8-24-2016 12:00 AM

Similarity, Adequacy, and Purpose: Understanding the Success of Scientific Models

Melissa Jacquart
The University of Western Ontario

Supervisor
Chris Smeenk
The University of Western Ontario

Graduate Program in Philosophy
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Recommended Citation

Jacquart, Melissa, "Similarity, Adequacy, and Purpose: Understanding the Success of Scientific Models" (2016). *Electronic Thesis and Dissertation Repository*. 4129.
https://ir.lib.uwo.ca/etd/4129

Abstract

A central component to scientific practice is the construction and use of scientific models. Scientists believe that the success of a model justifies making claims that go beyond the model itself. However, philosophical analysis of models suggests that drawing inferences about the world from successful models is more complex. In this dissertation I develop a framework that can help disentangle the related strands of evaluation of model success, model extendibility, and the ability to draw ampliative inferences about the world from models.

I present and critically assess two leading accounts of model assessment, arguing that neither is sufficient to provide a complete understanding of model evaluation. I introduce a more powerful framework incorporating elements of the two views, which can help answer these three questions: What is the target of evaluation in model assessment? How does that evaluation proceed? What licenses us in making inferences about the real world, based on the evaluation of our models as successful?

The framework identifies two distinct targets of model evaluation: representational similarity between the model and target system, and the adequacy of the model as a tool to answer questions. Both assessments must be relativized to a purpose, of which there are three general kinds: descriptive, predictive, and explanatory. These purposes differ in the way they inform the similarity relation, which is relevant for the similarity assessment, and the output they produce, which is relevant for the adequacy assessment. Any model can be assessed relative to any purpose, however a model encodes certain decisions made during the model's construction, which impact its ability to be applied to a new purpose or new domain. My framework shows that extending a model, and drawing inferences from it, depends on its representational similarity.

I apply this framework to several examples taken from astrophysics showing in detail how it can help illuminate the structure of the models, as well as make the justification for inferences made from them clear. The final chapter is a detailed analysis of a contemporary debate surrounding the use of models in astrophysics, between proponents of MOND and the standard ΛCDM model.

Keywords

Philosophy of Science, Philosophy of Astrophysics, Model Evaluation, Adequacy for Purpose, Similarity Relation, Domain of Application, Models, Modeling, MOND, ΛCDM.

Acknowledgments

It takes a village to raise a philosopher, and there are many people who have contributed to me becoming the philosopher I am. First and foremost I would like to thank my supervisor Dr. Chris Smeenk for his insightful feedback and generous support. There are also many faculty members from whom I have had the privilege of learning during my time at Western. I would like to thank Dr. Gillian Barker in particular for numerous conversations ranging from philosophical matters, to general discussion about professionalization and the writing process. Also Drs. Kathleen Okruhlik, Wayne Myrvold, Eric Desjardins, Robert DiSalle, Bill Harper, Carl Hoefer, Stathis Psillos, and Samantha Brennan for illuminating discussions and their contributions to the philosophy community at Western.

One of the keystones to my graduate education has been the Rotman Institute of Philosophy. I am incredibly grateful to Joseph Rotman for his vision, dedication, and generosity in establishing the Institute. The Rotman Institute has been an invaluable. Within the Institute, I have been surrounded by wonderful philosophers, colleagues, and friends. Everyone who is a member of the Institute contributes to the excellent environment, and philosophical community. In particular, I would like to thank Carol Suter for her support in keeping academics on track, focused, and supplied with coffee and cookies. My fellow graduate students (and officemates) have been are a wonderful source of discussion, support, and inspiration. In particular I want to thank Molly Kao, Emma Ryman, Craig Fox, Andrew Peterson, Yann Benétreau-Dupin, and Jessey Wright. But perhaps the most notable source of support, encouragement, and mentoring has been the Rotman Post-Docs: Alida Liberman, Rachael Brown, Kerry McKenzie, Alkistis Elliott-Graves, Robert Foley, and Dan Hicks. I am also grateful for coffee discussion with Nick Nash and Michael Walshots. It is a welcome challenge to discuss my work in the philosophy of science with philosophers who specialize in other areas.

I am also incredibly grateful to have had support and discussions with non-philosophers though my work at Western's Teaching Support Centre—Melanie-Ann Atkins, Leichelle Little, Kate Traill, and my other colleagues at the TSC. I am constantly inspired by them to better myself in all aspects of academia.

Thanks also to the Mary Routledge Fellowship and the Routledge family, for support in visiting the University of Pennsylvania, and Dr. Michael Weisberg for discussing my work with me during my time there. The Arts & Humanities faculty Graduate Thesis Award and The Carnegie Observatories for support in obtaining a philosopher's cross-training in astrophysics, and Dr. Barry Madore for sponsoring my trip the Las Campanas Observatory, and discussion with me about my case studies.

Finally, I want to thank my family for their unwavering support and encouragement through this process. My mom, for not only being an incredible role model, but for supporting me and always encouraging me to do my best. My sister, for being a little sister I look up to with awe and for inspiration. And my husband, partner in life, and fellow philosopher Lucas Dunlap for his constant support and encouragement. Through his insightful philosophical feedback, discussion, and support, he constantly helped push me to be the strongest philosopher I could be.

Table of Contents

List of Tables

List of Figures

Chapter 1

1 Introduction

1.1 Introduction

- What is the difference between a physicist and an astronomer?

- A physicist needs two data points to get a line of best-fit, but an astronomer only needs one.

This joke was told to me on many occasions throughout my time as undergraduate major in astronomy and physics by post docs and faculty members. The first time I heard it, the first day in my undergraduate astronomy lab, I was shocked. While I knew this was an exaggeration, there is undeniably an element of truth to it. I wondered how one of our best sciences could function if astrophysicists were able to develop models based on such sparse data. I asked how we know that the model we are working with is “good” and “successful” if we only have one data point. The answers I received were vague, unspecific, and never fully satisfying.

While this did not bother many of my peers, it troubled me greatly. In astrophysics, there is extremely limited access to observational data; we are limited to a small observable part of the universe, and will never be able to observe the universe in its entirety. Nonetheless, cosmologists make claims about the entire universe based on what they see locally[1]. If their models are only based on local observations, how can they possibly be used to make claims about the entire universe? Furthermore, astronomers will

[1] While cosmology and astrophysics are closely related, the difference between them can be characterized as follows: cosmology is the study of the entire universe, and aims at answering broader questions about the nature of the entire universe, and the laws that govern it. Cosmologists tend to makes global claims based on local extrapolations, given that it is impossible for them to observe the universe in its entirety. Astrophysics, on the other hand, is a branch of astronomy that studies various bodies in the universe, such as stars, black holes, galaxies, and the interstellar medium. Astrophysicists tend to makes claims regarding types of local systems. Claims about the systems of interest in astrophysics often involves samples of many instances of a particular type of system, which can then be used to test models to describe the system of interest.

never be able to run experiments on the universe as a whole, a common tool in other sciences for investigation of a system. How are any of our models to be tested, and how are they supported? This, to some extent, is the explanation of the joke: astronomy, because of the nature of what is being observed, often only has a small data set to work with, and so we must use what we have.

Yet this should strike one as a very perplexing situation. Scientists constantly use models to make ampliative inferences, to go beyond the data they have to learn more about the system than what they can see of it. The question in cosmology and astrophysics is whether we can ever be justified in doing this in spite of the fact that we have so few observational data. Obviously there are best practices that have developed in each science individually that guide the practitioners in developing their models. However, such reasoning is rarely analyzed, and the justifications for why practitioners make these decisions are rarely discussed in published work. What does make its way into the literature lacks detail about this process and the decisions made in developing the models. And the complex details about a model's accuracy are often presented as a simple numerical value of (un)certainty[2]. Questions like these continued to bother me throughout my time as an astrophysics major, and (as the story goes for many), I was not satisfied until I found my way into a philosophy of science classroom, where understanding of the process of scientific reasoning and justification was the main focus of inquiry.

In the philosophical literature on scientific modeling it is acknowledged that models are used as a tool to understand and investigate the world around us. Models, including physical scale models, mathematical equations, and computer simulations are indispensable for scientific practice. A central component to scientific practice is the

[2] There are various different quantitative measures of a "degree of fit of a model." For example, in some disciplines such as astronomy and physics, the standard deviation σ (sigma) measures dispersion of data around a mean. Practitioners often use "5 sigma" as the standard for discovery. The standard deviation measures the likelihood that data is the result of random fluctuation or error. "5 sigma" means it is extremely unlikely that data results from a random occurrence. In other disciplines, quantitative measures of a model's degree of fit take the form of a p-value, Akaike information criterion (AIC), or Bayesian information criterion (BIC) for model selection.

construction and use of scientific models, and it is through the use of models that scientists are able to make claims about what we know about the world around us, and how the world works. The challenge in understanding how we can generate knowledge from models stems from the fact that models are necessarily incomplete representations, and partial descriptions of the features of phenomena[3] in the world being modeled. Yet science proceeds on the assumption that we are effectively able to discover new things about the world through models. The interesting philosophical project is to develop an understanding of how we can discover true claims about the world, even though it is acknowledged that the models being used offer only an incomplete representation of the system under study.

There is a wide variety of models, and many different ways to think that they relate to the real world. Scientists ultimately think that by using models in science they are discovering and learning about the nature of the world. Yet how is it that they reason and discover things about the world using what they know are only partial representations? How is it that models allow for the ability to make seemingly true claims, and how should we understand the process of assessing a model that succeeds in this way? Most scientists believe that the success of a good model justifies making claims that go beyond the model itself. An important philosophical question arises here: what justifies these ampliative inferences? An understanding of how these inferences work is needed. Scientists want to be able to say they have discovered something about the nature of the world, not just about the model. But how is it that they are permitted to make this move from claims about a model, to claims about the real world?

My goal is to develop a framework that can help us precisely formulate such questions, and develop answers. This framework will allow for understanding the process of constructing and evaluating models, and what it means to say a model is good, or successful, or “fits”. Finally, it will provide an understanding of the justificatory process that allows scientists to make inferences from models to claims about the real world.

[3] “Phenomenon” refers generally, covering the general and stable features of the world that are of interest for the scientists or modeler.

1.2 Background: Modeling in Philosophy of Science

The philosophical examination of scientific models has predominantly focused on two key aspects. The first relates to ontology: what is a model? Attempts to determine what a model is also involve a second key question, how do models relate to theory? The syntactic view of theories (Carnap 1938; Hempel 1965) holds that a theory is a set of sentences in an axiomatized system of first-order logic. A model, then, is a system of semantic rules that offer an interpretation of those sentences. Most philosophers have abandoned this account in favor of the semantic view of theories. On the semantic view, a theory is constituted of a family, or set, of models (van Fraassen 1980; Giere 1988; Suppe 1989; Suppes 2002). While there are different versions of the sematic view, they all see models as the central unit of scientific theorizing. The function of the model is to represent part of the world. The scientific model is what represents the phenomena, features of the world, or the collection of data we obtain from observations. These are often treated as distinct types of models: theoretical models and data models.

A third account for understanding the relationship between models and theories argues for understanding models as "autonomous agents", relatively independent of theory, and functioning as "instruments of investigation" (Morgan & Morrison 1999, 10). A model is not something that is entailed by a theory. Rather, a model is a result of skilled construction on the part of the modeller, and through this construction it gains a partial independence from theory. In a sense, "models mediate between theory and the world" (Morgan & Morrison 1998, 242). On this "models as mediators" account, the role of models is understood as a tool that is used when theories are too complex to understand, or can be used in the development of a theory, or to complement a theory when the theory is incomplete.

Another approach to understanding what a model is—regardless of the relationship it holds to a theory—starts by looking at the sorts of models that exist in scientific practice, in order to determine the anatomy or possible forms they take. One recent analysis has identified at least three categories of models: Concrete, Mathematical, and Computational. Concrete models are physical objects that can stand in a representational relationship with the phenomena under investigation. Mathematical models are abstract structures whose properties can stand in a relation to mathematical representations of the phenomena. Computational models are sets of procedures that can potentially stand in relations to a computational description of the phenomena (Weisberg 2013, 7).

While the first aspect of the philosophical investigation of models focuses on what constitutes a model and its relationship to a theory, the second focuses on the relationship between the model and the world, as well as what goes into the construction of models. This line of investigation is often characterized as how models relate to phenomena, either directly or through data. Given the complexities of real-world phenomena, scientists often make judgments about what aspects of the phenomena are relevant to their questions or the investigation at hand. In order to develop a model, modellers often must first identify a *target system*. A target system refers to the selected part of the real-world phenomena that we seek to represent in our model (Suarez 2003; Giere 2004; Frigg 2010; Godfrey-Smith 2009; Weisberg 2013). The decision about what constitutes the target system can range from observations or a body of scientific evidence available (as is the case of models of phenomena), to more fine-grained set of data (as in the case of models of data) (Frigg & Hartmann 2012)[4].

Sometimes modellers will have a clear sense of a specific target system that a single model seeks to represent. This case can be referred to as target-directed modeling. There are also cases of nontarget-directed modeling, which comes in at least three varieties (Weisberg 2013). Generalized modeling occurs when a generalized phenomenon is chosen as the target. For example, rather than constructing a model of a specific black hole, we may want to construct a model for black holes generally. Hypothetical modeling involves modeling possible target systems. In this case, we construct models about a

[4] It is worth noting that while the literature on scientific modeling has focused substantially on the *how* models represent though approximations and idealizations, little work has been done on the role of evidence in the context of making decisions regarding *what* constitutes the target system that is then represented by the model. Hughes (1997) discuses this question tangentially though examining how we learn from models. More recently, in her doctoral thesis, *Target Systems and their Role in Scientific Inquiry,* Elliot-Graves (2014) begins to address the larger question of what constitutes a target system.

target system that might not actually be instantiated in the real world, such as a model of a perpetual motion machine. Finally, targetless modeling involves modeling in which the “target system” is not a real world target but rather a model itself[5].

Models represent target systems through means of approximations—an inexact description of a target system—and idealizations—the creation of a new system, some of whose properties approximate some belonging to the target system[6]. There are many types of idealizations that can be made in the construction of models. McMullin (1985), for example, distinguishes six types of idealization: mathematical, construct, formal, material, causal, and subjunctive. Weisberg (2013) reduces this to three: Galilean idealizations, minimalist idealizations, and multiple-models idealizations.

Regardless of the particular philosophical question under investigation in any of these discussions, there is general agreement that the greatest challenge of using models in scientific reasoning is related to the fact that models are partial and incomplete representations of their target systems. Models make approximations or idealizations, or they are highly simplified and incomplete representations of target systems. Parties to this debate consider models to be false[7]. Some philosophers have even argued that models should be thought of as fictions; they represent entities that do not actually exist and are never instantiated (Contessa 2010; Frigg 2010; Godfrey-Smith 2006; 2009). Nevertheless, we use these false models, and consider them effective tools for making predictions, providing explanations, and helping to establish true claims about the real world (Wimsatt 1987; 2002).

[5] For example, the “Game of Life” is a cellular automaton model. Each cell can be in one of two states, “alive” or “dead”, and must follow four simple rules (for more details see Conway 1970, Weisberg 2013). A targetless model has no real world target chosen at all. Rather, the system of interest is a model itself, in this case the Game of Life model. The “model” in this case is a model of the model the aims to explore the functioning of the model itself.

