( \sigma \) uncertainty for three of the four parameter sets). This
becomes less severe as the sample size is increased, but across all four parameter sets for the larger sample sizes $n_{\text{stars}}^{\text{sel}} \sim 500$ and $n_{\text{stars}}^{\text{sel}} \sim 5000$, the uncertainty on the estimated virial mass is consistently larger than for either the $R > R_{1/2}$ datasets or the benchmark case, demonstrating that to achieve an accurate virial mass estimate, it is important to have measurements of outer stars.

- The behavior of the posterior $\gamma$ distribution varies across different sample sizes and different parameter sets—in some cases, the posterior $\gamma$ distribution is biased high when the selection function is applied; in other cases, it is unchanged from the posterior distribution in the benchmark case. In all cases, the $R < R_{1/2}$ selection function does not improve the ability of the method to accurately constrain $\gamma$, relative to the benchmark case. Therefore, for the purpose of constraining $\gamma$, additional stars need to be measured across all radii.

- As we will discuss in Sec. 5.4, for the smallest sample size $n_{\text{stars}}^{\text{sel}} \sim 50$, for all parameter sets, the uncertainty on the $J$-factor estimate is larger than in the benchmark case. For the larger sample sizes, the uncertainty on the $J$-factor estimate is comparable to or slightly ($O(0.1 \, \text{dex})$) smaller than in the benchmark case.

- Outer stars analysis ($R > R_{1/2}$)

  - For all parameter sets and all sample sizes $n_{\text{stars}}^{\text{sel}}$, the estimated virial mass is consistent with the true value, and the uncertainty on the virial mass estimate is comparable to or slightly smaller than in the benchmark case, demonstrating that having measurements of inner stars is not crucial to the recovery of the virial mass.
- For all parameter sets and all sample sizes $n_{\text{stars}}^{\text{sel}}$, the posterior $\gamma$ distribution is comparable to (when the benchmark posterior distribution is already unconstrained) or less constrained than in the benchmark case.

- As we will discuss in Sec. 5.4, for all parameter sets and all sample sizes $n_{\text{stars}}^{\text{sel}}$, the uncertainty on the $J$-factor estimate is comparable to or larger than in the benchmark case, indicating that having measurements of inner stars is important for the purpose of constraining $J$-factors.

While the $R > R_{1/2}$ datasets perform slightly better in terms of the uncertainty on the recovered virial mass relative to the two other selection functions, the improvement is marginal (see Table 5.15 for values). Therefore, based on the overall performance at inferring the full DM density profile and the inner slope $\gamma$, especially for the smallest samples, we find that it is crucial to have measurements of stars across the full radial distribution of the dwarf galaxy. Doing so allows the fit to anchor the DM distribution across the full radial range, and consistently results in comparable or better performance at accurately reconstructing both the inner and outer profile of the DM distribution, relative to the cases when the data consists purely of stars in either the inner or outer region of the system.

5.3.4 Role of Degeneracies

One of the factors that limits the accurate recovery of the inner slope of the DM density profile is degeneracy between the different halo parameters—different combinations of the normalization $\rho_0$, scale radius $r_s$, and inner slope $\gamma$ can result in similar enclosed mass profiles, and therefore are equally valid descriptors of the kinematic data. This is manifest in the left half of Figure 5.16, an example triangle plot of the posterior halo parameters from analyzing a single 10,000-star dataset. In this clean example, the fit converges near the true values of $\rho_0$, $r_s$, and $\gamma$, but there are
Table 5.15: Inferred values of the virial mass $M_{200}$ and $J$-factor for the different parameter sets, selected sample sizes, and spatial selection functions, with $\Delta v = 0$ in all cases. The $J$-factors are in units of $\log_{10}\left[\frac{J(0.5^\circ)}{\text{GeV}^2\text{cm}^{-5}}\right]$. Each entry in this table represents the median across 10 realizations of the median and $\pm 1\sigma$ values.
Figure 5.16: Left: An example triangle plot of the posterior DM parameters from a scan of a 10,000-star sample from parameter set I, with $\Delta v = 0$ km/s. While the parameters are converged about their true values (red lines), there are significant degeneracies between pairs of parameters. Right: An example triangle plot of the posterior DM parameters from a scan of a 100-star sample from parameter set I, with $\Delta v = 0$ km/s. Compared to the posteriors from the larger sample size shown on the left, the parameters are much more poorly constrained in this case (note the wider axis ranges in this case), and $\gamma$ is unconstrained at low values.

It is increasingly difficult to constrain the value of $\gamma$ as the sample size is decreased.

For comparison, the right half of Figure 5.16 shows an example triangle plot of the posterior halo parameters from analyzing a single 100-star dataset. Once again, there are clear degeneracies between the pairs of parameters. In this case, all of the DM parameters are more poorly constrained (note the wider axis ranges compared to to 10,000-star case), and in particular the posterior $\gamma$ distribution is almost entirely flat down to the lower edge of our prior range. We emphasize that our choice of parameterization for the DM distribution is simpler than the Hernquist/Zhao parameterization widely employed in the literature [308, 317, 318], which has five parameters. Given the extra degrees of freedom in that model, the role of degeneracies
would present an even bigger challenge for constraining the inner slope of the DM distribution, especially in the case of statistics-limited datasets.

We can explicitly break the degeneracies in our halo model by holding $\rho_0$ or $r_s$ (or both) fixed to their true values and examining the resulting posterior distributions for $\gamma$. We show the results for parameter set I(III), for a sample size of 100 stars and $\Delta v = 2$ km/s, in the top(bottom) panel of Figure 5.17. Fixing $\rho_0$ (second column) or fixing $r_s$ (third column) results in a posterior $\gamma$ distribution which is peaked near the true value of $\gamma$, with slightly more constraining power in the case of fixing $\rho_0$. This makes sense intuitively because the inner region of the DM distribution is directly sensitive to $\rho_0$ and $\gamma$, whereas $r_s$ more directly influences the distribution at intermediate radii, and therefore breaking the degeneracy between the former two parameters is more effective at improving the constraint on $\gamma$. If we fix both $\rho_0$ and $r_s$ to their respective true values (fourth column), we recover the true inner slope with high accuracy.

We have thus demonstrated that, even for our simplified mock datasets and three-parameter DM halo model, the dimensionality of the problem is large enough that constraining the inner slope of the DM density profile for moderately sized stellar samples proves to be difficult. These challenges would be further exacerbated when one takes into account velocity anisotropy, which is difficult to accurately model and is also degenerate with the mass profile [300, 301, 302, 303, 304, 305].

While it may not be well-motivated to hold DM halo parameters fixed in an analysis on real data, one could ameliorate the effect of parameter degeneracies by setting model-informed priors on the halo parameters [297, 318]—for example, if one were to assume a specific mass-concentration relation, then there would be a specific relation between the normalization $\rho_0$ and scale radius $r_s$, and the priors for those parameters would no longer be independent of each other. Additionally,
Figure 5.17: Posterior $\gamma$ distributions for 100-star samples from parameter sets I (top row) and III (bottom row), with $\Delta v = 2 \text{ km/s}$. The lines (bands) show the median (middle 68%) in each $\gamma$ bin across the 10 realizations. We show the results for our fiducial setup (first column), fixing $\rho_0$ to its true value (second column), fixing $r_s$ to its true value (third column), or fixing both $\rho_0$ and $r_s$ to their respective true values (fourth column). Breaking the degeneracies between $\rho_0$, $r_s$, and $\gamma$ by holding $\rho_0$ and/or $r_s$ fixed gives rise to improved constraining power on $\gamma$.

Ref. [319] recently demonstrated that non-spherical mass models can alleviate the effect of parameter degeneracies.

A separate approach to mitigating the effect of parameter degeneracies is to jointly analyze multiple dwarf galaxies at once, under the assumption that the systems share certain properties—in the simplest case, one could assume that the systems all share the same value of $\gamma$. We discuss the joint analysis approach in more detail in Section 5.7.2. We note that a thorough study of the joint analysis method is computationally infeasible within our current analysis framework, because the dimensionality of the model quickly grows with the number of jointly analyzed systems, to the point that it is highly inefficient to use traditional MCMC or nested sampling methods to sample the posterior probability distributions.
Figure 5.18: $J$-factors as a function of the sample size and the measurement error on the line-of-sight velocities, $\Delta v$, for parameter sets I (top left), II (top right), III (bottom left), and IV (bottom right). We take the distance to the dwarf in each case to be 50 kpc and the angle of integration to be $0.5^\circ$. The results for $\Delta v = 2$ km/s in the top(bottom) left panel correspond to the recovered DM profiles shown in Fig. 5.1(5.3). For each realization of a given sample size and value of $\Delta v$, we build up a posterior $J$-factor distribution by calculating the $J$-factor for every set of posterior parameters, from which we can calculate the median and middle 68\% containment range of $\log_{10}[J(0.5^\circ)/(\text{GeV}^2 \text{ cm}^{-5})]$ for that realization. Each data point denotes the median across the 10 realizations of the median $\log_{10}[J(0.5^\circ)/(\text{GeV}^2 \text{ cm}^{-5})]$, and each set of error bars brackets the median of the 68\% containment across the realizations.

5.4 Implications for Indirect Detection

In this section, we cast the results of our study into the context of indirect detection by calculating the inferred $J$-factors for the tests discussed in Section 5.3, using the public code CLUMPY [320, 321, 322] to the perform the $J$-factor computations (as
defined in Eq. 5.1). We examine the effects of sample size and line-of-sight velocity measurement error (Sec. 5.4.1), choices of priors in the Jeans analysis (Sec. 5.4.2), and spatial selection functions (Sec. 5.4.3) on the inferred \(J\)-factors. In Sec. 5.4.4, we discuss the implications of our findings on indirect detection results and make recommendations for future observations.

### 5.4.1 Sample Size and Measurement Error

First, we examine the effects of sample size and velocity measurement error, \(\Delta v\), on the inferred \(J\)-factors. In the left column of Figure 5.18, we show the inferred \(J\)-factors for parameter sets I (top) and III (bottom), for which \(M_{200} \sim 10^9 M_\odot\), for the different sample sizes and values of \(\Delta v\). For an individual scan, we evaluate the \(J\)-factor for each set of posterior parameters, assuming a distance of 50 kpc to the dwarf. Each data point in Fig. 5.18 shows the median across our 10 realizations of the median and middle 68% containment range for the inferred values of \(\log_{10}[J(0.5)/(\text{GeV}^2 \text{cm}^{-5})]\). Within each cluster of three data points corresponding to a particular sample size, the blue circle, teal triangle, and green square show the results for \(\Delta v = 0, 2, 5 \text{ km/s}\), respectively.

As expected, the uncertainties on the \(J\)-factor decrease as a function of increasing sample size. Additionally, the \(J\)-factors are nearly independent of \(\Delta v\), which is expected for parameter sets I and III (see Sec. 5.3.2 for a discussion on the effects of \(\Delta v\)). For parameter set I, our estimates of the \(J\)-factor are on average consistent with the true value for all sample sizes and values of \(\Delta v\). For parameter set III, our estimates of the \(J\)-factor are systematically biased high, although the median values are within a factor of 2 of the true values for the 1000- and 10,000-star samples—this is consistent with the inner density profiles being biased high for the smaller samples.

---

\(^5\)We have verified that qualitatively, our results on the \(J\)-factor uncertainty are unchanged if we instead assume a distance of 100 kpc to the dwarf.
as shown in Fig. 5.3. The typical values of the $J$-factor we recover for the different combinations of parameter set, sample size, and $\Delta v$ are tabulated in Table 5.2.

For parameter sets II and IV (shown in the top right and bottom right panels of Fig. 5.18, respectively), the $J$-factor estimates are highly sensitive to $\Delta v$, in a manner that is consistent with the results discussed in Sec. 5.3.2 (the corresponding fractional recovered density and enclosed mass profiles are shown in the bottom half of Fig. 5.7 for parameter set II and Fig. 5.5 for parameter set IV). In particular, the data points that extend below the range of the right panels correspond to the cases of larger $\Delta v$ where the recovered DM abundance is significantly underestimated.