[6] This characterization of idealization and approximation is taken from Norton 2012. It should be noted, however, that he does not provide these characterizations in the context of how models represent target systems.

[7] In chapter 2, I will argue that models are not truth-evaluable themselves; so I strictly speaking reject the claim commonly made in this debate that models are false.

Given that models are incomplete, partial, and in some sense false, what justifies their use to make claims about the real world? A significant amount of the philosophical literature has focused on these two aspects: what a model is (and understanding the relationship models hold to theories) and how we construct models (and how to understand them as representing the real world given that they are incomplete, or partial representations). However, there are open questions in philosophy related to a critical third aspect related to scientific models, which arise after a model is constructed. What does it mean to say a model is good or successful, and how is this evaluated? How does our evaluation of models justify and inform our claims about the nature of the real world? Ultimately it is these questions that are of concern in providing scientific justification, and thus need to be answered.

The only means by which these questions have been discussed thus far in the literature is by extending what philosophers have argued about theory confirmation to model confirmation. Confirmation theory explains how empirical evidence confirms the truth of hypotheses and theories. In the context of model evaluation, confirmation theory will also explain how the empirical evidence confirms the model. For example, in evaluating a model the goal is to look for empirical or confirmatory virtues, "a virtue that indicates that a model or models are more likely to be used to represent accurate or true claims about the observable world" (Lloyd 2015, 58).

An alternative is to discuss model evaluation in terms of validation, rather than confirmation. Validation, employed in the context of modeling, refers to the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended use of the model (Calder et al. 2002, Jebeile & Barberousse 2016; Oreskes, Shrader-Frechette & Belitz 1994, Thacker et al. 2004). Given the idealizations and approximations contained in models, and that models are evaluated relative to a specific use, it may not be the case that model confirmation proceeds in precisely the same manner as theory confirmation[8] (Weisberg 2006; forthcoming).

In this dissertation, I will approach model evaluation without making any prior commitments to the process involving confirmation, or validation. Rather, I want to explore the question of the evaluation of success of a model by focusing simply on how it proceeds in scientific practice. In developing this framework, I will not assume the process of model evaluation is the same (or different) from theory evaluation. It is possible that the arguments contained in this dissertation will have implications for the debate about model confirmation versus validation. However, this is beyond the scope of the present work. This dissertation contains the positive argument for a framework for understanding the evaluation and justification of models, and does not directly address the implications of this view for other debates.

For these reasons, my framework will take the work of Michael Weisberg and Wendy Parker as a starting point. Their analysis of model evaluation does not rely on extending theory confirmation to models. Rather, they start by examining scientific practice. While they both do draw conclusions about validation or confirmation of models from their analysis, their work on model evaluation is separable from these claims. I can therefore separate their work on evaluation of models from the context of validation and confirmation, and discuss it instead in the context only of model *success*.

Furthermore, Weisberg and Parker have each respectively developed the current strongest starting positions for possible understandings of model evaluation: similarity and adequacy. Weisberg (2013, 2015, forthcoming)—following other philosophers such as Cartwright (1983), Giere (1988), Teller (2001), and Godfrey-Smith (2006)—argues that successful scientific models stand in a relation of similarity to their real-world target systems. Parker (2009, 2010, 2015), however, argues that successful models are evaluated as being adequate, or sufficient relative to a given purpose. Part of my goal is

[8] The relationship between models and theories has also been discussed above in section 1.2.

to assess if one, both, or neither of these approaches is the better understanding of model evaluation.

Following Parker and Weisberg, I understand models to be entities that represent parts of the world[9]. The representational relation a model holds to its target system is analogous to the relation a picture can hold to its subject. The model captures details at a certain level of resolution of the target system. Like a picture, a model can be used to learn things about the world, and in this sense is a tool. This view of models fits most naturally with the idea that models have a high level of independence from scientific theories. This is similar in certain respects to the "models as mediators" view. While models can rely on theory in their construction, the end product, or model itself, can be assessed mostly independently of theoretical considerations. Despite the fact that Weisberg does not explicitly endorse this view, I will show through close analysis of his weighted feature-matching equation, that he has this kind of relationship in mind. As I will detail in the following chapters, the model's similarity to a system in the world is one of the primary components of model assessment. Our understanding of that target system may, in some cases, have a high theoretical component[10]. However, this is the extent of the interaction between models and theory in the framework I develop in this dissertation.

I will provide an account of what it means to say a model is successful through answering three questions: What is the target of evaluation in model evaluation? How does that evaluation proceed? And finally, what licenses us in making inferences about

[9] In philosophy of science, the problem of representation focuses on two problems. The first problem is to explain in virtue of what a model is a representation of something else, the second focuses on what representational styles there are in science (Frigg 2006; Frigg & Hartmann 2012). However, there is no consensus on solutions to these problems. So, I have adopted what I take to be the most general account of representation currently utilized in the philosophical discussion of scientific models offered by Ron Giere (1988; 1996; 1999; 2004): the way in which models represent is similar to the way in which pictures represent. He provides further detail about this relation as that of similarity, which I will discuss in substantial detail throughout this dissertation. Steven French (2003; and French & Ladyman 1999) also provides a stricter position along this line by adopting a representational relation of isomorphism. Given a relation of isomorphism is stronger than similarity, I have adopted the more modest positon.

[10] Since similarity relation used in my framework is most closely based on Weisberg's work, his view of model-theory relation is the one I adopt.

the real world, based on the evaluation of our models as successful? Through analyzing these questions, I will develop an understanding of how a model can tell us true things about the world.

1.3 Overview of Thesis

What is needed in the philosophical discussion of scientific modeling is a fully general framework in which we can trace the path of justification of a given model. Such an account would be a unified way to discuss modeling, and a means for understanding the claims being made across various disciplines that use scientific models. My project is an attempt to develop such a framework. Even if such a fully general account, ultimately, is not possible, the attempt in itself is valuable. This is because it will succeed in unifying some evaluative processes, as well as help identify where current evaluative standards need to be disentangled and refined. It is plausible that in many cases of conflicting models, even a partial account can be used as a tool to determine where, precisely, the conflict lies.

I will argue that an important point has been overlooked thus far in philosophical work: the evaluation of a model actually contains two conceptually distinct parts, or components of model evaluation. The first part is evaluation of how similar the model is to the target system it is meant to represent. This part of evaluation happens primarily during the construction stage, when the model is being developed. This process is dynamic, and takes place over many iterations during the model's development. The second part of evaluation compares the output of the model with the analogous phenomenon in the target system. For example, if the model outputs a prediction, it is then compared to a later state of the target system it intends to model to see if the predicted phenomenon in fact came to pass. This evaluation assumes a completely constructed model that generates a particular output, and that we can obtain an analogous output from the target phenomenon. This second part of evaluation is different from an evaluation of similarity, in that it assesses the adequacy of a model. Evaluation of a model's adequacy is different from an assessment of a model's similarity in that adequacy is concerned with what the model does, while similarity is concerned with how the model represents. These two evaluations should not be seen as competitors, but rather be seen as targeting two different components of an overall assessment[11].

The central concept at play in each of these components of evaluation is *purpose*, the reason for which the model is created or for which the model is used. The general kinds of purposes that are relevant for modeling are prediction, explanation, and description. These purposes will determine features of the target system that are included in the model, the structure of the model itself, and the kinds of outputs the model generates. The purpose for which the model is intended will determine what counts as "similar" during the construction stage. Likewise, the purpose to which the model is being put (whether it be prediction, explanation, or description) will determine how we evaluate its "adequacy" at the output stage. Of course, a model that was constructed to provide predictions can also, for example, be evaluated for its adequacy with respect to its ability to give explanations. The purpose for which it was constructed does not constrain the purposes to which it can be put. But keeping track of the role purpose plays in both parts of the evaluation will allow us to develop a clearer picture of what I will call the overall *fit* of the model—a combination of its similarity evaluated at the construction stage, and its adequacy evaluated at the output stage.

Drawing these distinctions will provide the resources to better understand the success of our best models and the failure of those that fall short. But most importantly, it will allow for identifying how a model can fit well with respect to one purpose while failing to do so with respect to another. There is therefore a possibility that some models that were deemed unsuccessful for failing along one dimension of evaluation can be judged as successful, with the understanding that they are successful only with respect to a particular purpose.

[11] In the formal epistemology literature recent work has focused on how to characterize evidence in terms of accuracy and coherence. However, in science there is often a broader notion in mind where anything that has any bearing on the truth or falsity of theoretical commitments or hypotheses could potentially count as evidence. This issue is underdeveloped in the philosophy of science, and I will not be directly addressing it here. The framework I develop in this dissertation draws on examples in which what counts as evidence is unambiguous. While the framework likely will not help identify what counts as evidence, it useful for identifying when evidence makes a difference for the hypotheses.

There are debates ongoing today to which this framework will make an immediate difference. In particular, debates surrounding the Lambda Cold Dark Matter (ΛCDM) model in astrophysics could benefit from exactly the conceptual clarity provided by this framework. Many of the criticisms that this model faces would be appropriate only if it were intended, and therefore constructed, with a particular purpose in mind. They fail to have as serious an effect when the actual intended purpose of the model is identified.

The framework I propose begins by answering the question of what considerations the modeller needs to take into account when constructing a model. The most important question is why the model is being developed in the first place, or what the purpose of the model is. The process often proceeds by the modeller identifying a target system in the world that they want to model, and then subsequently considering the purpose of the model. However, it is consistent with this framework that the purpose partially determines the target system of interest.

The important consideration at this stage is *similarity*[12]. The modeller needs to build a model that is similar in the relevant respects to the target system. The purpose determines what that similarity relation looks like—the features, or ways in which the model is similar to the target system. If the purpose is to describe the target system, then simply having features of the models that stand in for the features of the target system may be sufficient. However, if the purpose is to predict, then the sorts of features of the target system that are included will be different. For this part of my framework I will draw on Michael Weisberg's weighted feature-matching account of similarity. This is because Weisberg has developed the most complete account of model-world relations in the context of assessing similarity. Any use of the term *similar* will be reserved for talking about evaluative considerations, like those identified by Weisberg, made at this stage.

[12] Similarity is roughly characterized as resembling without being identical. This definition will be made precise in section 2.4.1 where I introduce the weighted-feature matching definition of similarity.

In my framework, all models are considered to produce an output. The simplest case is a predictive model, where the output is a future (or past) state of the target system. But other things can count as outputs as well. These can be structures in the model itself that can feature in an explanation, or a description of the target system. Which outputs are of interest will depend on how we are using the model. It may seem more natural to reserve "output" for a prediction generated by a model. However, I am employing "output" in a broader sense. An output is purpose-dependent and can vary based on what question the model is used to answer. This can include, in addition to questions about predictions, questions related to interrelations of the structures in the model itself, or questions related to what the model represents. While this may seem like a strange usage, the reason it is employed here is so that I can talk about the different ways in which we use models, while employing the same terminology. While my framework could be developed with a more complicated terminology—reserving "output" for predictions and including terms for model structure relations and representation—it would not make a substantive difference.

The second sense of evaluation that comes up in this framework occurs when the output of the model is compared to the target system in the real world. At this stage, what we evaluate is whether the model is *adequate* for the purpose for which we are currently using it. It is important to note that a model that was constructed with one purpose in mind can be evaluated relative to its adequacy for another purpose. This discussion of adequacy for purpose will draw on the work of Wendy Parker. The reason for starting from Parker's work is that she provides an account of how modellers assess their model's adequacy in the context of climate change. I believe this idea is generalizable, and thus I want to apply it in a broader context. All discussion of *adequacy* will be reserved for this component of evaluation.

The final element of the framework is what I will call an overall assessment of model *fit*. The idea of model *fit* is used extremely loosely in the philosophy of science literature, as well as in science more broadly. The statement "the model fits" could refer to a model having a certain degree of similarity to the target system, to a model fitting data points, or to a model's prediction fitting with our observational data or best-supported theory. Such ambiguity can lead to confusion about what part of a model is being evaluated as “fitting”. Both Weisberg and Parker use the concept of fit. But as I will argue, fit has a different meaning for each of them as they are evaluating different aspects of the model. For these reasons, it is essential to provide a specific definition of what I will mean by model fit.

In my framework, model fit will refer to a claim about the assessment of both the similarity of the model, and the adequacy of the model, relative to a specified purpose. The concept of fit admits degrees; some instances of fit can be stronger than others. For example, cases in which we are able to include many features of the target system in the model, and minimize extraneous features, will be a stronger fit than cases in which we have not been able to include all important features of the target system of interest. Likewise, instances in which the model’s output is a closer match to the equivalent output in the real world are a stronger fit than those in which the output does not match as closely.

In the end, my framework will provide guidance about the kinds of inferences that scientists are justified in making about the world from the models. The justification for inferences is grounded by (1) establishing a positive assessment of similarity of the model relative to the intended purpose and (2) establishing a positive assessment of the model’s adequacy for a particular purpose. The details embodied in these assessments inform us about how the model makes connections to the real world, as well as the limits of the model’s successful use.

In what follows, I provide further details of the arguments I make for this framework in the subsequent chapters.

1.3.1 *Chapter 2: Constructing Hypotheses about Models*

When scientists conclude that they have a successful model, what exactly is it that they are evaluating, and what is it that they are gaining knowledge about? I begin this chapter by examining what the target of evaluation is in the case of models. Following other philosophers, I take it that model evaluation is not the evaluation of the truth of the *models themselves*, but rather the truth of *hypotheses* regarding the utility of models for different purposes. If this is the case, the second issue—and main focus of this chapter—is to determine what form these hypotheses should take.

Wendy Parker and Michael Weisberg have each provided possible formulations for hypotheses regarding model evaluation. Parker proposes model evaluation as a matter of a model's *adequacy* relative to a purpose to which the model is being applied (2009; 2010). Weisberg proposes model evaluation as a matter of establishing the model's *similarity relation* in the desired respects and degrees relative to its purpose (2013). This similarity relation is captured by a "weighted feature-matching" equation. I examine these two approaches for evaluating the success of a model, and their proposed formulation for hypotheses regarding model evaluation. In light of criticisms of the similarity account made by Parker (2015), I focus on further developing the understanding of the similarity relation hypothesis, and the weighted feature-matching equation offered by Weisberg.

The similarity relation hypothesis and weighted feature-matching equation elucidate an incredible number of valuable elements that modellers consider in constructing their models. However, I argue there are two main problems. Weisberg accounts for the various elements that go into model construction in his formalized weighted feature-matching equation, *S(m,t)*. From this *S(m,t)* equation we obtain a numerical value between zero and one quantifying how similar the model is to its target system. My first criticism is that, while conceptualizing model construction in this manner and formally accounting for the process is extremely valuable, it may not, in practice, be possible to obtain a numerical score. And even if we are able to obtain a score, it is not clear that we should want to use such a score, as it may have bad epistemic consequences. I argue that we should only consider the *S(m,t)* equation as an extremely informative tool and means for explicitly formulating the evaluative elements and decisions that occur at the model construction stage.