### 5.4.2 Dependence on Priors

It is crucial to emphasize the dependence of the $J$-factor inference on the priors assumed for the DM halo parameters. The uncertainties on our inferred $J$-factors are notably larger than values commonly quoted in the literature for the data, such as the ones found in Ref. [147], hereafter GS15, which were used to derive the constraints on DM annihilation by the Fermi-LAT collaboration in Ref. [148], hereafter A17. The $J$-factors from GS15 are listed in Table 5.5 for reference. For example, Ursa Major II, which has a sample size of 20 stars, is quoted to have a $\sim \pm 0.5$ uncertainty on $\log_{10}[J(0.5^\circ)/(\text{GeV}^2\text{cm}^{-5})]$, whereas on average, the uncertainties on $\log_{10}[J(0.5^\circ)/(\text{GeV}^2\text{cm}^{-5})]$ for our 20-star samples span $\sim \pm 1$–2 (when the values of $\Delta v$ are sufficiently small for the DM to be recovered). This discrepancy is especially surprising because GS15 models the DM density distribution with the Hernquist/Zhao profile, which has two additional slope parameters compared to the gNFW model we use, and additionally models the velocity anisotropy—a model with more free parameters, combined with the added degeneracy between the anisotropy and mass profiles, should give rise to larger uncertainties on the inferred $J$-factors.
Figure 5.19: Comparing the $J$-factor results for each of the ten 20-star datasets in parameter set III: using our fiducial analysis setup (dark purple squares), using the DM priors from Ref. [147] (medium purple triangles), and using the priors on the normalization $\rho_0$ and scale radius $r_s$ from Ref. [147] while setting the prior on the inner slope to be $\gamma \in [-1, 3]$ (light purple circles). This demonstrates that for these small sample sizes, the results are highly prior-dependent, which is consistent with our findings that the DM profile and inner slope are poorly constrained for the datasets with limited statistics. This additionally demonstrates that the remarkably small uncertainties on the $J$-factors from [147] for the ultrafaint dwarfs may be driven by their narrow choice of prior on $\gamma$.

The primary source of this apparent discrepancy is that in this work, we have assumed wider prior ranges on the halo parameters than what was assumed in GS15—in particular, the analysis in GS15 assumed a prior of $0 \leq \gamma \leq 1.2$ on the inner DM slope. When we repeated our analysis assuming the same priors on $\rho_0$, $r_s$, and $\gamma$ as the ones used in GS15, the uncertainties on our $J$-factors decreased significantly. In Figure 5.19, we show the median and middle 68% range on $\log_{10}[J(0.5^\circ)/(\text{GeV}^2 \text{cm}^{-5})]$ for each of our 10 different 20-star datasets from parameter set III (which is representative of the results for all parameter sets), assuming either our fiducial setup (squares) or the priors from GS15 (triangles). Implementing the GS15 priors reduced the $J$-factor uncertainty in all 10 datasets, by as much as a factor of $\sim 3$ in certain cases.
GS15 additionally takes the best-fit Plummer radius from the literature and fixes it in their fit. Analogously, we have also repeated our analysis fixing our light-profile parameters to their best-fit values while assuming the GS15 priors and found the results to be essentially unchanged from the case of GS15 priors without fixing light-profile parameters. Furthermore, GS15 truncates the $J$-factor integration at $r_{\text{max}}$, the galactocentric distance of the outermost star. We have tested this prescription as well, and found that it makes negligible difference to our values of the $J$-factor. This is expected, because the $J$-factor within the inner $0.5^\circ$ is dominated by the most central regions of the DM halo, and is therefore insensitive to the outer truncation radius of the integration.

As an additional test, we set the priors on the normalization $\rho_0$ and scale radius $r_s$ for the DM profile to the GS15 priors, but rather than using the GS15 prior of $\gamma \in [0, 1.2]$ on the inner slope, we assume a wider prior range of $\gamma \in [-1, 3]$, which is equivalent to our fiducial prior range with the exclusion of the unphysical values of $\gamma > 3$. This directly tests how a wider prior range on $\gamma$ affects the inferred $J$-factor. The results of this test are shown by the circles in Fig. 5.19, and are similar to our fiducial results (squares), indicating that the narrow prior range on $\gamma$ is indeed what primarily drives our fit to reproduce the small $J$-factor uncertainties found in GS15.

We have also verified that implementing the GS15 priors (with and without fixing the light profile parameters) on our 1000-star samples decreases the uncertainty on our estimated values of $\log_{10}[J(0.5^\circ)/(\text{GeV}^2\text{cm}^{-5})]$ by a factor of $\sim 2$, making them broadly consistent with the uncertainties quoted in GS15 for the classical dwarfs.

An important takeaway from this exercise is that the $J$-factors inferred through the Jeans modeling procedure, for the currently accessible stellar sample sizes, depend sensitively on prior assumptions on $\gamma$, and therefore should be treated with caution. Motivated by the prior-dependence of $J$-factor estimates from Jeans analyses, a com-
plemetary method that has been proposed in the literature is a frequentist approach to deriving $J$-factors [323, 324], which removes the prior-dependence but also loses the ability to construct full posterior probability distributions of the DM inner slope.

5.4.3 Spatial Selection

We can revisit the discussion of spatial selection functions detailed in Section 5.3.3, in the context of $J$-factors. In Sec. 5.3.3, we found that if we implemented a selection function of $R > R_{1/2}$, i.e., only included stars in the outer regions of the system, the resulting inferred DM density profile was more uncertain in the inner regions of the dwarf than in the benchmark scenario. We also found that if we implemented a selection function of $R < R_{1/2}$, i.e., only included stars in the inner regions of the system, the inferred DM density profile was more uncertain in the outer regions of the dwarf than in the benchmark scenario. Furthermore, for the $R < R_{1/2}$ datasets, we found that the inner profile could be biased high, especially when the sample size was small. The degree of such biases and increased uncertainties on the DM density profile can be quantitatively captured by evaluating the $J$-factor. These results are shown in Figure 5.20 and detailed in Table 5.15.

Overall, we find that for the datasets with $\lesssim 50$ observed stars (comparable to the current sample sizes of ultrafaint dwarfs), observing more stars which are distributed across the full range of the radial distribution would have the most potential to decrease the uncertainty on estimates of the $J$-factors. This is demonstrated in the bottom panel of Fig. 5.20, in which the uncertainties on the $J$-factor are always smaller in the benchmark case (green squares) than for either of the other two cases (teal triangles and blue circles). For the systems with hundreds or thousands of observed stars, there is room for slight improvement on the accuracy of inferred $J$-factors by measuring more stars in the inner regions of the systems. This is demonstrated in

197
Figure 5.20: $J$-factors as a function of the selected sample size, for all parameter sets and spatial selection functions. From top to bottom, the panels show the results for $n_{\text{stars}}^{\text{sel}} \sim 5000$, 500, 50. From left to right, each cluster of three data points shows the results for parameter set I, II, III, and IV. Each data point denotes the median across the 10 realizations of the median $\log_{10}[\frac{J/J_{\text{true}}(0.5^\circ)}{\text{GeV}^2\text{cm}^{-5}}]$, and each set of error bars brackets the median of the 68\% containment across the realizations. For the smallest samples, the benchmark case consistently has smaller uncertainties than either the $R > R_{1/2}$ or $R < R_{1/2}$ cases. For some of the larger samples, the uncertainties are slightly reduced ($O(0.1 \, \text{dex})$ smaller) for $R < R_{1/2}$ relative to the benchmark case. Note the different $y$-axis scale for the bottom panel.

the top and middle panels of Fig. 5.20, in which the uncertainties can be somewhat smaller ($O(0.1 \, \text{dex})$) for the $R < R_{1/2}$ datasets (teal triangles) than for the benchmark datasets (green squares) or $R > R_{1/2}$ datasets (blue circles). As expected, the behavior of the recovered $J$-factors is consistent with the ability of the Jeans modeling to accurately recover the inner density profile of the DM, as was discussed in Sec. 5.3.
5.4.4 Dwarfs in Need of More Measurements

Within the literature, there are two approaches to dwarf galaxy indirect detection analyses—individual dwarfs may be analyzed on their own [325, 326, 327, 328, 329, 330], or many systems may be stacked to obtain a more competitive limit on DM annihilation [281, 282, 160, 148, 279, 149]. In both cases, achieving robust indirect detection results is dependent upon accurately estimating the $J$-factors for the dwarfs that dominate the limits. Table 5.5 lists the confirmed dwarf galaxies used in the analysis from A17, in order of decreasing $J$-factor. We emphasize that while the dwarfs that give rise to the strongest constraints on DM annihilation are among those with the largest $J$-factors, having a larger $J$-factor does not necessarily imply that the resulting limit from a given dwarf will be stronger, due to effects such as different levels of background contamination in different regions of the sky. In the following discussion, we will emphasize future observations which are important for obtaining more accurate estimates of the $J$-factors for the systems that dominate the A17 results.

As shown in Fig. 5.18 and detailed in Tab. 5.2, the typical uncertainty on $\log_{10}[\frac{J(0.5^\circ)}{(\text{GeV}^2 \text{cm}^{-5})}]$ from our analysis is $\sim \pm 1$–2 for 20-star systems and $\sim \pm 0.5$–1 for 100-star systems (excluding the cases of small intrinsic dispersion and large $\Delta v$ where the fit drastically underestimates the abundance of DM), as opposed to $\sim \pm 0.5$ and $\sim \pm 0.2$, respectively, from GS15 (listed in Tab. 5.5). We determined in Sec. 5.4.2 that this discrepancy may be due to different prior choices on $\gamma$. To test the effect of larger $J$-factor uncertainties on the resulting indirect detection constraints on DM annihilation, we can use the likelihood functions provided in A17 to derive limits assuming different values of the $J$-factor uncertainty. Similarly to A17, we use Eq. 3 of Ref. [160] to profile over the $J$-factor uncertainty.

http://www-glast.stanford.edu/pub_data/1203/
The three dwarfs from A17 that provide the strongest limits in the mass range relevant for the DM interpretation of the Galactic Center Excess (GCE) are Ursa Major II, Ursa Minor, and Draco. We first focus on Ursa Major II, which has a sample size of 20 stars. We find that increasing the uncertainty on \( \log_{10}[J(0.5^\circ) / \text{GeV}^2 \text{cm}^{-5}] \) from 0.4 (which was assumed in the A17 analysis) to 1 weakens the limit by a factor of \( \sim 5-8 \) in the 10–100 GeV mass range for the \( b\bar{b} \) annihilation channel, resulting in

\[
\frac{J(0^\circ)}{(\text{GeV}^2 \text{cm}^{-5})} = \begin{array}{c}
\text{Dwarf} & N \text{ Stars} & \log_{10} J(0.5^\circ) & \text{Dispersion} & \text{References} \\
\text{Ursa Major II} & 20 & 19.42^{+0.44}_{-0.42} & 5.6^{+1.4}_{-1.4} & [331] \\
\text{Segue 1} & 70 & 19.36^{+0.32}_{-0.35} & 3.7^{+1.4}_{-1.1} & [332] \\
\text{Coma Berenices} & 59 & 19.02^{+0.37}_{-0.41} & 4.6^{+0.8}_{-0.8} & [332] \\
\text{Ursa Minor} & 313 & 18.93^{+0.27}_{-0.19} & 9.5^{+1.2}_{-1.2} & [333] \\
\text{Draco} & 292 & 18.84^{+0.12}_{-0.13} & 9.1^{+1.2}_{-1.2} & [333] \\
\text{Sculptor} & 1365 & 18.54^{+0.06}_{-0.05} & 9.2^{+1.1}_{-1.1} & [334] \\
\text{Bootes I} & 37 & 18.24^{+0.40}_{-0.37} & 4.6^{+0.8}_{-0.6} & [335] \\
\text{Leo II} & 126 & 17.97^{+0.20}_{-0.18} & 7.4^{+0.4}_{-0.4} & [336] \\
\text{Carina} & 774 & 17.87^{+0.10}_{-0.09} & 6.6^{+1.2}_{-1.2} & [334] \\
\text{Ursa Major I} & 39 & 17.87^{+0.56}_{-0.33} & 7.0^{+1.0}_{-1.0} & [331] \\
\text{Leo I} & 267 & 17.84^{+0.20}_{-0.16} & 9.2^{+0.4}_{-0.4} & [337] \\
\text{Fornax} & 2483 & 17.83^{+0.12}_{-0.06} & 11.7^{+0.9}_{-0.9} & [334] \\
\text{Canes Venatici II} & 25 & 17.65^{+0.45}_{-0.43} & 4.6^{+1.0}_{-1.0} & [332] \\
\text{Sextans} & 441 & 17.52^{+0.28}_{-0.18} & 7.9^{+1.3}_{-1.3} & [334] \\
\text{Canes Venatici I} & 214 & 17.43^{+0.37}_{-0.28} & 7.6^{+0.4}_{-0.4} & [332] \\
\text{Leo T} & 19 & 17.11^{+0.44}_{-0.39} & 7.5^{+1.6}_{-1.6} & [332] \\
\text{Hercules} & 30 & 16.86^{+0.74}_{-0.68} & 5.1^{+0.2}_{-0.2} & [332] \\
\text{Leo V} & 5 & 16.37^{+0.94}_{-0.87} & 2.3^{+3.2}_{-1.6} & [338] \\
\text{Leo IV} & 18 & 16.32^{+1.06}_{-1.69} & 3.3^{+1.7}_{-1.7} & [332] \\
\text{Segue 2} & 25 & 16.21^{+1.06}_{-0.98} & < 2.2 & [339] \\
\end{array}
\]

Table 5.5: List of dwarf galaxies used in Ref. [148] ordered by decreasing \( J \)-factor. The observed numbers of stars and \( J \)-factors are compiled from Ref. [147]. The dispersions are compiled from Ref. [331].
a limit that no longer excludes or is in tension with the regions of parameter space consistent with the GCE from Refs. [126, 130, 132, 129]. Similarly, for Draco and Ursa Minor (∼ 300 observed stars each), we find that increasing the uncertainty on $\log_{10}[J(0.5^\circ)/(\text{GeV}^2\text{cm}^{-5})]$ from their assumed values in A17 of 0.1 and 0.2, respectively, to 0.5(1) weakens the limit by a factor of ∼ 2(10). We note that a factor of ∼ 2 weakening of the strongest dwarf limits is sufficient to significantly reduce the tension with the DM interpretation of the GCE.