My second criticism is that Weisberg has provided insufficient detail with respect to how a model's purpose impacts the similarity relation, and the weighted feature-matching equation. I argue that the similarity relation hypothesis must be modified to explicitly include considerations of the domain of application of the model. Weisberg's account will encounter a challenge that, without a domain specification, it cannot solve: instances in which we have the same target system, same model, and same purpose, yet the model should obtain different similarity scores for the different domains of application. I provide a case study of a mathematical model of stellar implosions to make this case. Domain of application must be made explicit if one is to draw on the weighting feature-matching equation. Without it, it is not possible to properly specify under what conditions the model is, in fact, similar to the target system.

However, these critiques are not detrimental to the similarity account. Rather, with my proposed modifications—using the weighted feature-matching equation as a pragmatic guide, and including the domain of application explicitly in the hypothesis statement—I provide the strongest version of the similarity relation hypothesis for model evaluation. The similarity-relation hypothesis is an extremely informative and effective way to formulate a hypothesis statement. It forces us to enumerate the elements of the target system the modeller has chosen to include in the model explicitly. It also tracks what we choose to include in the construction of a model, and how the evaluation of the construction is relativized to a purpose within a certain domain of application.

1.3.2 *Chapter 3: Evaluating Hypotheses about Models*

With this modified version of the similarity-relation hypothesis, I turn to the question of whether one of the hypotheses—either the adequacy-for-purpose or a similarity-relation hypothesis—should be favored as best capturing the evaluation of model fit in scientific practice, as well as the question of how model fit should be evaluated through such hypotheses. By examining how Parker and Weisberg respectively propose to evaluate their hypotheses, I argue that each hypothesis actually has a different target of evaluation, and that, in the end, *aspects of both similarity and adequacy need to feature in an account of evaluation of model fit*.

I argue that the two hypotheses work together in the following way: An assessment of a similarity-relation hypothesis is involved when evaluating the relation between the model and the target system during the model's construction. For this component of model evaluation, we should employ my modified similarity-relation hypotheses. Evaluation of the model's adequacy for purpose is about evaluating the output of a model and comparing it to the equivalent output phenomena of the real world. These two aspects are different in that an assessment of adequacy is concerned with evaluating what the model *does* and its effectiveness for that aim, while assessment of similarity is concerned with evaluating how the model *represents*. I argue that *it is only when we take both hypothesis statements together, that we can evaluate the overall fit of the model.*

I propose a framework in which assessment of model fit is understood through four components. The first component involves constructing the model and establishing the similarity relation via the weighted feature-matching equation. The second component involves, through reasoning or calculation, obtaining an output from the model. This component also involves determining what would be observed as the output in a certain test situation if the model is effective, or adequate for the purpose. The third component involves comparing and evaluating the level of agreement of the model's output with the analogous output from the target system. The fourth component involves an assessment of the model's overall fit through a final evaluation of our two hypotheses. The assessment of the adequacy-for-purpose hypothesis addresses whether the model is qualitatively or potentially quantitatively satisfactory for the purpose at hand. The similarity relation hypothesis addresses the standards by which the model is assessed to be similar to, and to representative of, the target system for the given purpose.

1.3.3 *Chapter 4: Making Inferences from Models*

Having established a framework in which evaluation of model fit is done through assessment of a similarity and adequacy of the model for a given purpose, I return to the larger question at hand: What justifies making inferences from a model to knowledge claims about the world?

I argue that the justification for extending claims about a model to the world first requires explicit attention to the scientific purpose of the model, since both the assessments of similarity and adequacy are always made relative to a purpose. While the particular purposes to which any given model is put can be quite specific, I argue that there are three general kinds of purpose: descriptive, predictive, and explanatory. The difference is related to the kind of output obtained from the model when attempting to use it for a particular purpose. In the case of a descriptive purpose, the modeller obtains from the model an output that somehow represents the features present in the target system. In the case of a predictive purpose, the modeller obtains from the model an output corresponding to a future or past state of affairs about the target system that is not originally built into the model. In the case of an explanatory purpose, the modeller obtains from the model an output that can serve as an explanans in an explanation of some phenomenon.

The second part of the argument in this chapter is related to how inferences from a model about the world are justified. While assessments of adequacy for purpose can evaluate whether a model is successful for one application relative to one purpose, it is not what justifies extending models to a new purpose or new domain; nor does it ground inferences made from models. I argue that a model having a high degree of similarity relative to a purpose-dependent *S(m,t)* is what provides justification that allows for determining when a model should or should not be extended, whether the model must be modified in order to serve a different purpose, and ultimately the inferences that can be made about the world from the model. It is because one can identify how the model is similar to the target system in the relevant ways that one can determine the appropriate level of confidence in drawing conclusions about the target system that go beyond the information that was built into the model in the first place.

Through examples of modelling from astrophysics, I demonstrate how my framework can be deployed as a tool to gain insight into success claims about models and a means for understanding the connections between similarity, adequacy, fit, and justification for inferences about the world. The astrophysical examples are used to support both my argument for the three general kinds of purpose models can serve, and my claims about similarity grounding the extension of models.

1.3.4 *Chapter 5: Tracing the Path of Justification for ΛCDM and MOND*

The final chapter examines a case from astrophysics in which analyzing the debate in terms of the framework I propose can make an immediate difference. The Lambda Cold Dark Matter (ΛCDM) model is considered to be the current best model of large-scale structure formation. However, part of the model posits that 84% of the mass of the universe is made of matter we have never seen, *dark matter*. Some astrophysicists consider this strange matter to be an unjustified ad hoc addition to the model introduced to ensure the model fits the data. In response, some of these critics have proposed (contentious) alternative models, which fit the same data by Modifying Newtonian Dynamics (MOND), such that positing the existence of dark matter is not required. MOND proponents view their models as equivalent or superior in some respects to ΛCDM models.

This is a case in which there are two models that include different elements, and even differ fundamentally in terms of the theory on which they are based. Yet both models have been evaluated as models that successfully describe the observations, make adequate predictions, and even offer explanations. How can a claim like this be understood? How should we deal with situations in which there are two models, that seem to contradict one another, yet are both evaluated as having a good fit?

While one option is to regard this as a case of Kuhnian incommensurability, or as a case in which a purely subjective choice must be made, I argue that the debate should be understood as one primarily about choosing the purpose of models, and then assessing whether they are useful for that purpose. Through the framework I have developed, I demonstrate how both ΛCDM and MOND can be considered well-justified, high-fit models given different choices about what to prioritize. I argue that both models can be evaluated as having good fit, when considering their fit within their respective domains of application. The apparent conflict between the two models arises due to extending both models past the domains in which they are successful. In attempting to extend each model to new domains, the modeller relies heavily on the model's explanatory fit. But, extending claims of explanatory fit relies on strong commitments to the similarity relation established, particularly with respect to the way the model represents theoretical commitments (as will be seen in the case of both the ΛCDM and MOND models).

Chapter 2

2 Constructing Hypotheses about Models

2.1 Introduction

Models are used in science as a tool to understand and investigate the world around us. The scientific practice of modeling is the indirect study of real-world systems through the construction and analysis of models. One of the main goals in model-based scientific reasoning is to construct successful models. But what is the right way to understand claims about the success of models? One of the main ways in which a model is considered to be successful is when the model fits well with a part of the world, or a *target system* under investigation. However, it is not clear what exactly "fit" means, or how this fit is evaluated in practice. My first goal is to provide an account of what is meant by the claim that a model fits the target system under investigation.

Providing an account of what it means for a model to fit involves answering three questions. First, what is the target of evaluation in model fit? Ultimately, I argue that model fit is a complex notion; "fit" must be understood as a composite evaluation including assessment of both the *similarity* and *adequacy* of a model. The more complete picture of model fit is fleshed out in chapter 3. In the present chapter, I introduce and discuss these two possible options: evaluating model success in terms of assessing similarity and in terms of assessing adequacy.

The target in model evaluation, I argue, is not the evaluation of the truth of the *models themselves*, but rather of *hypotheses* regarding the utility of models for different purposes. If we do not evaluate the model itself but rather a hypothesis about the model, the second question regards the form hypothesis statements about models should take. After determining what form the hypothesis should take, the final question is how the hypothesis should be evaluated. In this chapter, I address the first two questions. The third follows in chapter 3.

The strongest arguments for the possible forms a hypothesis about model fit should take are offered by Wendy Parker and Michael Weisberg. Parker (2009; 2010) argues that successful model fit is about assessing the *adequacy* of a model for its particular purpose. A model is successful when the model is adequate for the purpose the modeller intends to use it for. Weisberg (2013, 2015 forthcoming) argues that evaluating model fit is about evaluating a *similarity relation* between the model and the target system the model is constructed to represent. Weisberg provides a detailed account as to how we should understand this similarity relation and offers a *weighted feature-matching* equation of similarity, in which the similarity of a model to a target system can be computed as a scalar value between zero and one.

My goal is to assess these two approaches for evaluating the success of a model. This chapter will proceed as follows: In §2.2, I briefly review why evaluating models ought to be conducted via evaluation of hypotheses about models, rather than evaluating the truth of the models themselves. I then provide an overview of the two candidate formulations for these hypotheses. Parker's account of evaluation of models via assessment of adequacy for purpose seems promising, as she addresses what she means by "adequacy" and "purpose". However, I claim that Parker's account also relies on some kind of assessment of similarity, although she is not explicit on this point. I examine how she understands similarity and the relation it holds to evaluations of adequacy for purpose.

With respect to Weisberg's similarity account of model evaluation, I argue that this account faces two problems related to the formalization of similarity relations via the weighted feature-matching equation. In §2.4, I provide details of his account, and identify the main weaknesses as it is currently formulated. Constructing Weisberg's weighted feature-matching account of similarity is highly complex. So in §2.5 I provide a detailed example of how the similarity hypotheses are constructed using a historical example—the black hole model constructed by J. Robert Oppenheimer and Harland Snyder. I use this example to argue that, at best, formalizing similarity relations via a weighted feature-matching equation that outputs a value between zero and one is not possible; at worst, this has bad epistemic consequences. I argue that it is better and more useful to use this equation as a heuristic tool for explicitly enumerating the elements that go into the construction of the model, rather than computing a numerical value of similarity. Second, and more importantly, I argue that the mechanics of the weighted feature-matching equation have not accounted for cases in which a model is applied outside of its intended domain.

In light of these criticisms I propose a modification to the form of the similarity relation hypothesis and weighted feature-matching equation in order to account for these problems. This modification requires explicit inclusion of the domain of application of a model. My proposed modification to the similarity account supports the assessment of model construction in terms of a similarity relation and weighted feature-matching equation as an extremely informative and effective means by which to form a hypothesis for model evaluation.

2.2 What is the Target of Model Evaluation?

Scientists attempt to learn more about the nature of the world around us through representing certain features of the world in a model. However, when they construct a model, they need to be able to evaluate whether the model is a successful representation of the part of the world the model is intended to represent. When they conclude that they have a successful a model, what exactly is it that they are evaluating, and what is it that they are gaining knowledge about?

One option is that the assessment of a model is concerned with the truth of the model[13]. On this view, a model somehow contains or embodies a truth evaluable claim about the system it represents. The truth of the model itself is supported[14] by various instances of the model output matching the observational data. Another option, however, is to argue that models by their very nature are false. In constructing a model, idealizations, approximations, and assumptions are made. On this view, models cannot be true representations of the target system (Wimsatt 1987; 2002).

[13] For example, Elizabeth Lloyd argues for such a position in the context of climate models in her 2009 paper, “Varieties of Support and Confirmation of Climate Models”.

[14] As mentioned in chapter 1, I am actively choosing not to present an account in terms of confirmation, a term that in this instance could easily be substituted in.

A third option is to argue that to attempt to evaluate models as being true or false is a category mistake, as they are not the sort of entities that are candidates for truth evaluation. On this view, a model is a representational tool that does not contain or embody truth evaluable claims about the system it represents. A good analogy for understanding this point is a painting: a painting may represent a part of the world, but is not itself truth evaluable. Statements about how the painting represents the world are truth evaluable, but these are not part of the painting itself. In the same way, models are *used* in various hypotheses, such as, "the model *m* is adequate for the purpose *p*" (Parker, 2009). What is supported by observational evidence is the truth of this hypothesis, not truth of the model itself. As such, when evaluating a model, we are not evaluating whether the model itself is true. Instead, we evaluate a hypothesis as a claim about the model.

I agree that models are not truthful representations; they are at minimum false or, at most, not candidates for truth evaluation. Nevertheless, we use models as effective representational tools in helping to establish true claims about the real world. For my purposes, I adopt the view that hypotheses are the direct target of evaluation because this is the more general case. Whether or not models are truth evaluable themselves, using a model to generate knowledge that goes beyond the truth of the model itself is still an important element of scientific reasoning. In order to do this, one must employ the model in a hypothesis about the real world[15]. In order to assess whether a model of fluid dynamics is successful in a new situation, such as modeling the dynamics of galaxies, we need to embed the model in a hypothesis statement[16] about this new situation, the content of which is not related to the content of the original model. As a result, to say that a model is successful is to evaluate some hypothesis about the model fitting with the target it intends to represent to a certain level of acceptability.

[15] It is not necessary that the model bear any prior relation to the target system before one undertakes this evaluation. That is to say, it is possible to employ a model that was constructed to represent a given target system in a hypothesis statement about an entirely different target system and assess its similarity and adequacy.

[16] By "hypothesis statement", I simply mean the form of a hypothesis, such as the hypothesis being formed as "model *m* is adequate of purpose *p*", or "model *m* is similar to target *t* for purpose *p*".

2.3 What Form Should the Hypotheses Take?

If model evaluation is a matter of evaluating hypotheses as claims about the model, then there are two important questions to answer: (1) What form should the hypothesis statements take? and (2) How do we evaluate the hypotheses? In this chapter, I begin to answer the first question by examining two candidate formulations for these hypothesis statements. Wendy Parker argues we are evaluating the hypothesis that the model is adequate for its intended purpose. Michael Weisberg thinks we evaluate a hypothesis with respect to whether a model is similar enough to the target system for its intended purpose. I assess the benefits of each of these views, and ultimately argue that we should adopt a view that combines elements of both[17].

2.3.1 *Similarity-Relation Hypotheses*

One possible form the hypothesis could take is framed in terms of assessment of a model's similarity to targeted aspects of the real world. When scientists want to investigate a certain phenomenon, they construct a model. Modellers choose certain aspects of the real world to be represented in the model. If representations of certain properties of the world are included in the model, it will be similar enough to the real world that the model's output can be trusted as telling us something about how the world will behave. A successful model will be evaluated as being *similar* enough to the target system for a certain aim.

Weisberg, following Cartwright (1983), Giere (1988), and Godfrey-Smith (2006), understands this central model-world relation to be that of *similarity*. Models are not truthful representations[18] or isomorphic to their target systems; rather, they stand in a relation of similarity with their real-world targets in relevant *respects* and *degrees*, where those respects and degrees depend on the information sought by the modeller. According

[17] I will return to discussion of the second question, of how we evaluate these hypotheses, in chapter 3. I will argue that, in examining the accounts of evaluating the hypotheses about model fit, it is evident that assessments of similarity and assessments of adequacy have different targets of evaluation. Any final account of model evaluation will need to include elements of both.