This demonstrates that for the current observed sample sizes, the dwarf galaxy indirect detection limits can be highly sensitive to the assumed priors for the inner DM slope $\gamma$. In order to derive robust indirect detection constraints from the dwarf galaxies, it is crucial to increase the number of observed stars in order to obtain more robust estimates of their $J$-factors. In particular, we emphasize the importance of increasing the sample sizes for Ursa Major II, Ursa Minor, and Draco, which dominate the indirect detection limits. Our results in Secs. 5.3.3 and 5.4.3 suggest that measuring more stars spanning the entire spatial extent of the galaxies would be most effective at achieving more accurate estimates of their $J$-factors (see bottom panel of Fig. 5.20). If sample sizes are increased beyond $\gtrsim 500$ stars, our findings suggest that focusing on measuring more stars in the inner regions of the dwarfs may provide additional constraining power on their $J$-factors (see top two panels of Fig. 5.20).

5.5 Conclusions

In this chapter, we performed a systematic study of the spherical Jeans analysis method in the context of inferring the DM content in dwarf galaxies. We focused on simulated data for spherical, isotropic systems, and assessed the performance of
the method at accurately recovering the overall dark matter density profile, the virial mass, and the inner slope of the dark matter density profile. Our primary conclusions are the following:

- For parameter sets I and III, which describe $M_{200} \sim 10^9 M_\odot$ halos (intrinsic velocity dispersion $\sim 10$–$15$ km/s) with inner density slopes of $\gamma = 1$ and $\gamma = 0$, respectively, we find that the virial mass we recover is always consistent with the true value, and is increasingly accurate as the sample size is increased. However, the inner profile of the DM density distribution is less well-constrained—for samples with $\lesssim 1000$ stars, the posterior distributions on the inner slope $\gamma$ are typically unconstrained, and there is no statistical preference for a cuspy or cored profile. We recover statistical evidence for the true (cuspy/cored) profile only for samples with $10,000$ stars. For these parameter sets, the results are generally insensitive to varying the measurement error of the line-of-sight velocity, $\Delta v$, over the range $\Delta v = 0$–$5$ km/s.

- For parameter sets II and IV, which describe $M_{200} \sim 10^7 M_\odot$ halos with inner density slopes of $\gamma = 1$ and $\gamma = 0$, respectively, we find that the virial mass estimates depend sensitively on $\Delta v$, particularly for the samples with fewer stars. For parameter set II (intrinsic velocity dispersion $\sim 3$ km/s), the inferred virial mass for the 20- and 100-star samples is consistent with zero when $\Delta v = 5$ km/s. Similarly, for parameter set IV (intrinsic velocity dispersion $\sim 2$ km/s), the recovered virial mass for the 20- and 100-star samples is consistent with zero for the cases of $\Delta v = 2$, $5$ km/s. This is rectified when the sample size is increased to 1000 or more stars.

- From our study of spatial selection functions, we conclude based on the overall performance at inferring the DM density profile, the inner slope $\gamma$, and the virial
mass, that it is crucial to have measurements of stars across the full radial distribution of the dwarf galaxy, especially for the smallest samples. Doing so allows the fit to anchor the DM distribution across the full radial range, and consistently results in comparable or better performance at accurately reconstructing both the inner and outer profile of the DM distribution, relative to the cases where the data consists purely of stars in either the inner or outer region of the system. For systems with \( \lesssim 50 \) observed stars, measuring more stars across the full radial extent can reduce the uncertainties on \( \log_{10}[J(0.5^\circ)/(\text{GeV}^2 \text{cm}^{-5})] \) by a factor of \( \sim 3 \) compared to measuring the same number of stars only within the half-light radius.

• Degeneracy between the DM halo parameters in our model makes it difficult to constrain the inner slope, \( \gamma \), especially when sample sizes are small. We emphasize that this is separate from the issue of the velocity anisotropy profile being degenerate with the enclosed mass profile. While datasets with larger sample size can help resolve these parameter degeneracies, it is unfeasible to measure upwards of 10,000 stars—the sample size required for constraining \( \gamma \)—in the dwarf galaxies in the near future. Instead, a potential method for increasing the constraining power of Jeans analyses on the core-cusp problem is to jointly fit to many dwarf galaxies simultaneously. This is computationally challenging to implement using standard MCMC or nested sampling techniques, so a thorough study of joint fits requires the use of other methods for approximating posterior distributions.

• Cast in the context of indirect detection, we find that for the 20-star samples across all parameter sets (in the cases of sufficiently small \( \Delta v \) for the DM to be recovered), the median 1\( \sigma \) uncertainty on \( \log_{10}[J(0.5^\circ)/(\text{GeV}^2 \text{cm}^{-5})] \) across our 10 realizations is \( \sim \pm 1-2 \), in contrast with the uncertainties of \( \sim \pm 0.5 \)
quoted for some of the current ultrafaint dwarf measurements (with \( \sim 20 \) stars) in GS15 (see Table 5.5), which were used to derive the dwarf galaxy constraints on DM annihilation in A17. We find that this discrepancy may be driven by the more restrictive prior ranges for the DM profile parameters in GS15—in particular the prior range on the inner slope \( \gamma \)—and note that the resulting indirect detection results should be interpreted with this prior-dependence in mind.

In our study, we have focused on the case of spherical, isotropic systems with the goal of understanding the limitations of Jean analyses even in the absence of challenges that are known to complicate the process of Jeans dynamical mass modeling, such as background contamination [170, 340, 341, 342], the effect of assuming equilibrium for systems which are not in equilibrium [310], the effect of non-spherical models [317, 343, 344], the degeneracy between the enclosed mass and velocity anisotropy [300, 301, 302, 303, 304, 305], and the presence of potentially large fractions of binary stars in the dwarf galaxies [345, 346, 336, 347].

With regard to the core-cusp problem, we have found that even for the idealized systems we consider, and a relatively simple three-parameter halo model, the Jeans modeling method is severely limited in its ability to constrain the inner slope \( \gamma \) of the dark matter density profile. A crucial reason behind this is that there are degeneracies between the three parameters that describe our DM profiles. The fact that \( \gamma \) is difficult to pinpoint is consistent with previous Jeans modeling results in the literature [333, 348, 305]; we have additionally determined that, in order to constrain \( \gamma \) within this framework, it is necessary to measure \( \sim 10,000 \) stars within a single dwarf galaxy, which is not practical within the near future. We therefore need to search for alternative methods for addressing the core-cusp problem using Jeans analysis methods.
While complementary mass modeling methods have claimed preference for cores or cusps in the dwarf galaxies, important caveats when interpreting such results have been identified in the literature. For example, while many rotation curve analyses have shown preference for cored DM distributions, studies have shown that systematic effects in rotation curve analyses can erroneously bias the inferred DM distribution towards a centrally cored profile (see, e.g., Refs. [274, 275, 276] and references within). It has also been demonstrated in Refs. [277, 278] that results using mass estimator methods such as the ones proposed in Refs. [349, 271] can depend sensitively on the specific line of sight that is chosen, and can result, for example, in predicting a cored profile when the true halo is cuspy.

The parameter degeneracy that limits our ability to reconstruct \( \gamma \) is a distinct from the well-known mass-anisotropy degeneracy which plagues Jeans analyses, for which a number of proposed solutions exist in the literature: using higher order moments of the velocity distribution [300, 350, 351, 304, 305] and incorporating proper motion measurements of stars [352, 353] are among the methods that have been demonstrated to ameliorate the mass-anisotropy degeneracy. It is worth exploring whether or not these methods would also lead to improved constraints on the inner slope of the DM density profile, the answer to which is not intuitively obvious. Ref. [354] recently used the framework described in Ref. [304], which parameterizes the DM density profile as a multiply-broken power law and employs higher order moments, to derive \( J \)-factors for the classical dwarfs. They obtained \( J \)-factor estimates which are consistent with the ones from GS15, but with reduced uncertainties. Additionally, jointly fitting to multiple dwarf galaxies at once is a potential method for leveraging moderately-sized datasets to achieve better constraints on \( \gamma \). While we have not yet explored this avenue systematically, due to computational challenges, it is a promising direction for future work.
Finally, we have used our results to make recommendations for future observations. For the purpose of achieving more accurate, less prior-dependent $J$-factor estimates for the systems that dominate the indirect detection results presented in A17, we identify Ursa Major II, Ursa Minor, and Draco as the dwarf galaxies that would most benefit from more stars being measured. Our preliminary analyses show that if we assume the typical $J$-factor uncertainties that we find in our work, the DM annihilation limits for these systems may be weakened to the degree of significantly affecting their implications on the DM interpretation of the GCE.

5.6 Light Profile Fitting Procedure

We take a binned likelihood approach to fit the stellar light profile in the initial step of our analysis, modeling the light profile as a projected Plummer profile (Eq. 5.25). For a sample size of $n_{\text{stars}}$, we bin the data in $\sim \sqrt{n_{\text{stars}}}$ logarithmically-spaced bins in the projected radius $R$. Because the measurement errors on the stellar positions are small—largely driven by the uncertainties on the distance to the galaxy, given the accurate measurements on the angular positions of stars on the sky—we take the uncertainty on the number of stars in each bin to be the Poisson uncertainty corresponding to the mean number of stars in that bin. For a $100(1-\alpha)\%$ confidence level, the lower and upper bound of the Poisson uncertainty are given by \[ \mu_{\text{lo}} = \frac{1}{2} F_{\chi^2}^{-1} \left(\frac{\alpha}{2}; 2\hat{n} \right) \] \[ \mu_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} \left(1 - \frac{\alpha}{2}; 2(\hat{n} + 1) \right), \] where $F_{\chi^2}^{-1}$ is the inverse of the $\chi^2$ cumulative distribution function and $\hat{n}$ is the mean number of counts. We then have $\sigma_{\text{lo}} = \hat{n} - \mu_{\text{lo}}$ and $\sigma_{\text{up}} = \mu_{\text{up}} - \hat{n}$, which need to be modeled in our likelihood. In order to account for the asymmetric uncertainties
that arise from this prescription, we use the following approximation to a Gaussian log-likelihood for $\hat{n}_i$ observed counts and $n_i(\theta)$ predicted counts in the $i^{th}$ bin, where $\theta$ are the model parameters [356]:

$$\ln L(\hat{n}|\theta) = -\frac{1}{2} \sum_i \frac{(\hat{n}_i - n_i(\theta))^2}{V_i - V'_i(\hat{n}_i - n_i(\theta))}, \quad (5.29)$$

where $V = \sigma_{lo}\sigma_{up}$ and $V' = \sigma_{up} - \sigma_{lo}$.

We find that this approximation works well for our purposes, and we can generally fit the light profile extremely well. We note that in practice, it can be numerically easier to fit for the stellar surface density in each bin rather than the star counts themselves, but the principles remain unchanged. We use the results of the light profile fit to set the priors on the surface brightness parameters in our Jeans analysis—conservatively, we set the prior ranges of the surface brightness parameters to be the middle 95\% containment range on their posteriors from the light profile fit, similar to the procedure in Ref. [304]. We show an example light profile fit for a 20-star sample in Figure 5.21 and a 1000-star sample in Figure 5.22. The fit results are generally in excellent agreement with the data, and become increasingly well-constrained as the sample size is increased.
Figure 5.21: An example light profile fit for a single 20-star sample. In the left panel, the data points in the show the binned data, with error bars corresponding to the Poisson error for each bin; the blue line shows the median recovered profile, and the blue bands show the middle 68 and 95% containment. The right panel shows the corresponding triangle plot on the light profile parameters, with the true value of $a$ indicated by the red lines. To convert the units of $L$ from star counts to luminosity, we have assumed that each star has luminosity $L_\odot$.

Figure 5.22: Same as Figure 5.21, but for a 1000-star sample. Compared to the case of the 20-star sample, the light profile is significantly better constrained (note the different axes scales on the triangle plot compared to Fig. 5.21).
5.7 Prior Selection and Joint Analysis

5.7.1 Implementing Narrow Priors

In our fiducial analysis, we choose conservative priors on the DM halo parameters. Specifically, we impose a wide prior of \( \gamma \in [-1, 5] \) due to the large theoretical uncertainty on the inner slopes of DM halos. While values of \( \gamma < 0 \) are not physically-motivated, because they predict a density profile that dips down in the central region of the halo, we choose the lower bound of \(-1\) to allow \( \gamma \) the freedom to converge at 0—this would not be possible if the lower bound on \( \gamma \) were set exactly at 0. The values of \( \gamma \) on the highest end are also unphysical, because for \( \gamma \geq 3 \), the enclosed mass (Eq. 5.20) diverges at finite radius. Based on the posterior \( \gamma \) distributions from our fiducial scans (Fig. 5.4), we do not expect that assuming a prior range of \( \gamma \in [-1, 3] \) instead would qualitatively change our results, because the posterior probability for values of \( \gamma \geq 3 \) tend to be negligible.