[18] Models are not truthful representations given that they make idealizations and approximations.

to Giere, this understanding can take the form of evaluating a *theoretical hypothesis*—a statement, claim, assertion, or conjecture, about the relationship between the theoretical model and certain aspects of the world (Geire, 2006, 25). These hypotheses are of the form:

> Model *m* is similar to the world in the desired respects and to implied degree of accuracy.

If the model is similar in the desired respects and degrees, the theoretical hypothesis is true[19].

Weisberg identifies three issues that must be addressed in order for this account to be developed into a more complete picture of the model-world relation[20]. First, there must be a precise formulation of what the similarity relation actually is. Second, given that there is a sense in which every model is similar to every target in some respect or other, there needs to be a principled way of specifying which respects are relevant. Finally, the view must account for the pragmatic dimensions of modeling (Weisberg forthcoming, 10). Weisberg proposes understanding these contextual factors through his *weighted feature-matching* account of similarity. He suggests generating Giere-like theoretical hypothesis of the following form (forthcoming, 12):

> Model *m* is similar to target *t* for scientific purpose *p* to degree *S(m,t)*,

where *S(m,t)* is his weighted feature-matching equation. A model can be successfully applied to the target when the model fits the target, where fit is understood as this model-world relation (2013, 93). Given that Weisberg considers the best account of the model-world relation to be that of similarity, model fit should be understood as a hypothesis of the form given above.

[19] Giere's terminology ('desired respects', 'implied degree', and 'theoretical hypothesis') reflects the fact that he considers models to be a subset of theories. For my purposes, I discuss a more general framework that is not committed to this relation. The points made in this chapter stand regardless of how one understands the model-theory relationship.

[20] Understanding the model-world relationship in terms of similarity was criticized by philosophers such as W.V.O. Quine (1969) and Nelson Goodman (1972). Weisberg's attempt to develop a more complete account aims to address these criticisms by improving Giere's account.

Weisberg's weighted feature-matching similarity equation attempts to formalize the model's representational relation of similarity to a target system as a function of the features that the model and the target share, penalized by the features they do not share. It is the modeller's *construal* that determines "the choice and weighting of these important features" (2015, 299). Weisberg claims that the broader scientific context will inform the modeller's construal. The construal of a model is composed of four parts: *assignment*, *scope*, and two kinds of *fidelity criteria*. Assignment and scope tell us how the real world phenomena are intended to be represented in the model. Fidelity criteria are the standards theorists use to evaluate a model's ability to represent phenomena.

2.3.2 *Adequacy-for-Purpose Hypotheses*

A second possible form the hypotheses could take is framed in terms of assessment of the model's adequacy. Scientists construct models with the goal of being able to obtain predictions or explanations. The success of a model is then related to whether the model is *adequate* for providing this desired result. Parker (2009; 2010) develops the strongest position for an understanding of model evaluation in terms of adequacy by examining how to understand instances of fit between observational data and model predictions in the context of climate models. She understands models to be representational tools: "A model is a representation in that it (or its properties) is chosen to stand for some other entity (or its properties), known as the target system" (2009, 235). A model is also a tool in that it is intended to serve some particular purpose. If a model can be said to embody a truth-evaluable hypothesis, it is one that is usually known from the outset to be false. However, she thinks models will nonetheless be evaluable as adequate for the purpose of interest to the modeller. As a result, she believes we should understand model fit in the following way:

> What these instances of 'fit' might be said to confirm (or support, or raise the probability of), if anything, are hypotheses about the adequacy of climate models for particular purposes. An example of such a hypothesis might be: This climate model, when run from these initial conditions, is adequate for the purpose of predicting whether Earth's global mean surface temperature would increase by more than 2°C between now and 2100 under this emission scenario (Parker 2009, 236).

Generalizing from this statement about climate models, I take the form of an adequacy-for-purpose hypothesis statement to be:

> Model *m* is adequate for intended purpose *p*.

These sorts of hypotheses are evaluated as true when the model constructed is adequate for the purpose at hand.

If we are to accept Parker's proposal as the proper form of the hypothesis statement, we need to know how to understand the concepts of *adequacy* and *purpose*. With respect to the latter, She says that the "purpose" of a model typically involves answering some limited range of questions about the target system. Further, she says the relevant purpose will often include *simulating* aspects of the past, *predicting* aspects of the future, or *explaining* (providing information about the causes of phenomena of interest) (2009, 235-7). With respect to the "adequacy" of a model, she understands this as the idea that a model, "when used in accordance with specified methodologies, will convey information about the target system that allows model users to infer correct answers to the target questions" (2009, 236).

However, as a footnote to this conceptual definition of adequacy, Parker indicates that adequacy also relies on some sort of similarity:

> As understood by this paper, an adequate model is one that is sufficient for the purposes of interest not just by chance or luck but because it is similar enough to the target system in relevant respects. Which similarities are relevant and what counts as similar enough is determined by the purposes at hand (Parker 2009, 236, footnote 6; also see 2010, 4, footnote 7).

It seems that Parker understands the adequacy of a model as somehow being related to whether the model is similar enough to the target system in certain respects. She says here that our model gets things right not because of luck, but by virtue of some sort of similarity to the target system. Does Parker mean for adequacy-for-purpose assessments to be *related to* or *grounded in* similarity? If so, she has not told us how one ought to understand the relationship of similarity; nor how purpose informs what counts as "similar enough". If not, why has she bothered to mentioned similarity at all?

In looking to her other work, Parker seems to take adequacy to go beyond what assessments of similarity offer. For example, she states,

> Demonstrating that a model is similar to its target in various respects and degrees is not enough; one should have some reason to think that these are sufficient respects and degrees, given the intended uses of the model (Parker 2015, 275).

I take it that Parker thinks that similarity *per se* does not answer the question of whether a model can reliably provide new information about the target system. A model being similar enough to a target system for one intended use does not necessarily mean that the model's similarity will still be relevant when a model is used in a new context, for a different hypothesis, or to answer a different question. Whenever a model is used in a new way to address a new question about the target system, the grounds for a modeller to consider a model similar will need to be "revisited", and the modeller will need to consider whether the model's similarity is sufficient to provide this new information about the target system.

I agree with Parker in that an account of similarity alone may not be enough; yet we want to say that a model is adequate not by chance, or luck. If the ultimate goal in using models is to make claims about the world, there likewise needs to be something that grounds or connects an adequate model with the target system. A model bears, in some sense, a similarity to its target system in that there is something that gives a modeller good reason to think the model stands in some relevant relation to the real world. Yet at the same time there is a need to guard against the fact that every model is similar to every target system in some respect or another. If similarity plays any role, there needs to be a way to say what similarities are *relevant*. Given that models make idealizations and approximations in their construction, we must be able to say more about what aspects of the target system the modeller has chosen to prioritize, and why.

Based on this, the following options are left for formulation of the hypotheses for model evaluation: similarity alone is not enough to justify extending a model, unless there are reasons for the particular choice for respects and degrees. If we want to extend the model, then we need some reasons grounding similarity judgements. Either this grounding is more clearly and accurately captured by adequacy-for-purpose assessments, or something has been overlooked in understanding how to characterize similarity.

For the time being, I will set aside the question of whether adequacy-for-purpose hypotheses are better than similarity hypotheses, as well as the question whether some notion of similarity must play a role in adequacy hypotheses. Instead, I will focus on further developing the similarity relation. Weisberg's account does include details about how modellers make decisions about what to include in their models, which might answer Parker's concern. However, his account lacks clarity in places, and needs to be further developed.

So what does the strongest characterization of the similarity relation look like? The following section provides an examination of the details of Weisberg's understanding of similarity. While Weisberg has provided extraordinary details on how to assess similarity, I argue first for a minor point: it may not be possible, nor in our interest, to reduce a model's similarity to the target system to a value between zero and one. My second, more significant criticism is that Weisberg has not provided sufficient detail about how a model's purpose affects the similarity relation. Such detail is needed, as the model's intended purpose directly constrains the weighted feature-matching equation, and thus overall evaluation of the similarity. Without sufficient detail as to how purpose constrains the similarity relation, his account is not complete. More specifically, I argue that Weisberg's view fails to account for the appropriate domain of application of a model. Without explicit specification of a model's domain of application, Weisberg's account fails to make important distinctions in modeling success that it should make. In §2.5, I introduce a historical example of modeling black holes to clearly illustrate this point. In light of my criticisms, I provide a stronger account of the similarity relation that I will use for the remainder of the dissertation.

2.4 Assessing Weisberg's Weighted Feature-matching Account

Weisberg understands the central model-world relation to be that of *similarity*. He takes similarity to be the best account of the model-world relationship because of its "flexibility to accommodate for the complexities related to the practice of modeling" while still having features that he identifies as being important for any model-world relation (2013, 135-7). In order to develop a strong account of similarity relations, Weisberg points out that he needs to be able to say what similarity supervenes on, how it depends on context, and how similarity judgments are to be evaluated. To this end, he provides his *weighted feature-matching account* of similarity. While the understanding of similarity offered by the weighted feature-matching account provides extensive detail as to how to understand and capture the similarity relation, it faces two important problems. First, reducing similarity to a quantitative score is at best not possible and at worst has bad epistemic consequences. Second, given the significant role a model's purpose plays in formulating the similarity relation, further details (which Weisberg has not provided) are required to have a clear understanding of how purpose affects the evaluation of this relation.

2.4.1 *The Weighted Feature-Matching Account of Similarity*

Recall that the weighted feature-matching account of similarity involves hypotheses of the following form:

> Model *m* is similar to target *t* for scientific purpose *p* to degree *S(m,t)*,

where *S(m,t)* is the weighted feature-matching equation. Weisberg's weighted feature-matching equation attempts to account for a model's similarity to a target system as a function of the features that the model and the target share, penalized by the features they do not share. He formally represents this as:

$$S(m,t) = \frac{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m)}{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) + \alpha f(M_a - T_a) + \beta f(M_m - T_m) + \gamma f(T_a - M_a) + \delta f(T_m - M_m)}$$

(Weisberg 2013, 148; forthcoming, 11).

To use this equation, a modeller must first determine their relevant feature set Δ, which is partitioned into two subsets, Δ_a and Δ_m. These are the *attributes* and *mechanisms* judged to be relevant to the goal of successfully modeling the target system. The equation is a normalized ratio of the features that the model and target system share (captured by the intersections of the model's attributes and target's attributes, and model's mechanisms and target's mechanisms), to the features that the model and target system do not share (denoted by the differences between the sets of features in the model, and the sets of features in the target system).

The model's scope[21], or inquiry goals, determine what the significant features in feature set Δ will be. Weisberg argues for a liberal account of what features can be included, ranging from qualitative, interpretive features (such as oscillations, or populations getting bigger or smaller), or strictly mathematical terms, to quantitative, physically interpreted terms (such as the existence of an equilibrium) (Weisberg forthcoming, 11). The feature set will contain statistical or dynamic properties, which include properties and patterns of behaviours, which he calls *attributes*. It also contains causal features, or the processes underlying and generating the attributes, which he calls *mechanisms*. These are designated with subscript *a* and *m*, and can be features of the model (M_a , M_m) or the target (T_a , T_m). The attributes and mechanisms in the feature set may include theoretical elements. That is, for some target systems, there is a large theoretical component to how it is characterized. In these cases, the similarity in the model partly depends on its similarity to theory. This is the sense in which models and theories interact on Weisberg's account[22].

The relations between these feature sets generate a real number via the *weighting function f*(•). This weighting function is not a fixed function, but rather determined by the context of the model. (I discuss how to determine the weighting function in more detail below). In fact, Weisberg says that a mature research program may have "a range of permissible weighting functions accepted by the community" (2013, 154). The weighting function's output is then multiplied by the coefficients, represented by θ, ρ, α, β, γ. These coefficients' values are based on decisions made by the modeller as to how important it is that the model has certain features, and for the model and target to have or share the attributes or mechanisms. This weighting is informed by background theory, goals, and

[21] As a reminder, scope is one of the four parts of the model's construal. For Weisberg the intended scope specifies which aspects of the potential target phenomena are intended to be represented in the model.

[22] A detailed example will be given in §2.5 to clearly explicate how this works.

pragmatics in modeling. In cases where there is a mature background theory, it will seriously constrain and inform the decisions made by the modellers with respect to this point. When modellers do not have a mature or well-supported background theory, it is more likely that there will be disagreement between the modellers with respect to determining just how similar the model should be to the target system. Through accounting for the feature set Δ, weighting function *f*(•), and term weights α, β, … , the equation will output a similarity score between 0 and 1 that can be used in comparative judgements of similarity.

Determining what to include as the feature set Δ, the weighting function *f*(•), and the values of the coefficients α, β, … , also depends on the scientific context, or what Weisberg calls the modeller's *construal.* For Weisberg, scope and assignment play a role in the construction of the weighted feature-matching equation, since they provide information on how the real world phenomena are to be represented in the model. Fidelity criteria are the standards modellers use to evaluate a model's ability to represent the phenomena. For now, I will set fidelity criteria aside and focus on how Weisberg understands scope and assignment.

For Weisberg, the scientific context or construal, is a way of formalizing the relevant scientific considerations at play in modeling and is flexible enough to account for all the various different scientific practices. However, he says very little to specifically explain how *scope***,** in practice, affects the development of a model. Weisberg provides only the following few passages to aid in determining how scope and assignment fit into these scientific contexts and judgements:

> The intended scope specifies which aspects of the potential target phenomena are intended to be represented by the model (2013, 40).
>
> When scientists choose a focus, or intended scope, they focus on some set of properties and abstract away the others. This yields a target system, a subset of the total state of the system (2013, 91).
>
> The modeler's intended scope takes into account the research question of interest, the context of research, and the community's prior practice (Kitcher, 1993). These elements of the modeler's scope, in turn, determine

> the contents of the feature set. So ultimately the choice of scope is equivalent to the choice of Δ (2013, 149).

It seems Weisberg is using "scope" as a very broad term to encompass and refer to anything that might be a consideration in the construction of the model. Scope plays a role in determining how we get from a general phenomenon to our target system by determining what, in a complex phenomenon, is of interest as a result of the research question. For example, if we are examining Tasmanian devil populations, we may want to learn about their population dynamics, or why the devil's immune system fails to recognize facial tumour disease (2013, 91). Depending on the research question, we will want to make sure certain aspects are accounted for in the model.

The second aspect of a modeller's construal, *assignment,* plays no role in determining what to include. Rather, assignment tells us how the relevant information has been included in the model. Assignments are explicit specifications of how parts of the target system are mapped onto parts of the model (2013, 39). For example, in the case of a concrete physical model of the Bohr atom, the small round balls in orbit around the nucleus are electrons. Alternatively, in a mathematical model the assignment specifies what the variables in the equation stand for with respect to the system. While assignment is often not made explicit during discussions of models, Weisberg argues assignment should be regarded as the formal record of this coordination.

These aspects of the modeller's construal are so important for Weisberg because, at the core, his understanding of the model-world relation is one of similarity. His account is an attempt to find a way to acknowledge that a model's construction and evaluation are very closely connected to what modellers take to be similar to their intended target:

> [The] similarity relation … already supervenes, in part, on the modeller's construal. When the context or scientific goals change, the construal will change, and aspects of the relation will change (2013, 149).
>
> Weighted feature-matching allows scientists to assess how close they have come to meeting their goals … Different goals can require different kinds of similarity relations, or at least the emphasis of different kinds of

> features. This is accounted for by the way in which the parameter values of each term of equation are set (2013, 150).