Additionally, we have assumed a wide prior on the DM scale radius \( r_s \) of \( \ln(r_s/\text{kpc}) \in [-10, 10] \) for the purpose of being fully agnostic. However, we can follow the example of GS15 and set the more physically-motivated prior range of on \( r_s \) of 1 pc to 100 kpc, i.e., \( \ln(r_s/\text{kpc}) \in [\ln(10^{-3}), \ln(10^2)] \sim [-6.9, 4.6] \). For reference, a commonly used value for the NFW scale radius of the Milky Way DM halo is \( \sim 20 \text{kpc} \) [120, 130, 133, 135, 143, 176]—because we expect the dwarf galaxy DM halos to have smaller scale radii than the Milky Way halo, the GS15 priors are still fairly conservative.

We explicitly verify that implementing a narrower choice of priors on \( \gamma \) and on \( r_s \) does not qualitatively affect the results of our study, focusing on the 20-star samples because the smallest samples are most sensitive to prior choices. In Figure 5.23, we show the posterior \( \gamma \) distributions for the 20-star samples for parameter sets I (top...
**Figure 5.23:** Posterior $\gamma$ distributions for the 20-star samples for parameter sets I (top row) and III (bottom row). From left to right, the columns show the results for the fiducial priors, the narrow prior on $r_s$ and fiducial prior on $\gamma$, and the narrow priors on both $r_s$ and $\gamma$. We assume fiducial priors on all other parameters and the fiducial velocity error, $\Delta v = 2\,\text{km/s}$. The left column of Fig. 5.23 corresponds to the left column of Fig. 5.4 (for ease of comparison between the different sets of priors, the vertical scale here is zoomed in compared to Fig. 5.4). While there are slight quantitative changes, the key result—that the posterior $\gamma$ distributions are unconstrained, and therefore do not give rise to statistical evidence for a cusp or a core—remains unchanged. We show the analogous results for parameter sets II and IV in Figure 5.24.

Similarly, we can examine the recovered density and enclosed mass profiles that result from the narrow prior choices and compare them to our fiducial results. We
show this comparison for the 20-star samples from parameter set I in Figure 5.25. Qualitatively, we find that the recovered distributions are insensitive to the prior choices on $r_s$ and $\gamma$. Quantitatively, the recovered virial mass is $M_{200} \sim 2.3^{+10.2}_{-1.5} \times 10^9 M_\odot$ for the case of narrow prior on $r_s$ and fiducial prior on $\gamma$ and $M_{200} \sim 2.2^{+7.5}_{-1.4} \times 10^9 M_\odot$ for the case of narrow priors on both $r_s$ and $\gamma$. For the fiducial analysis, this value is $M_{200} \sim 2.1^{+6.7}_{-1.3} \times 10^9 M_\odot$. In each case, the recovered virial mass is consistent within uncertainty with the true value of $M_{200} \sim 1.9 \times 10^9 M_\odot$. Additionally, imposing narrow priors on $r_s$ and $\gamma$ does not result in smaller uncertainties on the inferred virial mass. We show the analogous results for parameter set III in Figure 5.26. The results for parameter sets II and IV are qualitatively similar.

For brevity, we only present selected representative results here. We have verified that, for our spatially selected samples (see Section 5.3.3 for detailed discussion), the choice of narrow priors on $r_s$ and $\gamma$ also results in qualitatively unchanged results
Figure 5.25: Inferred DM density profiles $\rho(r)$ (left panels) and corresponding enclosed mass profiles $M(r)$ (right panels) for 20-star samples parameter set I. From lightest to darkest, we show the results for the fiducial priors, the narrow prior on $r_s$ and fiducial prior on $\gamma$, and the narrow priors on both $r_s$ and $\gamma$. The recovered distributions are overall insensitive to the prior choices on $r_s$ and $\gamma$. The recovered virial mass is $M_{200} \sim 2.3^{+10.2}_{-1.5} \times 10^9 M_\odot$ for the case of narrow prior on $r_s$ and fiducial prior on $\gamma$ and $M_{200} \sim 2.2^{+7.5}_{-1.4} \times 10^9 M_\odot$ for the case of narrow priors on both $r_s$ and $\gamma$. For the fiducial analysis, this value is $M_{200} \sim 2.1^{+6.7}_{-1.3} \times 10^9 M_\odot$. In each case, the recovered virial mass is consistent within uncertainty with the true value of $M_{200} \sim 1.9 \times 10^9 M_\odot$.

from the fiducial ones presented in the chapter. We have found, however, that the narrow priors have a regulating effect in our preliminary study of jointly analyzing multiple dwarfs simultaneously, relative to our fiducial priors—we therefore employ the narrow priors in our discussion of the joint analysis in Sec. 5.7.2.

5.7.2 Joint Analysis

In lieu of obtaining much larger datasets (up to $\sim \mathcal{O}(10,000)$ stars) for the measured dwarf galaxies, one potential way to gain more constraining power on the DM halo parameters using moderately-sized datasets is to jointly analyze multiple dwarf
\[ \gamma = 0, r_s = 1 \text{kpc}, M_{200} \approx 1.4 \times 10^9 M_\odot, n_{\text{stars}} = 20 \]

Figure 5.26: Same as Figure 5.25, but for parameter set III. The recovered virial mass is \( M_{200} \sim 0.7^{+2.0}_{-0.4} \times 10^9 M_\odot \) for the case of narrow prior on \( r_s \) and fiducial prior on \( \gamma \) and \( M_{200} \sim 1.0^{+3.3}_{-0.6} \times 10^9 M_\odot \) for the case of narrow priors on both \( r_s \) and \( \gamma \). For the fiducial analysis, this value is \( M_{200} \sim 0.7^{+2.0}_{-0.4} \times 10^9 M_\odot \). In each case, the recovered virial mass is consistent within uncertainty with the true value of \( M_{200} \sim 1.4 \times 10^9 M_\odot \).

galaxies at once. While it may not be feasible in the near future to increase the stellar sample sizes within measured dwarf galaxies by orders of magnitude, with the advent of digital surveys, the number of discovered dwarf galaxies has exploded over the past five years (see, e.g., Figure 1 of Ref. [331]), and is expected to continue to grow drastically in the era of future surveys such as the Vera C. Rubin Observatory (formerly known as LSST, see, e.g., Table 1 of Ref. [357]). We could therefore try to leverage a large number of measured dwarf galaxies, even if within the individual systems the number of observed stars is small.

Within our analysis framework, we can in principle perform a joint analysis on \( N_{\text{dwarfs}} \) of our simulated dwarfs. For simplicity, we assume all \( N_{\text{dwarfs}} \) systems are from the same parameter set and have the same number of stars, and we analyze them simultaneously, under the prior assumption that they all share the same value
of $\gamma$ but are otherwise independent. This simulates the scenario of having a group of similarly-sized dwarf galaxies for which one might expect *a priori*, based on the specifics of the DM and baryonic feedback models, to have the same inner DM profile shape.

The joint likelihood is the product of Eq. 5.14 over each dwarf,

$$L_{\text{joint}} = \prod_{j=1}^{N_{\text{dwarfs}}} \prod_{i=1}^{N_{\text{stars}}} \frac{(2\pi)^{-1/2}}{\sqrt{\sigma^2_{p,j}(R_{ij}) + \Delta^2_{v_{ij}}}} \times \exp \left[ -\frac{1}{2} \frac{(v_{ij} - \bar{v}_j)^2}{\sigma^2_{p,j}(R_{ij}) + \Delta^2_{v_{ij}}} \right].$$

(5.30)

As in the case of the individual analyses, we model each dwarf with a Plummer light profile and gNFW DM distribution, but fit for only one value of $\gamma$ for all the dwarfs, i.e., $\gamma_j = \gamma$. The joint analysis model therefore has $(N_{\text{dwarfs}} \times 5 + 1)$ free parameters.

We have tested this method by taking five 20-star samples from the same parameter set and maximizing their joint likelihood. We note that for the results shown in this section, we have used the narrow priors on $r_s$ and $\gamma$ described in Sec. 5.7.1 and assume a velocity error of $\Delta v = 0$ km/s for cleanliness. All other priors are the same as in our fiducial setup. We choose to focus on the narrow priors because we have found that, for the cases we have tested, the joint analysis results can be biased more often towards incorrect values of $\gamma$ when using our fiducial priors.

In Figure 5.27, we show example results for parameter set I (for which $\gamma = 1$), with each row corresponding to a different set of five samples. In the first column, we show the results floating all 26 free parameters. Within each panel, we show the posterior $\gamma$ distributions resulting from the individual as well as the joint analyses—the teal line(band) shows the median of the median(middle 68%) in each $\gamma$ bin across the five individual scans, while the red line shows the posterior $\gamma$ distribution from the joint scan. While the posterior $\gamma$ distribution from the joint scan is more constrained and peaked near $\gamma = 1$, the posterior probability at $\gamma = 0$ tends to be non-negligible.
Figure 5.27: Example results from jointly analyzing five 20-star samples from parameter set I, for which $\gamma = 1$. Each row corresponds to a different set of five samples. From left to right, we show the results for floating all free parameters (26 free parameters), fixing $\rho_0$ for each sample to the true value (21 free parameters), fixing $r_s$ for each sample to the true value (21 free parameters), and fixing both $\rho_0$ and $r_s$ for each sample to their respective true values (16 free parameters). Within each panel, we show the posterior $\gamma$ distributions resulting from the individual as well as the joint analyses—the teal line(band) shows the median of the median(middle 68%) in each $\gamma$ bin across the five individual scans, while the red line shows the posterior $\gamma$ distribution from the joint scan.

The degeneracy between the DM halo parameters still has a strong effect on these particular results, as demonstrated by the fourth column, in which we fix $\rho_0$ and $r_s$ to their respective true values for each of the five samples (i.e., we now float a total of 16 parameters). In this case, the posterior distribution is narrowly peaked and the posterior probability at $\gamma = 0$ is negligible in all cases (although depending on the specific set of samples, the location of the peak may be shifted away from
the true value of \( \gamma = 1 \). If we fix either \( \rho_0 \) or \( r_s \) individually, we find that the joint analysis can accentuate biases that are present in the underlying samples (most clearly demonstrated by the middle two panels of the bottom row).

In Figure 5.28, we show analogous example results for parameter set III (for which \( \gamma = 0 \)). In this case, when all 26 free parameters are floated, the posterior \( \gamma \) distributions from the joint analysis tend to also be peaked near \( \gamma = 1 \). In the examples shown here, fixing \( \rho_0 \) for all the samples in the joint analysis resolves this bias, resulting in posterior \( \gamma \) distributions which are peaked near \( \gamma = 0 \) and better-constrained than the corresponding posteriors from the individual scans. When both \( \rho_0 \) and \( r_s \) are fixed to their respective true values for each of the five samples, the
posterior distributions from the joint scans are peaked cleanly about $\gamma = 0$ in all three cases; however, the bias towards $\gamma = 1$ is again present if we only fix $r_s$.

Further detailed study is required in order to understand the source of the biases we see, and also to characterize how the output of a joint analysis depends on factors such as the value of $N_{\text{dwarfs}}$, the sample size and measurement precision in each dwarf, and the relaxation of the assumption that the dwarfs all share the same value of $\gamma$ (for example, by assuming a central value of $\gamma$ and some scatter about it for the population of dwarfs being analyzed). However, the dimensionality of the model quickly grows as $N_{\text{dwarfs}}$ is increased, making a joint analysis difficult to efficiently implement using standard MCMC or nested sampling techniques. In particular, the number of MULTINEST evaluations required for convergence scales exponentially above $\sim 30$ dimensions [358], making it computationally infeasible to perform a detailed study using the analysis framework presented here. Nevertheless, our preliminary results suggest that a joint analysis approach is a promising method for making the most of the data moving forward, and deserves its own dedicated study. This would require the use of newer inference techniques which are designed to approximate posterior distributions for high-dimensional likelihoods, such as stochastic variational inference [359].
Chapter 6

Mining Stellar Substructure in the Milky Way

This chapter is based on ongoing work done in collaboration with Adriana Dropulic, Bryan Ostdiek, Tim Cohen, Mariangela Lisanti, Hongwan Liu, and Lina Necib. Section 6.1 introduces the motivation behind searching for stellar substructure in the Milky Way. Section 6.2 describes the different types of machine learning that are relevant throughout this chapter. Section 6.3 presents previous results from clustering analyses in 5D phase space. Section 6.4 presents work in progress on inferring the line-of-sight velocities of stars using machine learning. Section 6.5 summarizes the preliminary results, places them in the context of comparison with existing methods in the literature, and discusses the present outlook of the project.