Establishing the similarity relation involves reasoning and justifying what parts need to be similar based on considerations from the relevant background theory, and from the norms and expected standards of the community of researchers. The weighted feature-matching equation formalizes how the modeller understands, relative to a model serving a certain purpose, how and why a model accurately represents the target system. Furthermore, it identifies how a model might need to be similar to a target system. This explains why the model is similar to the target system not by chance or by luck; for based on the target system and the purpose for which the model is being constructed, the model should include or not include certain features of the target system.

Thus far, I have explained how Weisberg's weighted feature-matching equation works, how we determine *S(m,t)*, and how model and target systems are accounted for in the *S(m,t)* equation. Returning to the hypothesis,

> Model *m* is similar to target *t* for scientific purpose *p* to degree *S(m,t),* where *S(m,t)* is,

$$S(m,t) = \frac{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m)}{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) + \alpha f(M_a - T_a) + \beta f(M_m - T_m) + \gamma f(\mathrm{T}_a - M_a) + \delta f(T_m - M_m)},$$

it is now clear that model *m* and target *t* feature in both the similarity equation and outside of it in the hypothesis statement. All that is left is to discuss how Weisberg understands scientific purpose *p*. To identify what he has in mind, I critically evaluate his account, addressing how Weisberg proposes we understand and account for purpose *p* and arguing that how he accounts for purpose is problematic. However, before turning to this issue, I offer a criticism related to computing the numerical similarity score.

2.4.2 *Criticism: Do We Want a Similarity Score?*

I think we should be skeptical of summarizing the complexity of the decisions modellers make with respect to constructing their models and the ways in which they understand their model to be similar to, and representative of, the real world in terms of a single value between 0 and 1. Can we really capture all the relevant features of similarity evaluation in a number and say one model has a similarity score of 0.789 while another scores 0.801, and have this comparison be meaningful? Is evaluating these elements of modeling with such a one-dimensional score actually worthwhile? The answer to both of these questions is no.

First, while in principle it may be possible to obtain such values, it is unclear that these numerical scores are obtainable in practice. Weisberg himself never actually computes a concrete similarity score for a model. What he does do, however, is discuss in detail how a modeller's desire for certain similarities will be important for the model to capture depending on the purpose. He demonstrates that when developing a model in order to say how a system might possibly work, there are certain features or properties (attributes and mechanisms) of the target system that one would expect a model to include. If the modeller thinks a certain background theory or dynamics, such as fluid dynamics, is critical for the target system, then the model must be similar in that it also relies on fluid dynamics. Modellers have reasons and justifications for including or not including certain aspects in the models they construct. These decisions are sometimes justified based on theory or on observations, or may be arbitrary[23]. What is valuable about the weighted feature-matching equation is that it formalizes and records these decisions.

Second, even if it is possible to obtain a similarity score, there is a danger that scientists could rely too heavily on these numbers and not pay attention to how the numerical score was calculated. Caring about the final output value from the equation erases all of the important similarity judgements made during the construction of the model. This risks negating the important work undertaken during the construction of the model to keep careful track of the ways in which a model is similar to a target system. If the focus shifts to prioritizing the number value rather than the information contained in the numerical value, modellers may lose sight of the importance of knowing how and why the model was created.

[23] For example, a modeller may introduce so-called "fudge" parameters that are exploratory in order to see if the model can be made to work.

For these reasons, the safer course of action is to take the weighted feature-matching equation as useful, but only as a means for explicit discussion and specification of what is prioritized and included in a particular model. The weighted feature-matching equation illuminates the pragmatic elements of modeling practice. Being explicit about these judgements helps in thinking through what is important and what is being prioritized (and to what degree) in a model. Such explicitness is particularly important when assessing the model and evaluating it compared to other models. Furthermore, if there are conflicts about the model, then modellers will be more readily able to evaluate differences and determine where modeling conflicts actually lie.

While computation of this score itself may not be useful, using the equation as a way to communicate what the modeller values and prioritizes *will* be useful. The process of undergoing the pragmatic exercise of enumerating the details that go into the modeller's judgements of similarity relative to the model's purpose is incredibly important. It explicitly identifies what is considered important for inclusion, to what degree, and why. The value of the weighted feature-matching equation is that we now have a guide to understanding why modellers evaluate a model as being similar to the target system for a specified purpose. It also allows for the evaluation of what has been identified as relevant and important for the modeller's assessment and understanding of similarity. Such explicitness is not just pragmatically useful but also important epistemically, as it accounts for how and to what extent a model can be considered to represent the target system. Again, this will be valuable—particularly when conflicts arise, in cases that are more complex, or in cases in which there is ambiguity in trade-offs between the weighting of a mechanism or attributes—for identifying what the modeller has chosen as important features to capture. The case study discussed below in §2.4 speaks to this point. In my modified version of the weighted feature-matching account, I take the *S(m,t)* equation as a means by which to communicate what the modeller thinks is important, prioritizes, and takes to be epistemically significant features of the target system. However, I refrain from computing a numerical similarity score.

2.4.3 *Criticism: What's the Purpose?*

My second, more substantive criticism of Weisberg's weighted feature-matching equation is related to the role purpose plays in the weighted feature-matching account of similarity. All models are constructed for some reason, for some *purpose*. If purpose affects how a model is constructed and evaluated (and both Weisberg and Parker take evaluation of their hypotheses to be made relative to a purpose), then it is critical to provide a careful understanding of what purpose amounts to. A survey of the modeling literature[24] indicates three main purposes or tasks for which we construct models: *describing* aspects of the target system, *predicting* aspects of the future[25], or *explaining* causal aspects of the target system. I will return to these three kinds of purpose in chapter 4, but for the time being will take these as the starting point in providing an understanding of how Weisberg accounts for purpose in his account.

Weisberg claims to have accounted explicitly for the model's purpose in his hypotheses, given that in order to understand claims such as "Model *m* is similar to target *t* for scientific purpose *p* to degree *S*(*m*, *t*)" we must understand the relationship between the scientific purpose *p* and the degree of similarity. For Weisberg, this is in fact the advantage of articulating hypotheses using the weighted feature-matching formalism.

> One advantage of articulating theoretical hypotheses using the weighted feature-matching formalism is that we can say a lot more about what is meant by scientific purpose *p*. Specifically, the weighting function *f*, as well as the relative weight (or inclusion at all) of the terms [in the weighted feature-matching equation], are contextual factors reflecting the scientific purpose (Forthcoming, 12-3).

Weisberg goes on to claim that different scientific purposes require the construction of different kinds of models, such as minimal models, how-possibly models, or hyper-accurate models (forthcoming, 12-4). He is attempting to provide a more fine-grained

[24] For example, these are the three purposes identified by Parker (2009). Weisberg also discusses model's purpose most frequently being to provide a prediction or explanation (2007, 2012, forthcoming).

[25] Or, retrodicting aspects of the past.

account of how these different purposes—describing, predicting, and explaining—play out.

In the case of a how-possibly model, the modeller will be interested in constructing a model that reproduces the target's static and dynamic properties. A good model in this context will be "one that shares many and doesn't lack too many of the target's static and dynamic properties" (forthcoming, 13). As a result, the modeller aims for the *S(m,t)* equation to have a high value with respect to the intersection of the attributes the model and target share, and a low value for what they do not share. Weisberg formulates the hypothesis for a how-possibly model as:

How-Possibly model *m* is similar to target *t* to degree $\frac{f(M_a \cap T_a)}{f(M_a \cap T_a) + f(M_a - T_a)}$.

While the aim is for the value of the weighted feature-matching equation component to equal 1, Weisberg acknowledges that this is often not achieved[26]. Nevertheless, the formulation allows for comparing models by allowing us to "see which features, among the ones that matter, are omitted by different models". In addition, "assuming a common feature set and weighting function, the models' relative deviation from the target can be assessed" (forthcoming, 13). The similarity between a model and its target system is understood in terms of the *S(m,t)* equation, and the *S(m,t)* equation depends on the weighting function *f*(•). The weighting function *f*(•) is partially determined by our scientific purpose and scientific context (or construal). The purpose is also encoded to some extent in the relative weight of the coefficient terms, θ, ρ, α, β, γ. It might seem that Weisberg has accounted for all relevant aspects of purpose in the *S(m,t)* equation, and that he has also accounted for purpose in the form of the similarity relation hypothesis.

Given these details about the role purpose plays in model similarity assessment, is Weisberg's weighted feature-matching equation effective? In the next section, I construct and examine a weighted feature-matching equation for the Oppenheimer-Snyder (O-S) black hole model. This example allows me to show more clearly how the weighted

[26] This is also an example of how Weisberg fails to actually provide an *S(m,t)* value between 0 and 1 but nonetheless provides an excellent account of what similarity amounts to in this context.

feature-matching similarity equation works. However, this example also illustrates a challenge that the weighted feature-matching equation and similarity relation hypothesis, as it is currently formulated, cannot solve. It highlights a case in which evaluation of a constructed model will have the same target system, same model, and same purpose, yet the model should obtain different similarity scores for the two different applications considered. This shows us that an important element has been left out of the similarity discussion thus far: domain of application.

2.5 Modeling Gravitational Collapse and Implosion of a Star

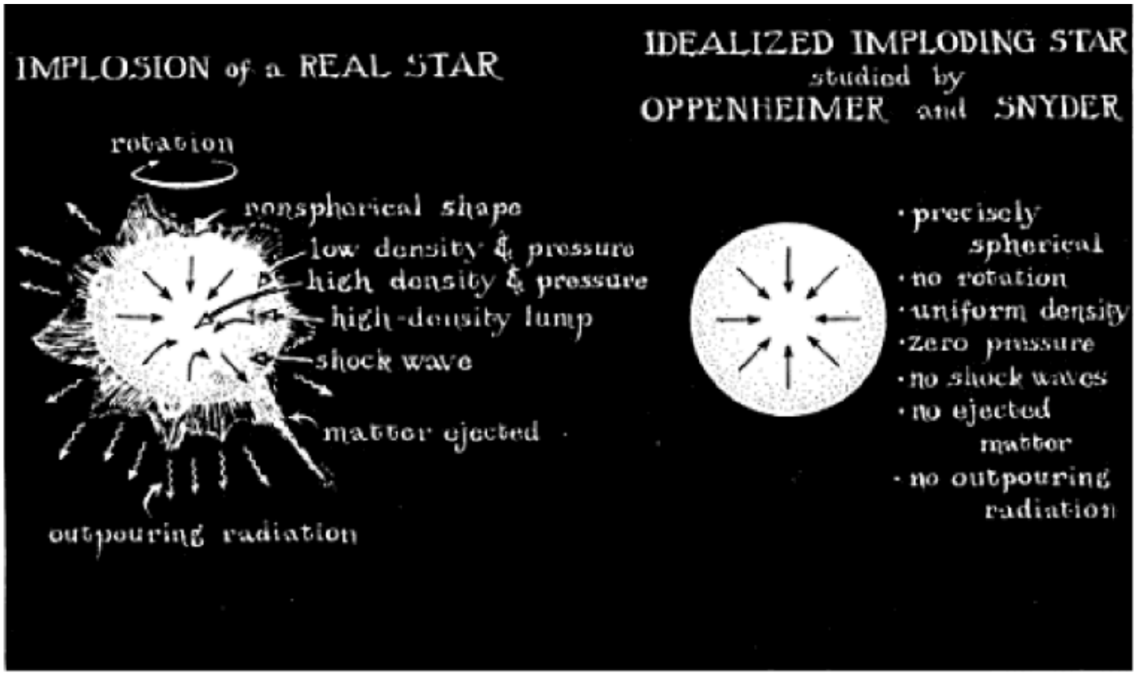
Early twentieth century astrophysicists were interested in what happens when massive stars run out of fusionable material. J. Robert Oppenheimer thought that they implode. However, investigation into what happens when a massive star implodes proved challenging. A real star rotates, giving it a non-spherical shape, high density, and pressure towards its center, and lower density and pressure further away from the center. When it implodes, it develops high-density lumps, as well as shock waves that may eject matter. There is also an outpouring of radiation (Thorne 1994, 215-7). In the 1930s, accounting for all of these features in computations would have been impossible. As a result, Oppenheimer and his student Harland Snyder constructed an idealized model of the imploding star in order to predict what might happen when the star implodes[27].

The most critical feature for Oppenheimer and Snyder (1939) was accounting for gravity as described by general relativity. Whichever way they chose to model the implosion, they determined this must be prioritized for inclusion. The star's spin and non-spherical shape, on the other hand, were ignored. For the target system of interest—massive stars that spin slowly—they considered rotation and shape not to have a strong effect (though for some other imploding stars, this might be crucially important). They also took the outpouring radiation, shock waves, and high-density lumps to be negligible. Lastly, since they had previously shown that gravity could overwhelm all pressure in

[27] Kip Thorne (1994) provides a detailed account of what Oppenheimer and Snyder prioritized in their idealized model, and this weighted feature-matching example is based on his account.

massive dead stars, they did not include thermal pressure, pressure arising from electron or neutron degeneracy, or nuclear force. While they thought the details of what happens in a real star during the implosion might differ slightly, they took it that the differences would not have a great enough effect on the outcome to require capture in their development of a mathematical model.

Figure 1: Sketch of Imploding Star.



Sketch of a real imploding star and relevant attributes and mechanisms (left), as compared to Oppenheimer and Snyder idealized imploding star and its relevant attributes and mechanisms (right). Image from Thorne 1994, 217; 454.

With these idealizations in place, Snyder worked out equations governing the entire implosion. From these equations one could read off aspects of how general relativity says stellar implosion would behave, as seen from the inside, outside, or surface of the star. These equations were the first to predict that for a static, external reference frame, as the star gets smaller, it implodes more slowly, until the point it becomes "frozen" at a critical circumference (now known as the event horizon). However, for an observer on the surface of the star, the implosion continues rapidly past that freezing point until "crunched" to infinite density and zero volume. These "frozen stars" are what we now refer to as black holes. This mathematical model predicted that when a sufficiently large mass star dies, it must implode to form a black hole. At the time, the model predicted what was considered a very bizarre outcome, and there was no way to test it experimentally. It was not until the late 1950s that mathematical computer simulations of imploding stars also produced results in favour of Oppenheimer and Snyder's claims (Thorne 1994, 218).

2.5.1 *Weighted feature-matching equation for the Oppenheimer-Snyder Black Hole Model*

With this background, I am in a position to construct a weighted feature-matching equation, *S(m,t),* as well as a similarity relation hypothesis for evaluation for the Oppenheimer-Snyder (O-S) black hole model. The target system under investigation is a massive imploding star. As mentioned, the purpose of the model is providing a prediction of the behaviour of these sorts of stars. With this target system and purpose in mind, the O-S model was constructed. Because every model is evaluable using *S(m,t)*, I will now explicate the similarity between the target system and the model using weighted feature-matching.