6.1 Introduction

In the ΛCDM paradigm, hierarchical structure formation dictates that large galaxies such as the Milky Way formed from the mergers of smaller satellites [360]. In
this model, galaxies like the Milky Way gained most of their present baryonic and DM mass through the gravitational capture and absorption of satellite galaxies and globular clusters. During the infall process, the satellites are stripped apart by tidal forces, leaving behind stars and dark matter which gradually virialize within the host galaxy. If a particular merger was recent enough that it has not had time to fully virialize, it may leave behind distinctive signatures in phase space, such as clumps or streams of stars, as well as trails of tidal debris [361, 362, 363, 364, 365, 366]. By detecting such fossil remnants of merger events, we could potentially glean insight into the history of the host galaxy’s evolution. Additionally, because the DM in the host galaxy is accumulated through these same merging processes, the stellar remnants of mergers can also give us information on DM substructure as well as the DM velocity distribution in the host galaxy [367, 368].

The Milky Way stellar halo is currently undergoing at least one merger, demonstrated by the Sagittarius Stream [369, 370, 371, 372], a vast stream of stars left in the trail of the Sagittarius dwarf galaxy [373] as it has completed multiple orbits during its infall process. Numerous other stellar streams have been found in the Milky Way, some of which are shown in Figure 6.1. Ref. [374] provides a review of tidal streams up until 2016. Since then, the Dark Energy Survey (DES) Collaboration has discovered several new streams around the Milky Way [375]; the Python library galstreams¹, which was used to make Fig. 6.1, maintains an up-to-date list of Milky Way stellar streams [376]. Some of the known stellar streams originated from dwarf galaxies merging with the Milky Way, such as the Sagittarius stream; other streams, such as the Palomar 5 (Pal 5) stream [377], are associated with infalling globular clusters. It is straightforward to determine whether a stellar stream originated from a dwarf galaxy or a globular cluster when the original structure can still be detected,

¹https://github.com/cmateu/galstreams
which is the case for both Sagittarius and Pal 5; when this is not possible, properties of the stream such as the velocity dispersion can be used to determine the nature of the original structure that deposited the stars in the stream. While stellar streams are a stunning visual demonstration of substructure from mergers due to their spatial coherence, remnants from older mergers which have lost spatial coherence may still retain coherence in velocity space [378, 379, 380].

The advent of the Gaia satellite [381] has provided, along with photometric information, the positions and two proper motion components for an unprecedented number of stars—over one billion in the second data release (DR2). This has made it a particularly exciting time to study stellar substructure in the Milky Way, and several new structures have already been discovered using data from Gaia [382, 383, 384, 385, 386, 387, 388, 389, 390, 391]. In practice, however, searching for substruc-
ture close to the Galactic midplane is complicated by the large background of stars in the stellar disk, which were formed within the Milky Way and do not carry direct information about past mergers. To tackle this challenge, Ref. [392] used machine learning to categorize stars in both simulated and real \textit{Gaia} data as either stars which were accreted onto the host galaxy or ones which were formed \textit{in situ}.

The output catalog from Ref. [392] was used to find new kinematic substructure in Refs. [390, 391] through clustering analyses in 3D velocity space—specifically, only the small subset of stars in the \textit{Gaia} DR2 data which have measured radial velocities were analyzed in those studies.\textsuperscript{2} The catalog itself contains \(\sim 737,000\) stars, out of which \(\sim 37,000\) have radial velocity information. The goal of this project is to leverage the results of Ref. [392] and establish a method for robustly identifying spatial and kinematic substructure within the much larger dataset of accreted stars with only 5D phase-space information, i.e. the stars without radial velocity measurements. This is a challenging task, and we make use of simulated datasets, for which we have truth-level information, with the goal of ultimately applying the methods to the actual \textit{Gaia} data.

This chapter presents our two-fold approach. Initially, we studied clustering in 5D phase space, exploring a number of clustering algorithms but primarily focusing on the HDBSCAN* algorithm [393, 394]; selected results are presented in Section 6.3. We found that the 5D clusters identified by the HDBSCAN* algorithm were not reliably reproducible, and depended sensitively on the parameters of the clustering algorithm, making it difficult to physically interpret the results of our 5D clustering studies. Motivated by the need for additional information to stabilize the clustering, we began to explore the possibility of using machine learning to infer the line-of-sight velocities of stars from their 5D phase-space information. This is ongoing work, and

\textsuperscript{2}In this chapter, we use the terms “radial velocity” and “line-of-sight velocity” interchangeably.
some preliminary results are presented in Section 6.4. The current conclusions and outlook of this study are summarized in Section 6.5.

6.2 Machine Learning Primer

Machine learning can broadly be divided into supervised learning and unsupervised learning.\(^3\) In supervised learning, the machine learning algorithms are trained on labeled data. Once trained, the task is then to predict the labels for unlabeled data. The labeled data is typically divided into the training dataset and validation dataset—as the name suggests, the training dataset is used to train the model to be able to predict the correct labels; the validation dataset is not used directly in training the parameters of the model, but is used to evaluate the performance of the model during the training process. Together, the training set and validation set ensure that the model learns the information that is needed to predict the labels well, without falling into the trap of learning patterns which are specific to the training data and not generalizable to other datasets (“overfitting”). Finally, when the training is complete, another separate dataset, the test dataset, is used to evaluate the performance of the trained model. The test set can be specifically curated to be representative of the data that the model would ultimately be applied to, to establish confidence in the model.

Supervised learning can further be divided into two broad classes of problems—classification and regression. In classification, the labels on the data are discrete. For example, in Ref. [392], classification was performed to categorize the stars in the Gaia DR2 data as having been accreted or formed \textit{in situ}. Conversely, in regression, the labels on the data are continuous. For example, in the study presented in Section 6.4,

\(^3\)For an introductory review on machine learning for physicists, see Ref. [395].
the goal of the machine learning is to predict the line-of-sight velocity, a continuous parameter, for each star in the data.

In unsupervised learning, the algorithm seeks to learn patterns and detect structure in the data, without being provided labels \textit{a priori} to train on. Clustering is an example of unsupervised learning, the goal of which is to group together data points based on their similarities. For example, in the study presented in Section 6.3, we used the HDBSCAN* clustering algorithm to identify clusters of stars which were close to each other in 5D phase space.

### 6.3 Clustering in 5D

Our first approach to the broad goal of finding an optimal method for detecting substructure was to directly perform clustering analyses on the accreted stars in the \textit{Gaia} data identified by Ref. [392], specifically focusing on clustering in 5D phase space: three position coordinates and two proper motion components. While 5D clustering on \textit{Gaia} data has been explored in the literature, the existing works either focus on identifying members of previously detected structures [396, 397, 398, 399, 400, 401] or rely on additional supervised learning to assess whether the overdensities identified by the clustering algorithm are physical or statistical [402, 403, 404]. Our approach to 5D clustering differs from those existing in the literature because our starting point was not the full \textit{Gaia} DR2 dataset, but rather the subset of stars identified by Ref. [392] as likely to be accreted. The neural network used in Ref. [392] to classify stars as accreted or \textit{in situ} assigned a network score, \( S \in [0, 1] \), to each star, where \( S = 1 \) corresponds to classifying a star as having been accreted and \( S = 0 \) corresponds to classifying a star as having been born \textit{in situ}. In practice, the network-predicted score values were typically somewhere between 0 and 1; for this
study, we chose to use as our data the stars with score $S > 0.75$, i.e. greater than 75% probability of being accreted.

We used the package HDBSCAN [405], which provides a Python implementation of the HDBSCAN* algorithm. A key difference between this algorithm (as well as the DBSCAN algorithm, from which many properties of HDBSCAN* are inherited) and other clustering algorithms such as K-means clustering or Gaussian mixture modeling is that one does not need to specify the number of clusters to be fit for.\textsuperscript{4} This is advantageous when performing clustering analyses on data in which we do not know the true number of clusters. There are two main parameters that go into an HDBSCAN clustering scan: \texttt{min_cluster_size} and \texttt{min_samples}. \texttt{min_cluster_size} is intuitive: it sets the minimum size of what gets classified as a cluster; the remaining parameter, \texttt{min_samples}, is less intuitive—according to the documentation, for the same value of \texttt{min_cluster_size}, a larger value of \texttt{min_samples} would lead to more of the data points being classified as noise and therefore could be thought of as corresponding to more conservative clustering.

We applied HDBSCAN to our dataset, varying over the HDBSCAN parameters to try to develop intuition for how the results would depend on the values of $(\texttt{min_cluster_size}, \texttt{min_samples})$. In Fig. 6.2, we show some example results from running HDBSCAN on input data consisting of the three heliocentric Cartesian position coordinates $(x, y, z)$, the proper motion in the right ascension direction (pmra, also denoted as $\mu_\alpha$), and the proper motion in the declination direction (pmdec, also denoted as $\mu_\delta$) for each star. Each row corresponds to a different set of parameters $(\texttt{min_cluster_size}, \texttt{min_samples})$. In each row, the left panel shows the detected cluster(s) of stars in declination (dec) versus right ascension (ra),

\textsuperscript{4}A comparison of various clustering algorithms implemented in Python can be found in the HDBSCAN documentation at https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html.
Figure 6.2: HDBSCAN results on Gaia DR2 stars with network score cut $S > 0.75$ \cite{392}, clustering in the parameters $(x, y, z, \text{pmra}, \text{pmdec})$. 

(a) HDBSCAN with min_cluster_size = 50 and min_samples = 10.

(b) HDBSCAN with min_cluster_size = 50 and min_samples = 15.

(c) HDBSCAN with min_cluster_size = 60 and min_samples = 10.

(d) HDBSCAN with min_cluster_size = 60 and min_samples = 15.
with arrows indicating the direction and magnitude of the proper motions and each color corresponding to a cluster identified by HDBSCAN. The middle panel shows the detected cluster(s) of stars in (pmra, pmdec) space, and as respective proxies for the thin and thick stellar disk, the stars from Ref. [392] with network score $S < 0.05$ (black circles) and $0.3 \leq S < 0.5$ (grey circles). Finally, the right panel also shows the detected cluster(s) of stars in (pmra, pmdec) space, but this time for comparison, the stars from the Nyx (black triangles) and Nyx-2 (grey triangles) streams from Ref. [390]. In general, we found that the clustering results depended sensitively on the choice of the parameters ($\text{min}_\text{cluster}_\text{size}$, $\text{min}_\text{samples}$), as demonstrated by the different rows in Fig. 6.2. While there was good shot-to-shot reproducibility of the clustering results for a fixed set of parameters, there seemed to be a lack of reproducibility in the specific clusters that were identified by HDBSCAN across different sets of parameters. Additionally, we found that the results also depended on the coordinate system in which the data was represented: Figure 6.3 shows examples of results from clustering in (ra, dec, parallax) spatial coordinates, rather than the heliocentric ($x, y, z$) coordinates used in Fig. 6.2. While clustering in Cartesian position coordinates\textsuperscript{5} yielded clusters that were coherent in proper motion—and typically had rather large proper motions, clustering in (ra, dec, parallax) coordinates resulted in some clusters which had very small overall proper motion and were comoving with the disk (and were correspondingly not coherent in proper motion).

In order to understand how the different sets of parameters compared to each other, and to find an optimal set of parameters for our analysis, we generated 100 bootstrapped datasets from our sample and repeated the clustering analysis on each bootstrapped dataset to statistically evaluate the performance of different choices of parameters. In particular, we wanted to evaluate the following: for a given choice of

\textsuperscript{5}We show here the results for clustering in heliocentric Cartesian coordinates, but note that the results using galactocentric Cartesian coordinates were very similar.
(a) HDBSCAN with min_cluster_size = 25 and min_samples = 10.

(b) HDBSCAN with min_cluster_size = 50 and min_samples = 10.

Figure 6.3: HDBSCAN results on Gaia DR2 stars with score cut $S > 0.75$ [392], clustering in the parameters (ra, dec, parallax, pmra, pmdec).

(min_cluster_size, min_samples), how reliably does the clustering algorithm pick up the same clusters across the different bootstrapped datasets? In order to ascribe robust physical meaning to a given group of stars found in the data, it would be crucial that the group can be found in a reproducible fashion across bootstrapped datasets.\(^6\)

For each of our chosen sets of (min_cluster_size, min_samples), we repeated the HDBSCAN analysis on each bootstrapped dataset, then compared the clusters found in each dataset. We classified two recovered clusters as belonging to the same star group if, in each of the five dimensions, the distance between their centers was less than 0.5 times the smaller of their extents in that dimension. A code snippet implementing

\(^6\)In the following discussion, we will use the term “clusters” to refer to the structures identified by HDBSCAN within individual datasets, and the term “star groups” to refer to recurring groups of stars which are picked up as clusters across multiple datasets.
import numpy as np

def sameblob(d1,d2,thresh=0.5):
    cen1 = np.median(d1, axis=0)
    cen2 = np.median(d2, axis=0)

    width1 = np.quantile(d1, axis=0,q=1)-np.quantile(d1, axis=0,q=0)
    width2 = np.quantile(d2, axis=0,q=1)-np.quantile(d2, axis=0,q=0)

    dist = np.abs(cen1-cen2)
    width = np.minimum(width1,width2)

    return np.all(dist < thresh*width)

Listing 6.1: Code for classifying two clusters of stars recovered by the clustering algorithm as either being within the same star group or not.