Both the choice of mechanisms and attributes will be informed by the scope component of the model's construal. Recall, "scope" refers to the aspects of the target that are intended to be represented by the model. The features of the target system that are intended to be represented in the O-S model example are the star's mass and gravitational behaviour. The O-S model does not seek to represent features such as the star's non-spherical shape or the star's rotation. The assignments are explicit specifications of how parts of the target system are to be mapped onto parts of the model (Weisberg 2013, 39). In the case of the O-S black hole model, this involves specifying what in the mathematical model is intended to represent the aspects of our target system. As Weisberg points out, modellers often do not make these assignments explicitly. However, the assignment should be regarded as the formal record of this type of coordination.

With the scope and assignment acknowledged, the next step is to identify what attributes and mechanisms are to be included in the model. The attributes are the static and dynamic properties, and so in the target system this includes the star's density, pressure, charge, spin, and mass. However, for the purpose of providing a prediction, the model did not need to be similar with respect to all of these attributes. In fact, the point was to make the idealization that the star is perfectly spherical and has uniform density. Priority in constructing the model is given to the mass of the star being large enough, so this will be included as an attribute in *S(m,t)*. We need not, however, be concerned with the model having similarity with respect to accurately representing all of these attributes of the target system.

The mechanisms are the processes underlying and generating the attributes, the causal features, M_m and T_m. For the purpose of providing a prediction in the context of general relativity, whatever happens with respect to attributes, the model must have the underlying mechanism of gravity as described by general relativity[28]. Oppenheimer and Snyder were not concerned with any other mechanisms[29], such as possible quantum mechanical effects.

Now that I have the relevant mechanisms and attributes identified, I can establish the weighting function and the values of the coefficients. Recall that for Weisberg, weighting is informed by our background theory, goals, and pragmatic considerations. Oppenheimer and Snyder's purpose in constructing their mathematical model was predicting how a massive star would implode and how its implosion would appear to various observers. They wished to explore what the predictions of general relativity would be in these astronomical cases. Oppenheimer and Snyder identified this as one of the significant items for our feature set **Δ**, and therefore believed that it should receive the greatest weight. Their justification for this was in some sense an appeal to background

[28] For Weisberg, the dynamical considerations in a model are captured in its mechanisms. One intuitive way to think about how this applies to general relativity is to consider the "shape" of spacetime to be the cause of gravitational effects. However, this is not essential, and should not be seen as an endorsement of a substantivalist view of spacetime, or commitment to the idea that spacetime has causal powers. General relativity as the "mechanism" for gravity could be redescribed as an attribute.

[29] Recall that in the context of the weighted feature-matching equation, we are to understand "mechanism" in a loose sense.

theory. The current best picture of our physical laws indicated that general relativity most likely governs the physics of how these stars implode. In other words, they wanted to see how general relativity would predict an imploding star would behave. Therefore, we must weight this mechanism heavily in the *S(m,t)* equation.

With this information in place, we can sketch the weighted feature-matching equation, and determine the weighting of the mechanisms and attributes that comprise the feature set. High weighting should be given to the model sharing the mechanism of spacetime and gravity described by general relativity $f(M_m \cap T_m)$. There is high weighting for a model sharing the correct attributes of the mass of the star in the target system $f(M_a \cap T_a)$. We give a very low weighting for the model being penalized for not sharing any other attributes such as the spin, charge, or density of the star $f(M_a - T_a)$. The low penalty is a result of the desire for the model to make idealizations about spin, charge, and density for the sake of computation. Of course, this low weighting is also justified by the significant amount of background theory and dynamical arguments that these effects are negligible.

There is also only a low penalty for the model not sharing other mechanisms, such as inclusion of other probable theoretic elements, like quantum effects, that might be at play in the target system $f(M_m - T_m)$. Given that this model is intended to be heavily idealized, there must be a heavy penalty if the model includes attributes or mechanisms that the target system does not share $f(M_m - T_m)$ or $f(M_a - T_a)$. In other cases, models will be assessed using different coefficient values, which determine different penalties for dissimilarity. But given that the model is constructed almost entirely from idealizations and background theory, these are the appropriate values.

It is important to bear in mind that these decisions related to the penalties are not arbitrary or left up to the modeller's whim. When evaluating what features are relevant given a particular modeling goal, theory plays an important role. In cases in which there is strong background theory to draw on, what the model will be penalized for including or not including is not a free decision on the part of the modeller. What is important to include is determined by the theory. However, there are instances in modeling where there is not much, or any, background theory. Models that have less background theory and are constructed almost entirely from data will have different priorities and weighting. However, by requiring explicit formulation of the weighted feature-matching equation, the justification for these decisions is more transparent. In cases in which there is not a strong commitment to a particular background theory, it is even more important to pay attention to why a modeller might include certain features as necessary (and thus the modeller might judge a model to be more or less similar).

Taken together, this information comprises the content of a weighted feature-matching equation, which can then be joined with our other elements to compose the following hypothesis statement:

> The Oppenheimer-Snyder model is similar to the target system, a massive imploding star, for the purpose of predicting behavior of the system, to similarity degree *S(m,t)*, where *S(m,t)* =
>
> $$\frac{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m)}{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) + \alpha f(M_a - T_a) + \beta f(M_m - T_m) + \gamma f(T_a - M_a) + \delta f(T_m - M_m)}$$
>
> where θ, ρ, γ, δ >> α, β.

However, this does not yet fix a value of *S(m,t)*, as we have not determined the weighting function *f*(•).

As I argued in the previous section, it is clear that the weighted feature-matching account of similarity is a useful tool for capturing the details of the model construction process, particularly the judgements and justifications of what to prioritize for inclusion, and why. However, it is not clear what additional value is gained by calculating the similarity score. In fact, most of the interesting, valuable information has been detailed, and calculating a numerical value may lead to ignoring that information. To this extent, my first criticism of the weighted feature-matching account is justified: Although the weighted feature-matching equation is an extremely useful tool for forcing explicit discussion, recording the relevant considerations that went into the judgements made by the modellers during the construction of the black hole model, and showing how the construction of the O-S model was justified as being similar to the target system given its intended purpose, providing a numerical score does not offer anything new, anything that has not already been captured by the analysis above.

2.5.2 *Same Model, Target, and Purpose: Different Weighting*

The Oppenheimer-Snyder model provided an account of the gravitational collapse of a star. The model contained features that were weighted highly and did not include extra features that were not needed. Overall, the O-S model can be evaluated as having a high degree of similarity with respect to the target system for the purpose of predicting the behaviour of the system. The Oppenheimer-Snyder model is simple enough to be able to compute predictions about what happens to the spacetime both on and outside the star, and for this purpose it was not necessary to include other features in the feature set.

Table 1: Weighted Features of the Oppenheimer-Snyder Model.

Weighted Features of the Oppenheimer-Snyder Model			
Feature Set Δ	**Weight**	**Model Feature**	**Target Feature**
Features Shared in Feature Set			
Mechanism: Gravity	High	GR Spacetime	GR Spacetime
Attribute: mass	High	At least 3 $M_{\odot}$	At least 3 $M_{\odot}$
Features Not Shared in Feature Set			
Attribute: non-spherical	Low	spherical	non-spherical
Attribute: angular momentum (spin)	Low	No	yes
Attribute: density	Low	Uniform	Center: high & lumpy Outer: low & lumpy
Attribute: shock waves	Low	No	yes
Attribute: outpouring radiation	Low	No	yes

This table provides a summary of the mechanisms and attributes of the model and target system, and their relative weightings in our formation of *S(m,t)*.

However, the community of researchers interested in black holes soon wanted to investigate what would happen to these stars over a longer timeframe and to make predictions about the system on the timescale of billions of years. Their purpose remained the same—modeling what would happen to their target system of an imploding star for the purpose of predicting how this system would behave. Their target system also stayed the same, a massive imploding star. Yet a correct analysis would should now evaluate the O-S model as having a low similarity with the target system. What has changed?

When the similarity of the O-S model to the target system is assessed relative to a longer time scales, the O-S model should receive a lower similarity score. While spin and charge are not relevant features for the model on short timescales, they may be needed when we are interested in longer timescales[30]. But perhaps most importantly, on longer timescales, quantum vacuum fluctuations have an important effect as a mechanism in the target system, leading to black hole evaporation (Hawking 1974). When examining longer timescales, these features start to have an impact and thus are necessary inclusions for a mathematical black hole model. They require a heavy weighting as attributes in the feature set[31]. The model should be penalized for not sharing this attribute with the target system. The modeller is dealing with the same target system, a massive imploding star; so the model has the same feature set, same attributes, and the same mechanisms. It even has the same purpose: predicting the behaviour of the target system. As such, nothing in these three elements (target system, model, and purpose) permits us to change our weighting in light of the model missing the spin and charge attributes and quantum fluctuation as a mechanism.

The motivating change was a desire to be able to predict and describe what happens to the imploding star over a longer timescale. In order to see what happens over a longer timescale, we need to change the weighting of some of these attributes and mechanisms that compose the feature set. We need to see a change in the weighting function $f(\bullet)$, or the coefficients[32]. But since, according to Weisberg, the similarity

[30] Hawking radiation depends only on the mass, angular momentum, and charge of the black hole. However, spin and charge need not be accounted for in the model. Page (1976) calculated the power produced and the time to evaporation for a nonrotating, non-charged Schwarzschild black hole.

[31] Accounting for spin and charge in the mathematical model for black holes is what led to the prediction that over extremely long periods of time, black holes undergo black hole evaporation, also known as Hawking radiation.

[32] If Weisberg allows us to change $f(\bullet)$ without changing the purpose, it is hard to see how we could compare the similarity of different models that were developed for the same purpose since their similarity score could depend on different $f(\bullet)$.

relation supervenes on the construal, and on the purpose, changing the elements of the equation must reflect a change of purpose. Yet in this case we are not changing our purpose, as we are understanding purpose to be about providing a prediction. What is changing is another element that is not explicitly included in Weisberg's account—the domain of application.

According to Weisberg, the possible purposes of models are to describe, to predict, or to explain. However, I believe there is an additional component that must be accounted for in a discussion of purpose: the domain of the model's application. By domain of application I mean a scale at which the model is intended to apply—for example, over a particular time or distance scale. Every model makes certain approximations and idealizations that make it the case that its predictions, explanations, or descriptions are only accurate relative to a certain time, distance, or size scale. For this reason, I argue that each model has an implicit time, distance, or size relevancy that makes it potentially inappropriate to apply to other scales. This is the domain relativity of the model. It may not be obvious that this is something important to include, or even that it is different from some of the elements that have already been discussed. However, we want to say how constructing a new, different hypothesis statement for the same model—representing the same target system and constructed for the same purpose—must acknowledge and account for a shift in the intended domain of application.

In the case of the O-S model, on Weisberg's understanding, we have the same target system and the same purpose. However, because we want to look at more than the initial collapse, our domain of application needs to inform and provide us with a different weighting to both $f(\bullet)$ and the coefficient values in the construction of the model. However, Weisberg does not seem to allow for this kind of flexibility in his account. For example, he says:

> For a given target and scientific purpose, the equation lets us evaluate the relative similarity of a number of models by scoring them. Moreover, when multiple plausible models have been proposed, this expression helps us isolate exactly where they differ (Forthcoming, 12).

This indicates that when one fixes the purpose and target system, that fixes the weighting function and coefficient values in *S(m,t)*. The resources that Weisberg uses to characterize purpose have not included a distinction between different domains of application. As currently formulated, Weisberg's account does account for the domain of application as a factor that can change our weighting.

The O-S model's intended purpose is to predict the behaviour of the target system. In the first application, we only want to predict what happens when the star collapses. In the second application, the question under investigation is what happens when to the collapsed star over a much longer period of time. In the second application we still have the same model, the same purpose, and the same target system. But there is some reason the model is not going to be similar enough to the target system in the second application. This reason is related to the change of the timescale, not a change in purpose. It is only over a longer time scale that you need to account for aspects such as spin, charge, and quantum fluctuations. However, in the way Weisberg has formulated how to determine the weighting, it is the purpose alone that determines what goes into the model, without any mention of purpose being relative to a domain. He has not explicitly allowed for a change in a domain of application to change the weighting.

One possible way Weisberg might attempt to solve this problem is by building the domain of application into the purpose (e.g. the purpose of predicting the behavior of the black hole over the scale of millions of years). While this strategy might be successful, Weisberg has discussed purpose simply as it relates to the action of describing, predicting or explaining, and not in a more fine-grained notion. Purpose is the role the model is intended to serve, but one should also explicitly acknowledge the *bounds* at which the model successfully aims to serve that purpose. As I propose below, such a qualifying modifier should be included explicitly to supplement the purpose as characterized by Weisberg. Furthermore, I advocate for the explicit inclusion of domain of application in the hypothesis because its effect on the weighting function *f*(•) is comparable to that of purpose. Therefore, it is just as important to make it a component of the hypothesis explicitly.

Another possible reply is that Weisberg has accounted for domain of application in his understanding of scope, which informs the modeller which aspects of the potential target phenomena are intended to be represented by the model, or even in the definition of the target system itself. If this is the case, Weisberg has given little in the way of textual evidence that this is his intention. Furthermore, given the points I have raised through my case study, the domain of application has not been accounted for in purpose either; yet domain needs to be accounted for explicitly. Either Weisberg has omitted the domain of application from his discussion of purpose and scope, or it is not clear that this is something he has considered and included in his account.

However, I do not think my critique here is detrimental to Weisberg's account. The similarity account he offers certainly will allow for the domain of application to be part of a modeller's intended scope. Yet there must be a way for the domain of application aspect to feature in the equation. As currently formulated, scope alone allows for such information to bear on the equation. If domain of application should permit us to change the weighting of the similarity relation, then this it must be accounted for explicitly.

An improved account must explicitly include the importance of the domain of application. I propose that this is done through inclusion explicitly in the form of the similarity relation hypothesis:

> Model *M* is similar to target *T* for scientific purpose *P* over the domain of applicability *d* to degree *S(m,t)*.

Domain of application should be seen as the final element that will specify the components of *S(m,t)*. Providing a different *S(m,t)* equation (i.e. different *f*(•), and coefficient values) can be done simply by considering a different domain of application. This modification to Weisberg's formulation of the hypothesis addresses the problems raised by the black hole case study. This addition makes my formulation the strongest candidate for an understanding of model evaluation at the construction stage as an assessment of similarity relations. A more complete hypothesis will pay explicit attention to the impact domain of application has on similarity judgments. The means by which we evaluate models is specific to domains, and modifying the similarity relation to capture these complexities is going to allow for modellers to have a better tool to use, and a better sense of how their model's construction has been justified.

2.6 Conclusion.

With these modifications in mind—using the weighted feature-matching equation only pragmatically and including the domain of application explicitly in the hypothesis statement—I have provided the strongest version of the similarity relation for model evaluation at the construction stage. A similarity relation hypothesis is an extremely informative and effective way to formulate a hypothesis statement. It forces us to enumerate the elements of the target system we have chosen to include in the model explicitly. It also requires the modeller to provide a justification of what has been included in the construction of a model, and of the way our evaluation of the construction is relativized to a purpose within a certain domain of application. All modellers must have these bounds defined in order to evaluate their models with respect to their intended purpose and to guide further application of a model beyond the original purpose.

In the next chapter I return to evaluating the two possible formulations of the hypotheses—as adequacy for purpose hypothesis statements and the modified similarity relation hypothesis statements—in order to determine if one of these forms for hypotheses should be preferred over the other. I will argue that they are both essential elements of model evaluation but must be understood as evaluations of different components of a model.

Chapter 3

3 Evaluating Hypotheses about Models

3.1 Introduction

Understanding model fit requires answering three questions: What is the target of evaluation? What form should the hypothesis statements take? And finally, how should the hypothesis be evaluated? In the previous chapter, I established that model evaluation is not about evaluating the truth of the model itself, but rather evaluating a hypothesis about the model's effectiveness or utility. I introduce two ways in which to form hypotheses about models. One focuses on evaluative standards concerning whether a model is *adequate* for a certain purpose; the other assesses if the model is *similar* enough for the purpose. I have argued that Weisberg's similarity relation hypothesis, "model *m* is similar to target *t* for scientific purpose *p* to degree *S(m,t)*", must be modified to account for domain of application, and I have offered the following reformulation: "model *m* is similar to target *t* for scientific purpose *p* over domain of application *d* to degree *S(m,t)*". I have also argued that the weighted feature-matching equation, *S(m,t)*, is an extremely useful tool for explicitly capturing the modeller's justifications of the relation between the target system and model, although its utility does not stem from calculating a numerical 0 to 1 score.

In this chapter, I will return to the question of whether the adequacy-for-purpose hypotheses should be favored over the similarity-relation hypotheses as best capturing evaluation of model fit in scientific practice, as well as the question of how model fit should be evaluated through such hypotheses. At first, it may seem that the two possible forms for the hypotheses are both evaluating the same thing about the model. By examining how Parker and Weisberg respectively propose to evaluate their hypotheses, however I argue that each hypothesis actually has a different target of evaluation.

This argument proceeds as follows: Weisberg's proposed evaluation of model fit involves two types of fidelity criteria—dynamical and representational. I argue, however, that there is a conceptual distinction between the evaluation done by assignment and scope on the one hand, and the work done by the fidelity criteria on the other. Assignment and scope provide information about how the model *represents* the target, whereas fidelity criteria evaluate how closely the output of the model must *fit* the real world phenomena in order to be considered an *adequate representation*. Given that the fidelity criteria are concerned with evaluating the adequacy of the model, there may be reason to think that the fidelity criteria are actually assessing adequacy in Parker's adequacy-for-purpose sense. I argue this is not the case either, given that Parker is not evaluating adequate representations, but rather whether a model is an adequate tool, effective for the purpose to which it is put.

In light of this analysis, I argue that *aspects from both similarity and adequacy need to feature in an account of evaluation of model fit*. The two hypotheses work together in the following way: An assessment of a similarity-relation hypothesis is involved when evaluating the relation between the model and the target system during the model's construction. For this component of model evaluation, we should employ my modified similarity-relation hypotheses (from Chapter 2). Evaluation of the model's adequacy for purpose is about evaluating the output of a model and comparing it to the equivalent output phenomena of the real world. These two aspects are different in that an assessment of adequacy is concerned with evaluating what the model *does* and its effectiveness for that aim, while assessment of similarity is concerned with evaluating how the model *represents*. I argue that *it is only when we take both hypothesis statements together, that we evaluate the overall fit of the model.*

I propose a framework in which assessment of model fit is understood through four components. The first component involves constructing the model and establishing the similarity relation via the weighted feature-matching equation. The second component involves obtaining through reasoning or calculation an output from the model. This component also involves determining what would be observed as the output in a certain test situation if the model is effective or adequate for the purpose. The third component involves comparing and evaluating the level of agreement of the model's output with the analogous output from the target system. The fourth component involves an assessment of the model's overall fit through a final evaluation of our two hypotheses.

The assessment of the adequacy-for-purpose hypothesis addresses whether the model is qualitatively or potentially quantitatively satisfactory for the purpose at hand. The similarity relation hypothesis statement addresses the standards by which the model is assessed to be similar to, and representative of, the target system for the given purpose.

3.2 Methods of Evaluating the Hypothesis

The two kinds of hypotheses under consideration are the adequacy-for-purpose hypotheses and my modified similarity-relation hypotheses, which take the following forms:

> **Modified Similarity-Relation Hypotheses**: Model *M* is similar to target *T* for scientific purpose *P* over the domain of applicability *d* to degree *S(m,t)* (where *S(m,t)* is the weighted feature-matching equation).
>
> **Adequacy-for-Purpose Hypotheses**: Model *M* is adequate for intended purpose *P*.

At first glance, it might seem that these two hypotheses intend to evaluate the same thing—whether the model is good enough for its intended purpose. That is, while the hypotheses have different forms, it might be the case that, in examining how model fit is understood, what is being evaluated, and how that evaluation proceeds, it turns out that the two different hypotheses actually have the same target of evaluation. In this section, I will examine how the hypotheses are to be evaluated. Parker and Weisberg have each provided accounts of how to evaluate these hypotheses, which I will use as the basis for my analysis. As mentioned in Chapter 1, Parker and Weisberg have thus far provided the most detailed discussion of these two approaches to assessing model evaluation, which makes their positions the natural starting point for my discussion. In §3.3, I examine what the differences might be in their evaluation processes and argue that the two hypotheses are not evaluating the same aspects of a model and have different targets of evaluation. In light of this, I argue that rather than thinking one hypothesis is better than the other, there is plausible reason to think both are useful as components in understanding the success of models. The similarity hypothesis is about the model's representational success, while the adequacy-for-purpose hypothesis is about the model's output and assessing its usefulness as a tool. These components of assessment are related but conceptually distinct. Their distinctness follows from the fact that they have different targets of evaluation[33].

3.2.1 *Evaluating Similarity-Relation Hypotheses*

While I have proposed a modified form for the similarity-relation hypotheses, I will take Weisberg's proposed methods for evaluations of his version of the similarity-relation hypothesis as the basis for my account of evaluation[34]. For Weisberg, evaluation of the similarity-relation hypothesis is about assessing the "goodness of fit" between the model and the target system. The aspects of this evaluation are captured through the model's construal—the relevant intentions of the modeller. Recall, the construal of a model is composed of four parts: assignment, scope, and two kinds of fidelity criteria. Assignment and scope track how the real-world phenomena are intended to be represented in the model. The fidelity criteria provide the standards modellers use to evaluate a model's ability to represent the phenomena (Weisberg 2013, 39; 2007, 123). On this view, similarity assessment is a central component to fit. For a model to fit, and therefore be successful, it must be grounded in the similarity relation.

It is important to note that the notion of similarity employed in Weisberg's account is not a strict sense of similarity—it is not to be understood as a one-to-one mapping or as requiring that all features or relations of the target system must be preserved or "fit" with the model. Rather, the fidelity criteria provide the acceptable standards:

> … model-target fits do not necessarily put equal weight on all aspects of the model and target, nor are they uniform in the degree of fit that must be established between each property of the model and of the target. The modeler's fidelity criteria will specify which properties must fit, and to what degree they might fit (2013, 93).

[33] It is true that a model that is successfully similar for a purpose is more likely to be adequate for that purpose. However, this connection is not necessary and needs to be assessed in each case.

[34] The changes that I proposed are relevant only in that they add to the specification in the construction of the model (via domain of application) and exclude evaluation of a numerical score.

While fidelity criteria play a central role for evaluation of model fit, Weisberg unfortunately provides little specific detail about the fidelity criteria. What he does say, however, is that they are the standards for evaluating a model's ability to represent phenomena (2007, 219; 2013, 39), in that they tell us "how similar the model must be to the world in order for it to be an adequate representation" (2013, 41; 2007, 221).

He also identifies two types of fidelity criteria—dynamical and representational. Dynamical fidelity criteria focus only on evaluating the output of the model and dictate how close the output of the model must be to the output of the real world phenomena (2007, 221; 2013, 41). One way in which these criteria are specified is as error tolerances; for example, the output of the model must be within ±5% (2013, 41; 2007, 221). While dynamical fidelity criteria determine if the model's outputs are close enough (for example, if the model is making the right predictions), representational fidelity criteria "go beyond this" and "assess whether the output is being provided for the right reasons, e.g., predictions are made for the right reason" (2007, 221; 2013, 41). While Weisberg calls this "representational fidelity", it really has more to do with grounding the output of the model in the representation that has been built into the model via the similarity relation. These "right reasons" for the model producing its output must be because the model accurately represents the target system in certain ways. This relation is captured in the weighted feature-matching equation, *S(m,t)*. What is being assessed by this fidelity criterion is whether the model's representation of the target system continues to be adequate when the output of the model is considered.

It is important to notice that there is a conceptual distinction between the work done by assignment and scope, and the work done by the fidelity criteria. Assignment and scope provide information about how the target system is intended to be *represented* in the model. Fidelity criteria do importantly different work; they evaluate how closely the output of the model must *fit* the real world phenomena. It is only after examining the model's output that one can determine if the modeller's decisions about what to include (or not include) in the model were sufficient.

Drawing this distinction between assignment and scope, on the one hand, and the fidelity criteria on the other much more sharply than Weisberg does also brings out two features: First, this distinction indicates that there are two different components to the evaluation of a model. Assignment and scope should be thought of as doing some evaluative work, as they indicate what aspects of the model are intended to be representative of aspects of the target system. However, this evaluation is different from the evaluation connected to the fidelity criteria. Assignment and scope assess the model's similarity relation relative to the target systems during construction of the model and establish what the modeller considers to be important properties of the target system that the model should (or should not) have. The fidelity criteria assess the success of the model, but only after we obtain an output from it.

Second, this distinction also provides a clearer understanding of what similarity means in an evaluation of the similarity relation. It is grounded in reasoning about the properties of the system and determining which of those features are to be included or not. This claim is not about similarity in a strong sense. Rather, it is about a modest similarity focused on identifying what properties and features of the target system stand in relation with each other for the intended purpose of the model, and if the model does indeed have those properties. A modest similarity relation grounds the reasoning about the properties or features that are in the model. What the similarity relation specifies is that the model has properties that make it suitable for the purpose for which the model is constructed. For example, if the model was constructed to serve the purpose of predicting the way tree leaves move in the wind, then, even before assessing how well the model performs, we need to establish that something in the model stands in a similarity relation to the leaves and the wind. The weighted feature-matching equation captures these decisions. Should a model be evaluated as not doing a good enough job at predicting the way in which leaves move in the wind, the modeller can determine if this is a result of the model failing to have a feature that is similar enough to the relevant features of the target system.

The target of evaluation in similarity relation hypotheses, then, is an evaluation of the ways in which the model is similar to the target system at the construction stage through establishing *S(m,t)*, as well as an evaluation of whether an appropriate level of similarity is maintained once an output from the model is obtained. A modeller first characterizes and determines what are the important features (attributes and mechanisms) of the target system to include in the model's representation. These features are established in the similarity equation through our weighting of the features we have evaluated to be important and the degree to which their presence, or absence in the model should be weighted or penalized. This evaluation takes place in the process of constructing our weighted feature-matching similarity equation. At this point, the modeller establishes how the model represents or *fits* with the target system. This aspect of fit provides epistemic justification for the model having properties that make it suitable for use. Only after we have constructed this equation do we turn to evaluating if the established degree of similarity is similar *enough* with the aid of our fidelity criteria—they fix the level of accuracy demanded.

Recall that I have argued that we should understand the fidelity criteria as identifying a very different component of the overall assessment of the model. Rather than assessing the ways in which the model is *similar* to the target system, the fidelity criteria are concerned with evaluating how closely the model output must match the world in order to be considered an *adequate representation.* Fidelity criteria assess the success of the model, but only after we obtain an output from it. What is unclear is whether fidelity criteria are best understood as assessing similarity in Weisberg's sense, or assessing adequacy for purpose in Parker's sense. I want to set this question aside for a moment, and instead examine how Parker characterizes the target of evaluation of model fit as a comparative assessment of the output of a model relative to the equivalent output phenomena of the real world.

3.2.2 *Evaluating Adequacy-for-Purpose Hypotheses*

To assess if a model is adequate for its purpose, Parker proposes the following means of evaluation:

> In order to argue that we have confirmed or disconfirmed such an adequacy hypothesis, we will need to (i) determine what we are likely to observe if it is true that the model is adequate for the purpose(s) of interest

> and then (ii) check how well what is actually observed fits with what we are likely to observe if the model is adequate. If what is actually observed fits well enough, then the observation confirms the hypothesis that the model is adequate for the purpose(s) of interest (2009, 237; see also discussion from Parker 2010, 7-8).

With respect to (i), Parker argues that determining what should be observed in a chosen test situation, assuming the model is indeed adequate, is not simple. Providing an argument for what should be observed if the model is adequate is relatively straightforward in cases in which the model output can be compared directly to the equivalent output data. That is to say, it is relatively easy to check that a model is able to reproduce the data that have already been obtained. However, it is much harder, or even impossible, to provide an argument about what should be observed if the model is to serve an explanatory or predictive role. This is due to there being no simple, general principle that can be applied to help identify what one is likely to observe if the model is adequate for the purpose of explaining and predicting (2009, 242). Since a modeller would not know what to expect as an output from the model, that would undermine their ability to assess the model and thus have confidence in its outputs. In the case of simulating, or accounting for data already obtained, the modeller already knows what they are likely to observe—the model will be able to reproduce the data. However, if the purpose of the model is to serve an explanatory or predictive role, they often do not know what they are likely to observe.

Parker notes that the evaluation of the hypothesis may be quite challenging; without being able to assess the first part (i), we cannot move on to the second part of comparing what is expected to what is actually observed (ii). That is to say, if it cannot be determined what to expect as an adequate output provided by the model, then one does not know what to be looking for in actual observations to compare the model's output against. Adequacy is about selecting a model with properties that are suitable for the tasks at hand.

The target of evaluation, then, is a comparative assessment of the output of a model to the equivalent output phenomena of the real world in order to assess if it is suitable for a particular purpose. This evaluation, however, is extremely difficult due to the fact that little else follows from a successful model about either the target system or model, leaving it unclear what the modeller should find in various test situations if a model is adequate for a purpose, unless they have additional information about the model and the target system. Parker notes that in such situations, “Perhaps the best that scientists can do is to draw on what led them to think that the model might be adequate in the first place” (Parker 2010, 9). This theme will be taken up in chapter 4 in discussing what justifies inferences about a given model being applied in novel situations. For now, I will take this account of evaluating adequacy and compare it to the fidelity criteria involved in assessments of similarity.

3.3 Why Fidelity Criteria and Assessment of Adequacy for Purpose are Different

Adopting the means of evaluation of adequacy-for-purpose hypotheses from Parker, the evaluation has two parts,

i. determine what we are likely to observe if it is true that the model is adequate for the purpose(s) of interest, and then

ii. check how well what is actually observed fits with what we are likely to observe if the model is adequate.

Returning to my earlier question, do fidelity criteria evaluate the same aspect as adequacy for purpose? If so, it might be the case that similarity and adequacy have the same target of evaluation. If not, it may be that there is more to model evaluation than just assessment of adequacy. I will address this issue, one kind of fidelity criterion at a time.

It might seem as though the dynamical fidelity criteria are attempting to formulate something similar to part (i), in which the goal is to determine what the model would predict if it were adequate. However, the dynamical fidelity criteria are used to evaluate how close the output of the model must be to the output of the real world phenomenon. This makes it seem as though the dynamical fidelity criteria are intended to make a comparison similar to part (ii), in which we compare the model’s output to the actual observations of the real world.