Figure 6.4: The clusters from analyzing the 100 bootstrapped datasets with (\texttt{min\_cluster\_size},\texttt{min\_samples}) = (90,10) which were classified as belonging to the same star group, called “Star Group 0.” The individual clusters are overlaid on top of each other in different colors. Star Group 0 was picked up as one or more clusters in \(n_{\text{MC}} = 35\) of the 100 datasets, the highest number out of all the identified star groups for this choice of parameters. It accounted for \(n_{\text{clust}} = 42\) out of the 527 total clusters that were identified by \texttt{HDBSCAN} across the 100 datasets—the fact that this number is larger than \(n_{\text{MC}}\) implies that Star Group 0 was identified as more than one overlapping cluster in some of the datasets.

this cluster-matching process is shown in Lst. 6.1. An example of a star group found via this matching procedure is shown in Figure 6.4, with its constituent clusters from the different bootstrapped scans overlaid on top of each other in the various colors.

After matching the clusters found across different bootstrapped datasets, we can count the number of times the members of a given star group is identified out of the
Figure 6.5: The number of bootstrapped datasets ($n_{MC}$) in which a given cluster is returned. “Cluster #” is the label for each unique cluster, and the clusters are numbered in descending order of $n_{MC}$. Each line corresponds to a different choice of min_cluster_size; min_samples = 10 for all lines shown here.

100 scans. Figure 6.5 shows the number of bootstrapped datasets ($n_{MC}$) in which a given star group was detected. “Star Group #” on the $x$-axis denotes the label ascribed to each unique star group, in order of decreasing $n_{MC}$. Each line corresponds to a different value of min_cluster_size: 50 (blue), 60 (orange), 70 (green), 80 (red), 90 (purple), or 100 (brown); min_samples = 10 for all the lines. In an ideal case, $n_{MC}$ should be close to 100—however, we found that across the board, $n_{MC}$ at most reached $\sim 35$. While not shown here, we found that no other combination of parameters scanning across min_cluster_size = \{50, 60, 70, 80, 90, 100\} and min_samples = \{10, 15, 20\} performed any better than this. In retrospect, this is perhaps not entirely unexpected if we are trying to detect small substructures on top of a dominant background of stars, because the process of bootstrapping would capture the properties of the background population of stars while the stars in substructures are by definition outliers or “rare events” which would not be captured as effectively by the bootstrapping process.
Finally, in order to evaluate the performance of the method in the presence of a substructure that we know is in the data, we artificially injected a “signal” substructure consisting of 100 stars which were spatially and kinematically coherent into the bootstrapped data and repeated the analysis procedure. We chose to center the signal substructure at \((x, y, z, \mu_\alpha, \mu_\delta) \sim (0, 0, 0, 250, -250)\). Importantly, we did not bootstrap the signal itself—the same signal stars were injected into each of the bootstrapped background star datasets. In this case, the injected substructure was the most consistently recovered star group (i.e. had the largest \(n_{MC}\), and therefore was labeled as “Star Group 0”) for all of the setups with \(\text{min\_cluster\_size} \leq 100\). This
Figure 6.7: Left: The number of bootstrapped datasets ($n_{MC}$) in which the injected cluster is recovered, as a function of $\text{min_cluster_size}$. Each line corresponds to a different choice of $\text{min_samples}$. Right: The number of stars in the injected cluster which are recovered in Star Group 0 ($n_{\text{stars}}^\text{rec}$) in any of the bootstrapped samples, as a function of $\text{min_cluster_size}$. Each line corresponds to a different choice of $\text{min_samples}$.

is expected, because the same set of 100 stars was injected into each bootstrapped dataset. Figure 6.6 shows the two most consistently recovered star groups for the specific case of $\text{min_cluster_size} = 60$ and $\text{min_samples} = 15$. The top panel shows Star Group 0, the star group detected in the most bootstrapped datasets and the one corresponding to the true injected substructure, and the bottom panel shows Star Group 1, the star group detected in the second most bootstrapped datasets.

Although the true injected substructure was consistently recovered as Star Group 0 across the parameter choices we tested, the injected signal was recovered in at most $n_{MC} \sim 75$ of the 100 datasets (left panel of Fig. 6.7) even for the best-performing parameter choices. Moreover, the number of injected stars which were identified as being associated with this star group in any of the 100 datasets plateaus at $n_{\text{stars}}^\text{rec} \sim 85$, i.e. even for the best-performing sets of parameters, at least 15 of the injected stars were never identified as being part of the star group corresponding to the true signal substructure (right panel of Fig. 6.7). Figure 6.7 also shows that in most cases, the
performance of the clustering at recovering the injected signal (quantified by $n_{MC}$ and $n_{\text{stars}}^{\text{rec}}$) declines as $\text{min\_sample\_size}$ is increased, especially for $\text{min\_sample\_size} \gtrsim 100$. This is expected, because as the minimum size of what gets classified as a cluster is increased, it becomes increasingly difficult to reproducibly cluster together the same set of stars across different bootstrapped datasets.

These limitations on the reproducibility and stability of HDBSCAN clustering on our 5D data, even in the presence of a relatively large and clean injected stellar substructure, combined with the lack of reproducibility in the absence of the injected signal, make it very difficult to make physical interpretations of any clusters found in the data using this method. Therefore, we instead turned to using supervised learning methods to try to infer the radial velocities of stars for which we only have 5D phase-space information, with the goal of then incorporating the inferred radial velocities into the clustering procedure.

## 6.4 Machine Learning the Sixth Dimension

Having established that it is a challenging task to robustly identify stellar substructure using only 5D phase-space information, we set out to utilize supervised learning methods to predict the line-of-sight velocities of stars using their measured properties such as positions on the sky and proper motions. We first apply our methods on simulated data, in which we have truth-level information for the kinematic and photometric properties of each star.

### 6.4.1 Galaxia Mock Gaia Catalog

The simulated data that we use is the mock Gaia catalog from Ref. [406], which was generated from Galaxia, a code for generating synthetic surveys of the Milky
Figure 6.8: Properties of the stars in the training set in ICRS coordinates. These distributions are representative of the distributions of stars within the validation and test sets as well. The top row, from left to right, shows the right ascension (ra), declination (dec), and parallax ($\varpi$); the bottom row, from left to right, shows the proper motion in the right ascension direction (pmra), proper motion in the declination direction (pmdec), and the line-of-sight velocity ($v_{\text{los}}$).

Way [407]. The entire mock catalog has a total of $\sim 1.6$ billion stars, which is comparable with the size of Gaia DR2. For our purposes, we do not need to use the full catalog at once. Instead, we first placed a cut on the relative parallax ($\varpi$) error of $\delta\varpi/\varpi < 0.1$; $\sim 75$ million stars remain after placing this cut. Even though the parallax values quoted in the catalog are the true values, without any convolved errors, and therefore $1/\varpi$ can be taken as the exact value of the distance to a star, we perform this cut to mimic what would be done on the real data. We then generated three separate random, mutually exclusive 500,000-star subsamples from the parallax-selected dataset to use as our training, validation, and test datasets in the machine learning. Figure 6.8 shows the distributions of all six phase-space parameters that are included in the datasets: ra, dec, parallax, pmra, pmdec, and line-of-sight velocity.
Figure 6.9: Properties of the stars in the training set in galactocentric coordinates, with positions shown in Cartesian coordinates and velocities shown in spherical coordinates for convenience. These distributions are representative of the distributions of stars within the validation and test sets as well. The top row, from left to right, shows the galactocentric $x$, $y$, and $z$ coordinates; the bottom row, from left to right, shows the galactocentric velocity components in spherical coordinates, $v_r$, $v_\theta$, and $v_\phi$, where $\theta$ is the polar angle and $\phi$ is the azimuthal angle, oriented such that the galactic disk rotates in the $-\phi$ direction.

$v_{\text{los}}$. It is often convenient to instead use a galactocentric coordinate system, especially when we are trying to make physical interpretations of results. We implement the coordinate transformations detailed in the Appendix of Ref. [408], and show the resulting galactocentric phase-space distributions for the test set in Figure 6.9—for physical intuition, it can be helpful to use Cartesian coordinates for the positions and spherical coordinates for the velocities, as shown here. In this Cartesian coordinate system, the “sun” is located at $(-8, 0, 0.015)$ kpc. We use the convention where $\theta$ is the polar angle and $\phi$ is the azimuthal angle, and orient the spherical coordinates such that the galactic disk rotates in the $-\phi$ direction. As shown in the bottom right
panel of Figure 6.9, the mock catalog is dominated by the disk, which rotates at a
median velocity of $\sim 216$ km/s.

6.4.2 The Case of the Helmi Streams

As a point of comparison, we consider the work in the literature that has sought
to approximate the line-of-sight velocities of stars in DR2 which lack radial velocity
measurements, and subsequently used the estimated stellar kinematics to search for
substructure in the Milky Way stellar halo. The most notable example is that of
the Helmi streams, first discovered over 20 years ago in the space of action-angle
variables [378]. The Helmi stream members are prominent in the $L_y-L_z$ plane, so an
approximation to the line-of-sight velocity that is effective for the goal of identifying
more member stars must be able to perform well at predicting $L_y$ and $L_z$. In the
directions directly towards and away from the Galactic Center, the line-of-sight ve-
locity, $v_{\text{los}}$, is purely along the cylindrical radial direction. Therefore, for stars lying
along those directions, $L_z = rv_\phi$ and $L_y = -xv_z$ are independent of $v_{\text{los}}$. As the
angular distance away from these directions is increased, the assumption that $L_z$ and
$L_y$ are independent of $v_{\text{los}}$ decreases in accuracy. In Ref. [386], it is assumed that
$v_{\text{los}} = 0$ for stars that lie within $15^\circ$ of the lines of sight towards and away from the
Galactic Center—we will refer to this spatial selection as the $15^\circ$ “cone” region, de-
fined by $\sqrt{\ell^2 + b^2} \leq 15^\circ$ or $\sqrt{(\ell - 180^\circ)^2 + b^2} \leq 15^\circ$. Ref. [386] found that within this
region, the maximum difference between the true angular momenta and the angular
momenta predicted by setting $v_{\text{los}} = 0$ is $\sim 1000$ kpc km/s. In that work, tentative
new stream members were selected by defining the “boxes” in action-angle space

$$
A : 1750 < L_{\perp} < 2600 \text{ kpc km/s}, \; 1000 < L_z < 1500 \text{ kpc km/s}
$$
$$
B : 1600 < L_{\perp} < 3200 \text{ kpc km/s}, \; 750 < L_z < 1700 \text{ kpc km/s}, \quad (6.1)
$$
where \( L_\perp = \sqrt{L_x^2 + L_y^2} \). Because the extent of the bigger box B is \( \sim 1000 \) kpc km/s, the authors took the 15° cone region to be the maximal region for their search.

The primary limitation of this method is that it is only valid within a narrow spatial region (only \( \sim 5\% \) of the stars in our datasets lie within the 15° cone). We are interested in building a method which is valid over the full spatial region covered by the data; in Section 6.5, we compare the performance of our method to the performance of setting \( v_{\text{los}} = 0 \).

### 6.4.3 Building a Neural Network

We use the Keras package [409] to implement our regression problem of predicting \( v_{\text{los}} \). We begin by building a basic model made up of densely connected hidden layers and an output layer; the model takes as input the 5D phase-space variables (unless otherwise specified, we take these to be the galactocentric Cartesian positions and the two proper motions) and returns a single, continuous variable, in this case \( v_{\text{los}} \).

Due to the black box nature of deep neural networks, our approach to the following

```python
from keras.models import Sequential
from keras.layers import Dense, Dropout
from keras.optimizers import Adam

MyModel = Sequential()
MyModel.add(Dense(units=100, activation="tanh", input_dim=X_train.shape[1]))
MyModel.add(Dropout(0.1))
MyModel.add(Dense(units=100, activation="tanh"))
MyModel.add(Dropout(0.1))
MyModel.add(Dense(units=100, activation="tanh"))
MyModel.add(Dropout(0.1))
MyModel.add(Dense(units=100, activation="tanh"))
MyModel.add(Dropout(0.1))
MyModel.add(Dense(units=1))

MyModel.compile(loss="mean_squared_error", optimizer="adam")
```

Listing 6.2: Code for setting up our basic model.
tests will be to choose an arbitrary initial configuration which is reasonable within the realm of standard neural network architectures employed in the literature. We can then use this initial choice as a starting point from which to adjust aspects of the configuration, and with which to compare the performance between the different architectures at predicting \( v_{\text{los}} \). As our starting point, we build a model that has four hidden layers with 100 nodes each. We implement dropout at a rate of 0.1 between the hidden layers to avoid overfitting—namely, 10% of the nodes in each hidden layer are ignored during the training process. This initial setup is shown as a code snippet in Lst. 6.2.

We assume a mean squared error loss model and focus on the tanh activation function here—we have also tested relu and elu activation functions, and find that generically, the tanh activation function performs best at returning a predicted \( v_{\text{los}} \) distribution that is not significantly biased. In Figure 6.10, we show the results after training this model for 5000 epochs, using a batch size of 10,000 stars in the training process—the number of epochs is initially chosen to be large to ensure the training has converged. The first panel shows the training (blue) and validation (orange) losses, namely the mean squared error on the training set and validation set,
Figure 6.11: Output of our basic neural network for the test dataset. The top left panel shows a 2D histogram of the predicted \( v_{\text{los}} \) versus true \( v_{\text{los}} \) for all the stars in the test set. The dashed black line along the diagonal indicates one-to-one correspondence. The fact that there is a clear correlation between the predicted and true values of \( v_{\text{los}} \) indicates that the neural network is learning more than just the overall distribution of \( v_{\text{los}} \) for the full sample. We also transform the predicted results into galactocentric spherical coordinates, and show the predicted versus truth histograms for \( v_r \), \( v_{\theta} \), and \( v_{\phi} \) in the second, third, and fourth panel of the top row, respectively. The bottom row shows the 1D histograms of the true (shaded) and predicted (shaded) velocities corresponding to the top row.

respectively. The second, third, and fourth panels show the true \( v_{\text{los}} \) in the shaded histograms and the inferred \( v_{\text{los}} \) in the outlined histograms for the training, validation, and test set, respectively. Reassuringly, the predicted distributions for the training, test, and validation sets look very similar. However, the neural network overpredicts the frequency of small line-of-sight velocities and underpredicts the tails of the \( v_{\text{los}} \) distribution. While the overall predicted distribution of \( v_{\text{los}} \) is one measure of how well the method is performing, it is also important to assess the correlation between the true and predicted values of \( v_{\text{los}} \), which we show for the test set in Figure 6.11. In the top left panel, we show a 2D histogram of the predicted \( v_{\text{los}} \) versus true \( v_{\text{los}} \) for all the stars in the test set. The dashed black line along the diagonal indicates one-to-one correspondence. The three remaining panels in the top row show the results after
Figure 6.12: Same as Figure 6.11, but instead of using the neural network to predict $v_{\text{los}}$, we naively generate values of $v_{\text{los}}$ by inverse transform sampling the overall probability distribution of $v_{\text{los}}$ for the full sample. In this case, there is very good agreement between the 1D distributions of the predicted $v_{\text{los}}$ and true $v_{\text{los}}$, but there is no correlation between the predicted and true values.

It is promising that there is correlation between the predicted and true values of $v_{\text{los}}$—in a naive scenario, one could imagine predicting $v_{\text{los}}$ for each star by randomly sampling a value from the overall probability distribution of $v_{\text{los}}$ for all the stars; in that case, the 1D distribution of the predicted values of $v_{\text{los}}$ would match the true distribution very well, but there would be no correlation between the predicted and true values. This can be seen in Figure 6.12, where we have used inverse transform sampling to generate a predicted value of $v_{\text{los}}$ for each star in the test set and made an analogous figure to Figure 6.11. In this case, the 1D distribution of the predicted line-of-sight velocities is in much better agreement to the true distribution, compared
to the output from our model, but all correlation is lost between the predicted and true values of $v_{\text{los}}$. We therefore know that the machine learning is doing more than simply learning the overall distribution of our variable of interest.

6.4.4 Optimizing the Neural Network

Having established our basic framework, we then want to vary both our methods for preprocessing the data and the specifics of our network setup, in order to find an optimal configuration which most accurately and consistently predicts the line-of-sight velocities.

Weighting Functions

Because the network overpredicts the frequency of small values of $v_{\text{los}}$, we next try to de-emphasize the values of $v_{\text{los}}$ near 0 in the training procedure by introducing weights such that the peak of the $v_{\text{los}}$ distribution is flattened. One way to implement this is to take the probability distribution of $v_{\text{los}}$ as the reciprocal of the weighting function, i.e. we histogram $v_{\text{los}}$ for all the stars in the sample, then interpolate to get the probability $p$ corresponding to a certain star’s value of $v_{\text{los}}$, and take $w = 1/p$ to
be the weight for that star. We will call this scheme the “linear” weighting scheme to distinguish from the “log” weighting scheme discussed below. When we incorporate linear weights in this fashion into the training process, we find that the peak of the $v_{\text{los}}$ distribution is overly de-emphasized—so much so that the predicted distribution of $v_{\text{los}}$ has a dip in the central region. This is shown in Figure 6.13. We therefore want to choose a weighting function that is less steep than $1/p$—instead, we can try to weigh each star by $\log(1/p)$. While this is simple in principle, in practice $\log(1/p)$ is negative for most stars so we need to choose a scheme for shifting the values to be positive. We choose the scheme $w = \log(1/p) - \min(\log(1/p)) + 1$, such that $w \geq 1$ for all stars. We denote this scheme as the “log” weighting scheme. When we implement log weights in the training procedure, the resulting peak height of the predicted $v_{\text{los}}$ distribution is closer to that of the true distribution than in the case of no weights. However, the predicted distribution appears to be bimodal—there is one peak in the distribution at negative $v_{\text{los}}$ which is too high relative to truth, and another peak at positive $v_{\text{los}}$ which is too low relative to truth. This can be seen in the leftmost panels of Figure 6.14. This bimodal feature was also present in the case of no weights, but appears to be more prominent when log weights are incorporated.

Figure 6.14: Same as Figure 6.11, but with log weights implemented.
Network Configuration & Training Length

In our process of exploring different configurations, we have tested various network depths; in particular, we have found that going from three to four hidden layers has negligible effect on the output. Therefore, we next try to increase the width of the network, i.e. increase the number of nodes in each hidden layer. While the double peak feature in the predicted $v_{\text{los}}$ is present when we use hidden layers with 200 or 500 nodes, we find that when we increase the size of each hidden layer to 1000 nodes, the double peak feature is significantly ameliorated, the predicted $v_{\text{los}}$ distribution is close to the true distribution, and the predicted and true $v_{\text{los}}$ are well-correlated (results shown in Figure 6.15). However, when we examine the losses throughout the training process, it becomes clear that the network is overfitting to the training data. This is demonstrated in the left panel of Figure 6.16. The training loss being smaller than the validation loss is a possible indicator of overfitting, but in this case it is only

---

7While we have tested the effects of varying the activation function and the weighting scheme in conjunction with changing the network configuration, across the setups that we have tested, the combination of the tanh activation function with log weights has consistently had the best performance, so we only present those results for the remainder of this chapter.
Figure 6.16: Output of the wider neural network (which has 1000 nodes in each hidden layer) for the test dataset. The training loss being smaller than the validation loss, combined with the dip in the validation loss—which reaches a minimum value and subsequently increases before plateauing to a larger loss—indicates that the model is overfitting to the training data.

smaller by an $\mathcal{O}(1)$ factor, so it is difficult to determine based on this whether or not the model is overfitting the data. The more telling feature is the behavior of the validation loss: as training progresses, the validation loss decreases to its minimum very quickly, then increases again and subsequently plateaus to a larger loss value while the training loss is continuing to decrease. This is a clear sign that the network is hyper-optimizing to the patterns in the training data which don’t generalize to the validation data, i.e. overfitting.

We can therefore implement early stopping, i.e. stop the training process before the specified number of epochs has been reached, to prevent the network from training further once the validation loss has reached its minimum value. When we do so, the predicted $v_{\text{los}}$ distribution becomes bimodal again (Figure 6.17).

**Spatial Structure in the Data**

Given how ubiquitous the double-peaked feature is across many variations in the network configuration, it does not seem to be specific to any particularly suboptimal network setup that we have tested—other variations which have been tested have included different dropout rates, L2 regularization instead of dropout, and training
Figure 6.17: Same as Figure 6.15, but with early stopping implemented such that the training stops once the validation loss has reached its minimum value.

Figure 6.18: Distributions of the stars in the training set. There is structure in the $x$–$v_{\text{los}}$ plane and $y$–$v_{\text{los}}$ plane which could be contributing to the double-peaked features in the predicted $v_{\text{los}}$ distributions.

first with linear weights for some number of epochs followed by log weights or vice versa. Upon reexamining the data, which seems to be symmetric about $v_{\text{los}} = 0$ by eye, we found that there was a very small asymmetry in the data itself: there are $\sim 5\%$ more stars with $v_{\text{los}} < 0$ than with $v_{\text{los}} > 0$ in the mock catalog, and therefore in each of our subsampled datasets. Moreover, due to the rotation of the disk, there are nontrivial structures in $v_{\text{los}}$ as a function of the galactocentric positions $x$ and $y$, shown in Figure 6.18. In particular, the line-of-sight velocities for stars with $y > 0$ are slightly skewed negative, and the ones for stars with $y < 0$ are slightly skewed
Figure 6.19: Output of the neural network with 100 nodes per hidden layer for the test dataset, but only using the stars with \( x < -8 \) and \( y < 0 \). The true \( v_{\text{los}} \) distributions are skewed positive, but the prediction from the model overemphasizes the skew.

Positive. To determine whether the model is producing a double-peaked prediction because it is trying to fit to this feature in the data, we repeated the analysis with spatial subsets of the data. For example, we take only the stars on one side of the sun, with \( x < -8 \) and \( y < 0 \), and feed them into our neural network. What we find is that while the prediction from the network is not double-peaked in this case, it still overemphasizes the skew that is present in the data itself (Fig. 6.19).

If we examine the dependence of \( v_{\text{los}} \) on the galactic latitude \( b \) and longitude \( \ell \), we can also identify an overdensity of stars with \( v_{\text{los}} \) skewed slightly positive near \( \ell \sim 300^\circ \) and another overdensity of stars with \( v_{\text{los}} \) skewed slightly negative near \( \ell \sim 100^\circ \) (shown in the left panel of Fig. 6.20). Additionally, there is a very specific dependence of \( v_{\text{los}} \) on \( \ell \). We can try to de-emphasize these overdensities by making use of weights like before, this time implementing weights that depend on both \( v_{\text{los}} \) and \( \ell \), i.e. we now build the joint distribution of \( v_{\text{los}} \) and \( \ell \), interpolate to get the probability \( p \) corresponding to a star’s value of \( v_{\text{los}} \) and \( \ell \), and take either the linear weight \( w = 1/p \) or log weight \( w = \log(1/p) - \min(\log(1/p)) + 1 \). In Figure 6.20, we show the unweighted joint distribution of \( v_{\text{los}} \) and \( \ell \) in the left panel, the linearly weighted distribution in the middle panel, and the log-weighted distribution in the right panel. The linear weights entirely wash out any structure in the \( v_{\text{los}}-\ell \) plane,
Figure 6.20: The joint distribution of $v_{\text{los}}$ and $\ell$ in the training set, with no weights (left), linear joint $v_{\text{los}}$–$\ell$ weights (middle), and log joint $v_{\text{los}}$–$\ell$ weights (right).

Figure 6.21: Same as Figure 6.17, but using log weights calculated from the joint distribution of $v_{\text{los}}$ and $\ell$, i.e. the weighting scheme demonstrated in the right panel of Fig. 6.20. In contrast with Fig. 6.17, which used log weights from the distribution of just $v_{\text{los}}$, the predicted distribution in this case is no longer double-peaked.

while the log weights somewhat smear out the overdensities while preserving a lot of the correlation between $v_{\text{los}}$ and $\ell$. When we incorporate the log joint weights into the model, we find that the predicted $v_{\text{los}}$ distribution no longer has a clear double peak feature in it (Figure 6.21). Still, the predicted distribution is peaked too high near $v_{\text{los}} \sim 0$ (and is skewed slightly negative) compared to the true distribution. However, thus far we have not introduced any notion of confidence (or uncertainty) into the network’s prediction.
6.4.5 Predicting the Network Confidence

We are interested in not just what the neural network predicts a given star’s $v_{\text{los}}$ to be, but also how confident the network is in its prediction, and thereby an uncertainty associated with the network-predicted $v_{\text{los}}$. To do so, rather than using a mean squared error loss function, we implement a custom loss function which is simply the negative of a Gaussian log-likelihood (negative because the network tries to minimize the loss function):

$$\text{Loss}^i = - \left( \log(2\pi\sigma_i^2) - \frac{(v_{\text{los},i} - \mu_i)^2}{2\sigma_i^2} \right)$$

(6.2)

for each star $i$, where we have omitted the overall factor of $1/2$ because we are not concerned with the actual value of the loss itself. The total loss function is the sum of Eq. 6.2 over all stars. We then train the network to predict a central value $\mu_i$ along with an uncertainty $\sigma_i$ for the line-of-sight velocity of each star. In practice, this is done in multiple stages because we find that it is difficult for the network to predict both parameters at once. We first hold $\sigma$ fixed to 0 and train the network to predict $\mu$.\(^8\) Once this training is complete, we then fix $\mu$ to the output from the first stage of training, and train the network to output $\sigma$ instead. This process is repeated twice through, and as a final step $\mu$ and $\sigma$ are trained simultaneously.\(^9\) The process helps the neural network to iteratively approach the true minimum of the loss. We refer to this neural network, with uncertainty built into the loss function, as our “confidence network”.

We apply the confidence network to a specific example, which is slightly different from the examples discussed earlier in this chapter. Rather than taking only the galactocentric coordinates $(x, y, z)$ as the input to the model, this example takes

\(^8\)Boldfaced symbols here denote the vector quantities over all stars.

\(^9\)The choice to iterate twice through the process of training $\mu$ and $\sigma$ separately is arbitrary. We have found that the results are significantly improved relative to iterating only once through, but have not yet tested the effect of iterating more times.
Figure 6.22: The output of our confidence network when trained on input data in the form of $(\ell, b, \varpi, x, y, z, \mu_\alpha, \mu_\delta)$ for each star, using the log joint $v_{\text{los}}-\ell$ weights. These panels show the results derived from only taking the network prediction of the central values $\mu_i$ for each star $i$. The overall distribution of $\mu$ does not agree well with the true distribution of $v_{\text{los}}$ (left panels). However, when we account for the network output uncertainties $\sigma_i$, the resulting predicted $v_{\text{los}}$ distribution is in much better agreement with the true distribution (shown in Fig. 6.23). Data courtesy of Adriana Dropulic.

$(\ell, b, \varpi)$ as spatial inputs in addition to the galactocentric $(x, y, z)$. We choose to present this example here because we have found that providing redundant spatial information in the form of $(\ell, b, \varpi, x, y, z)$ improves the performance of the confidence network relative to only providing $(x, y, z)$ or $(\ell, b, \varpi)$. The configuration of the network itself consists of four dense hidden layers with 30 nodes each, with a dropout rate of 0.1 between hidden layers, and uses the tanh activation function and the log joint $v_{\text{los}}-\ell$ weights. In this case, the distribution of the predicted central values of the line-of-sight velocity, $\mu$, does not match well with the true $v_{\text{los}}$ distribution (shown in the left column of Figure 6.22).

However, when we account for the uncertainty $\sigma$ in the neural network’s output, the resulting predicted distributions are much more consistent with the true distributions. This is shown in Figure 6.23: to obtain the grey distributions shown in the lower row, for each star $i$ we generate an estimate of $v_{\text{los}}$ by sampling the Gaussian...
Figure 6.23: The output of our confidence network when trained on input data in the form of \((\ell, b, \varpi, x, y, z, \mu_\alpha, \mu_\delta)\) for each star, with the log joint \(v_{\text{los}} - \ell\) weights, accounting for the network predicted values of the uncertainty \(\sigma\). To make the grey predicted histogram in the lower left panel, we estimate the \(v_{\text{los}}\) for each star \(i\) by sampling it from the Gaussian centered at \(\mu_i\) with width \(\sigma_i\), and construct a histogram of the estimated \(v_{\text{los}}\) for all the stars. We repeat this process 100 times, and show the median and minimum/maximum in each bin with the solid line and shaded band, respectively. For the grey histograms in the three other panels of the bottom row, we transform the generated \(v_{\text{los}}\) samples into galactocentric spherical coordinates and again take the median and minimum/maximum in each bin from the 100 iterations. The top row of 2D histograms show the median in each bin across our generated samples. Data courtesy of Adriana Dropulic.

centered at \(\mu_i\) with width \(\sigma_i\), and histogram the estimated \(v_{\text{los}}\) for all the stars. We repeat this procedure 100 times, and obtain the median and minimum/maximum value in each bin across the 100 iterations, which we show by the grey solid line and shaded band, respectively. A separate method for building a distribution of \(v_{\text{los}}\) that accounts for the predicted uncertainties is to build a kernel density estimate (KDE) using the predicted values of \(\mu_i\) and \(\sigma_i\) for each star. The median distribution we obtain from the sampling method matches the KDE result. The top row of Figure 6.23 shows the median 2D histograms we get from our 100 generated samples. After taking into account the network output uncertainty, the predicted \(v_{\text{los}}\) distribution we end up with is slightly skewed positive relative to the true distribution, and the predicted \(v_{\phi}\)
distribution does not fully capture the tails of the true distribution. However, if we additionally place a cut on $\sigma$ and only take the output results for those stars whose predicted value of $\sigma_i$ is less than 30 km/s—approximately the median of the predicted $\sigma_i$ values in this particular example—the resulting predicted $v_{\text{los}}$ and $v_\phi$ distributions are in better agreement with the truth (shown in Figure 6.24). Decreasing the cut on $\sigma_i$ to 25 km/s results in further improvement in the agreement between the predicted and true distributions (shown in Figure 6.25).

A remaining question to be addressed is why the predicted and true $v_{\text{los}}$ are not better correlated in this example as well as other setups we have tested using the confidence estimating network. This could be a generic feature that is present when we model the uncertainty in the regression, or it could be an artifact of the representation of the data that is input into the neural network. It could also be driven by the fact that we are training on all of the stars in our dataset, when the line-of-sight velocity of a certain star may be much better correlated with other stars that are spatially close by than with stars that are much farther away—to account for this possibility,
Figure 6.25: Same as Figure 6.23, but only showing the results for stars with predicted uncertainty \( \sigma_i < 25 \text{ km/s} \). There is improved agreement between the true and predicted \( \mathbf{v}_{\text{los}} \) and \( \mathbf{v}_{\phi} \) distributions compared to the case of placing the confidence cut \( \sigma_i < 30 \text{ km/s} \). The aliasing features in this figure are due to limited statistics resulting from the confidence cut. Data courtesy of Adriana Dropulic.

we have also established a method in which the model only trains on the \( N \) nearest neighbors of a given star. Preliminarily, it is unclear whether the inclusion of nearest-neighbor information into our neural network has a significant effect on the output predictions.

### 6.5 Conclusions and Outlook

In the era of the *Gaia* mission, we have access to kinematic data for an unprecedented number of stars in and around the Milky Way—over one billion stars in DR2 with measured positions on the sky, parallaxes, and proper motions. This wealth of data makes it possible to systematically search for stellar substructure in the Milky Way’s stellar halo, which can shed light on the formation history of our Galaxy. However, only a small subset of these stars additionally have measured radial velocities, and therefore full 6D phase-space information. For the remaining majority of the measured stars, we have access to three spatial coordinates and two proper motion components.
In the upcoming years, the *Gaia* mission will continue to measure the radial velocities of more stars—with the goal of all stars in the data having measured radial velocities by end-of-mission—but for the upcoming third data release (DR3), which has been delayed due to the COVID-19 global pandemic but is currently on track for release in late 2020, the main improvements relative to DR2 are not focused on measuring more radial velocities.

While there have already been a plethora of new stellar substructures detected in the *Gaia* DR2 data, many of these detections have relied on the use of measured radial velocities [383, 384, 386, 387, 388, 390, 391]. In addition to the missing radial velocities, another challenge that arises in the search for stellar substructure is the presence of a large background of disk stars, especially near the Galactic midplane, which were born in the Milky Way and therefore do not hold information on our Galaxy’s merger history. This challenge was addressed by Ref. [392], which developed a machine learning-based method for classifying stars in *Gaia* data as either accreted or born *in situ*.

We are interested in developing methods which would allow us to systematically search for stellar substructure within the large set of *Gaia* data which does not have measured radial velocities. Doing so would allow us to potentially discover new stellar substructure as well as identify more stars which belong to already known structures such as *Gaia* Enceladus (also referred to as the *Gaia* Sausage) [410, 411] and the Nyx stream [390]. We first approached this task by directly applying clustering algorithms to the catalog of accreted stars from Ref. [392], using only 5D phase-space information. We found that it was very challenging to physically interpret the clusters identified by the algorithm, and also found that results were highly sensitive to user-defined parameters of the clustering method. We concluded that having radial velocity information is essential to the process of systematically and robustly iden-
tifying stellar substructure, and therefore began to explore whether or not we could infer the radial velocities of stars which do not have measured radial velocities.

Our preliminary results show that, using our neural network-based approach, we can predict the line-of-sight velocities of stars in our simulated dataset to better accuracy than more simple methods—for example, our predictions show stronger correlation between the predicted and true values of the line-of-sight velocity $v_{\text{los}}$ than the naive approach of predicting velocities by randomly sampling the probability distribution of $v_{\text{los}}$ for the full sample. One method that has been used in the literature to approximate the velocities for stars missing radial velocity measurements, which can serve as another point of comparison, is to assume the radial velocity is equal to 0 for those stars [386] (see Sec. 6.4.2 for more discussion).

We can compare the performance of our approach to the approach of setting $v_{\text{los}} = 0$, shown in Figure 6.26. In the first row, we show in grey the predictions for the galactocentric spherical velocity components $v_r$, $v_\theta$, and $v_\phi$, as well as the $z$-component $v_z$, from our confidence network discussed in Section 6.4.5. $v_\phi$ and $v_z$ are the velocity components that the method from Ref. [386] was optimized for predicting. For comparison, we also show the velocities we get if we set $v_{\text{los}} = 0$ for all the stars (shown in red). In the second row, we only show the stars that lie in the “cone” region defined by $\sqrt{\ell^2 + b^2} \leq 15^\circ$ or $\sqrt{(\ell - 180^\circ)^2 + b^2} \leq 15^\circ$; this cut is implemented in Ref. [386] when making the assumption $v_{\text{los}} = 0$. With the $15^\circ$ cut in place, both methods predict $v_\theta$ and $v_\phi$ distributions that are in excellent agreement with the true distributions. In the third row, we do not place any spatial cuts on the results, but only show results for the stars with uncertainty $\sigma_i < 30$ km/s predicted by our confidence network. Finally, the fourth row is the same as the third row, except for a cut of $\sigma_i < 25$ km/s instead on the predicted uncertainty. In all four rows, our confidence network performs better than the method of setting $v_{\text{los}} = 0$ at
Figure 6.26: Comparing the predicted results from our confidence model (grey histograms) to the results from setting $v_{\text{los}} = 0$ for all stars. In each panel, the shaded blue histogram shows the true distribution, the grey line and band show the median and minimum/maximum in each bin from our 100 sampled distributions, and the red line shows the result from setting $v_{\text{los}} = 0$. From left to right, the columns show the three galactocentric spherical velocity components $v_r$, $v_\theta$, and $v_\phi$, as well as $v_z$. The first row shows the results for all stars; the second row shows results only for stars which lie in the regions defined by $\sqrt{\ell^2 + b^2} \leq 15^\circ$ or $\sqrt{(\ell - 180^\circ)^2 + b^2} \leq 15^\circ$; the third row shows results only for stars whose network predicted uncertainty $\sigma$ is less than 30 km/s; and the fourth row shows results only for stars whose network predicted uncertainty $\sigma$ is less than 25 km/s.
predicting the $v_r$ distribution. Predicting $v_r$ accurately is important for finding more stars in *Gaia* Enceladus, for example, whose orbit is highly radial (see e.g. Figure 5 of Ref. [391]).

Thus far, we have tested our methods on mock *Gaia* data, training and testing on simulated stars for which we have the true $v_{\text{los}}$ values. Ultimately, on the real *Gaia* data, we would train our neural network on the subset of data with measured radial velocities, and subsequently apply the trained network to the remaining stars to obtain an estimate of their radial velocities along with the uncertainty on our predictions. There remains much work to be done in this ongoing project, but our preliminary results suggest that this novel approach of inferring the radial velocities of stars using machine learning may provide us with the missing information needed to mine the wealth of available stellar data for hidden structures, and thereby shed light on the formation history as well as the dark matter distribution of the Milky Way.
Bibliography


263


273


275


279


present and future γ-ray observatories - I. The classical dwarf spheroidal


[365] B. Robertson, J. S. Bullock, A. S. Font, K. V. Johnston, and L. Hernquist,
“A-cold dark matter, stellar feedback, and the Galactic halo abundance

Wiersma, and C. Dalla Vecchia, “Cosmological simulations of the formation of the
stellar haloes around disc galaxies,” *MNRAS* 416 no. 4, (Oct., 2011)


Tracers of the Local Dark Matter Velocity Distribution in the Milky Way,”


[370] SDSS Collaboration, Z. Ivezic *et al.*, “Candidate RR Lyrae stars found in

Richards, C. Stoughton, J. Anderson, John E., J. Annis, J. Brinkmann,
B. Chen, I. Csabai, M. Doi, M. Fukugita, G. S. Hennessy, Ž. Ivezić, G. R.
Knapp, R. Lupton, J. A. Munn, T. Nash, C. M. Rockosi, D. P. Schneider,
J. A. Smith, and D. G. York, “Identification of A-colored Stars and Structure
in the Halo of the Milky Way from Sloan Digital Sky Survey Commissioning
[astro-ph].

Streams: Evidence for a Nearly Spherical Massive Dark Halo around the
[astro-ph].