Should the dynamical fidelity criteria be understood as Parker's (i), (ii), or as something else? I do not think the intentions behind the dynamical fidelity criteria are to be understood as the same as step (i). And Parker's step (ii) is not exactly the evaluation that these fidelity criteria intend to capture either. Given that both kinds of fidelity criteria are used in conjunction with a similarity relation, it follows that the dynamical fidelity criteria are not directly comparable to these two steps. Rather, the dynamical fidelity criteria, given their role as part of a modeller's construal, are judgements about the degree to which the modeller would consider there to be a close enough similarity between the model and the target. The only difference is that the fidelity criteria evaluate the similarity of a model output and analogous target system output, i.e., the comparable output in the real world. Weisberg says that this judgement is similar to that of error tolerances. The evaluation connected to the dynamical fidelity criteria assess to what extent the model's output can be different from our real world observations. This analysis then further informs, and supports the assessment of the ways in which the model is similar to the target system.

If the dynamical fidelity criteria are used to evaluate what is acceptable similarity between a model's output and the real world, perhaps the representational fidelity criteria are comparable to the adequacy-for-purpose evaluation components (i) and (ii). This is not the case either. The representational fidelity criteria "give us standards for evaluating how well the structure of the model maps onto the target system of interest" (Weisberg 2013, 41) and "specify how closely the model's internal structure must match the causal structure of the real-world phenomenon to be considered an adequate representation" (2013, 73). The representational fidelity criteria aid in determining whether the model makes the right *predictions* for the right *reasons*. This commits to more than simply adequacy for purpose, as there is nothing in the assessment of adequacy for purpose that directly addresses the complicated evaluation of causal structure matching. We see here that similarity assessments again aim at a different kind of evaluation than adequacy for purpose assessments.

3.4 More Than One Evaluation of Model Fit

In constructing a model, modellers make evaluative decisions in determining what about the target system is important for the model to include and how the model's inclusion of certain elements of the target system, or lack of others, should be weighted in the weighted feature-matching equation *S(m,t)*. As I argued in the previous chapter, we can make an assessment of the model construction, and the model's similarity to the target system through understanding how the scope and purpose of the model inform the establishment of an *S(m,t)* equation. Yet Weisberg thinks it is the fidelity criteria that allow for the evaluation of whether *S(m,t)* is similar enough, and ground our evaluation of the degree in the hypothesis statement, "model *M* is similar to target *T* for scientific purpose *P* over the domain of applicability *d* to degree *S(m,t)".*

Should the fidelity criteria be understood as identifying a very different component of assessment? Rather than assessing in what ways the model is *similar* to the target system, the fidelity criteria are concerned with evaluating how closely the model output must match the world in order to be considered an *adequate representation.* Fidelity criteria aid in determining if a model is an accurate representation of the real world from the perspective of the intended *uses* of the model.

However, before we even can evaluate if the output of the model is useful, we must have already made evaluative claims about the model's construction. If the fidelity criteria are our standards for evaluating how close the model output must be to the target system in order for it to be adequate representation, then we need to be clear about how claims of this kind of adequacy are grounded. These are grounded on our assessments of similarity that are made during the construction of the model. What needs to be recognized is that there are two components to the assessment of the model fit—similarity as it relates to the evaluation of the model's construction relative to the target system and purpose, and the adequacy evaluation of the model's output for a particular purpose.

I argue that we should consider assessing similarity and assessing adequacy as two conceptually distinct components of evaluation. The modified similarity-relation hypotheses provide an understanding of how to assess similarity through the weighted feature-matching equation. The fidelity criteria ultimately are not the same sort of evaluation as the assessment of adequacy-for-purpose. They are different because their target of evaluation is different. Fidelity criteria support the assessment of the similarity relation being adequate and do different work than assessments of a model's adequacy for its intended purpose.

3.5 Assessment of Adequacy for Purpose, or Similarity?

This brings us back to the question of whether one of these assessments should be preferred over the other. Thus far, I have argued that the account of the similarity relation I am drawing on has implicitly identified two components of evaluation, both related to assessing similarity. The first relates to establishing similarity during the model construction; the second relates to comparing the output of the model to the comparable output of the world[35]. This second kind of evaluation, which utilizes the dynamical and representation fidelity criteria, may seem comparable with an assessment of adequacy for purpose. However, I have argued that the dynamical and representational fidelity criteria achieve something different as a result of the target of their evaluation being the similarity relation.

Yet, adequacy for purpose includes a kind of evaluation that has not yet been accounted for. Adequacy provides the means which enable a modeller to evaluate more than just how a model is similar to the target system (and likewise for the model output). It is through assessments of adequacy that the model is evaluated for its *usefulness* or

[35] Recall from Chapter 1, all models are considered to produce an output. The simplest case is a predictive model, where the output is a future (or past) state of the target system. But other things can count as outputs as well. These can be structures in the model itself that can feature in an explanation, or a description of the target system. Which outputs are of interest will depend on how we are using the model.

I am employing "output" in a broader sense than is traditional. An output is purpose-dependent and can vary based on what question the model is used to answer. This can include, in addition to questions about predictions, questions related to interrelations of the structures in the model itself, or questions related to what the model represents. While this may seem like a strange usage, the reason I am doing this is so that I can talk about the different ways in which we employ models, using the same terminology.

effectiveness for a given purpose for which it has been constructed. An assessment of usefulness is importantly different from assessment of representation.

My discussion thus far has identified two ways in which a model might be understood to fit, or be successful. On one hand, it can fit when it is similar in certain respects to the target system it intends to represent, and the output of the model can be similar to the real world. On the other hand, a model can fit and be considered successful if the model is adequate or useful for the job or purpose it was designed to serve. It is plausible that scientific models can be evaluated relative to each of these two dimensions, and the different evaluations can have different epistemic value.

For example, consider a model constructed to have an extremely high amount of similarity, such as a 3-D paper model of a tree in which the model contains exactly the same number of leaves as the tree, same patterns in the trunk. And if wind passed through the leaves, the model provides a similar enough demonstration of the ways in which the leaves move. However, this model would not be adequate for providing an explanation of how the tree came to have its particular shape. Likewise, it is possible to design a mathematical model that provides extremely accurate predictions for a system, yet was designed ad hoc with no consideration for having any similarity with the actual target system[36].

For these reasons, I propose a framework for assessment of model fit that involves components of both similarity and adequacy. The framework assesses how the real world phenomena are intended to be represented in the model. This is the initial evaluation of the similarity relationship between the model and the target system. This is assessed in the context of constructing the weighted feature-matching equation *S(m,t)*. After the *S(m,t)*, similarity is assessed a second time when comparing the model output and the comparable target system output. This evaluation of whether the similarity is enough for the purpose at hand is captured through the fidelity criteria, which specify the modeller's standards of similarity. My framework also includes assessing model fit in terms of

[36] For more on this issue see discussion in Chapter 4.

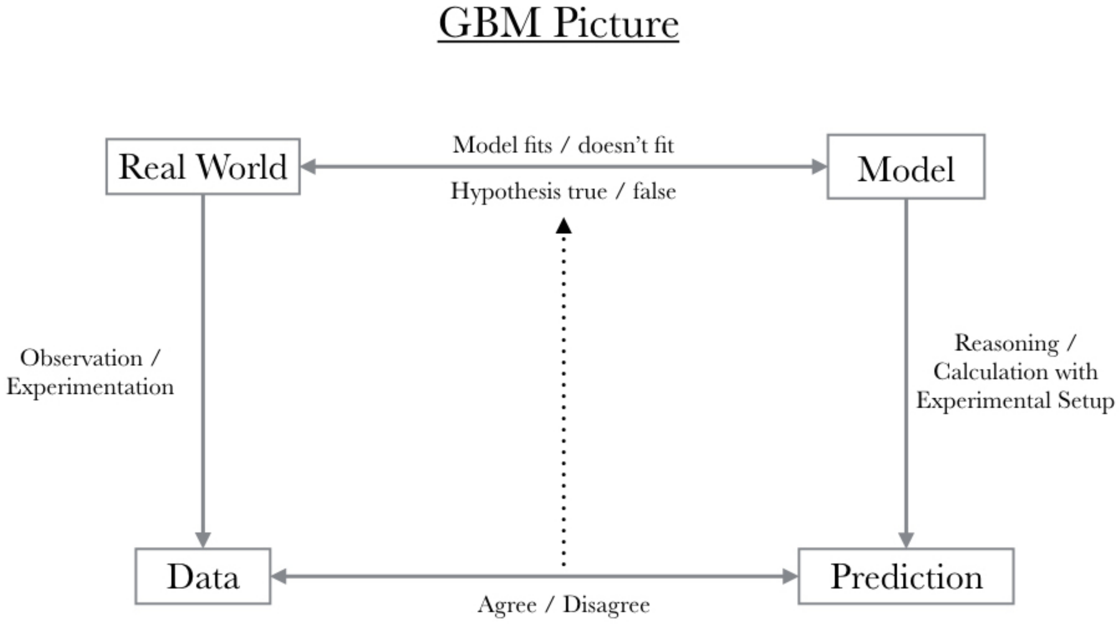
adequacy for purpose, as it relates to the corresponding evaluative components (i) and (ii) discussed above. This assessment of adequacy offers to my framework something that the fidelity criteria do not: an evaluation of the usefulness of the model for a particular purpose.

The remainder of this chapter will detail my proposed framework. Chapters 4 and 5 will offer support for this framework as being a meaningful way to capture the scientific practice of evaluating model fit, as I finally turn to examining the role purpose plays in model evaluation (chapter 4), and then provide a case study of competing models from astrophysics (chapter 5). I will argue it is through assessment of both similarity and adequacy that one can determine if it is permissible to extend the model beyond its initial purpose, and justify making inferences from models to claims about the real world. I propose a framework for model fit evaluation that incorporates elements from both accounts in order to form a stronger, more complete framework for these reasoning processes.

3.6 Proposed Framework

In order to provide my proposed framework in a clear manner, I will be adapting a figure from Ronald Giere, John Bickle, and Robert Mauldin's book, *Understanding Scientific Reasoning*, in which they detail four components of a scientific episode involving models. I have chosen this account, summarized in the figure, as my starting point, as it represents a fairly standard view of reasoning with models. But I will modify it in order to account for the points I have made thus far. It will also allow me to clearly indicate where the different aspects of evaluation of hypotheses about model fit come into play.

Figure 2: Giere, Bickle, & Mauldin (GBM) Scientific Episode.



Complete picture of a scientific episode involving models for Giere, Bickle, and Mauldin (GBM) (adapted from 2006).

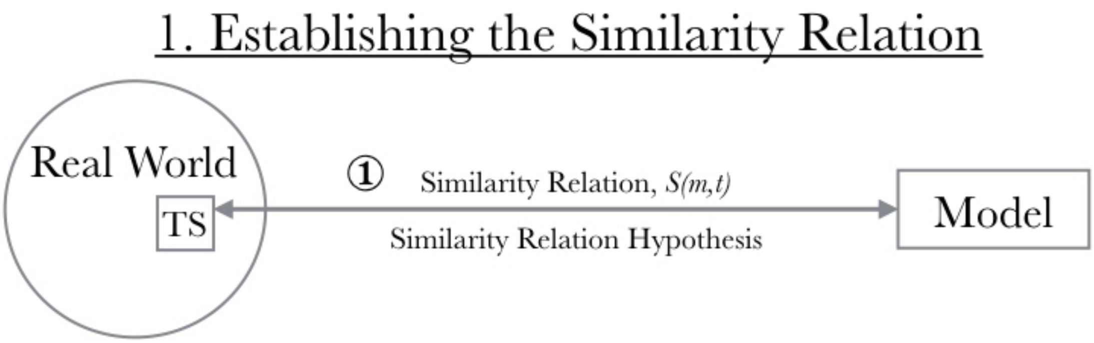
3.6.1 *Component 1: Establish the Similarity Relation*

According to the GBM account, when evaluating a model, we first establish a relationship between the real world and the model, expressed by a theoretical hypothesis. The theoretical hypothesis asserts that the model "fits" the real world. The authors take it that the model will fit only in some respects and to a certain degree of accuracy. If it does not fit to the specified degree of accuracy, then the hypothesis is false. Recall that Weisberg starts from Giere's conception of a theoretical hypothesis in developing similarity relation hypotheses about model fit. However, models are not directly compared to the real world but rather to target systems, which are understood as parts of real-world systems[37]. In light of this, I modify my diagram to reflect the fact that the

[37] For Weisberg, we can also generate hypothetical or abstract targets, or even conduct targetless modeling. However for hypothetical, abstract, and targetless modeling of the target system, regardless of if the process is that of abstractions over many phenomena, imaginary systems, or nothing at all, the process should still be understood as parts of real world systems.

comparison of fit is between the model and a target system, which is part of the real world.

Figure 3: Component 1 – Establish the Similarity Relation.


1. Establishing the Similarity Relation
Real World
TS
①
Similarity Relation, S(m,t)
Similarity Relation Hypothesis
Model

Component 1 of evaluating model fit establishes the similarity relation between the target system and the model.

I have also argued that the most general way to understand models is as being used in various hypotheses, rather than being true or false themselves. We are assessing the hypothesis statements, not the truth of the model itself. As such, the comparison of the model to the target in the above image to be understood as related to a hypothesis about the model. The form of the hypothesis involved here is a hypothesis that includes a similarity relation. Therefore, it should be understood that the first component in an overall account of evaluating model fit is about establishing the similarity relation between the model and the target system. The first part of model evaluation involves evaluating the initial relation between the target system, and the model being constructed.

The first component in model fit evaluation is about determining how the real world phenomenon is intended to be represented in the model. This relation is characterized in part by the modeler's construal, understood as the assignment and scope appropriate to that modeling task[38]. The scope specifies which aspects of the target

[38] It is possible when assessing a model to do a rational reconstruction of the hypothetical decision-making process that lead to the weightings or choice of a particular similarity relation. In this sense, it can be a component of epistemic assessment in cases where the particular decision-making process is unavailable.

system are to be represented by the model, and assignment provides information on how that representation is achieved. This first component involves a judgement of a similarity relation that holds between the model and target and is captured in the establishment of the *S(m,t)* equation.

As I argued in Chapter 2, this first aspect of evaluation involves determining what in the model will represent elements of the target system, and takes place via the process of constructing our weighted feature-matching equation for the model. In this way, assignment and scope are related to the choice of elements of the feature set Δ. As I explained in the black hole modeling example, given a certain application, the model-target fit might be evaluated as highly similar relative to one context, and relative to other contexts, with different applications, the same model might be evaluated as having a low similarity. Therefore, specifications of the model's domain of application at this stage must be made clear.

The similarity between a model and a target system is captured by the pragmatic use of the elements of the *S(m,t)* equation. Given that similarity can be accused of being a subjective judgment, there is a need to be very clear about the reasons for considering the model to be similar to the target system. Therefore, it is important to formulate statements about similarity precisely. This, I argue, is best accomplished by the modified similarity hypothesis statement:

> Model *m* is similar to target *t* for scientific purpose *p* over the domain of application *d* to degree *S(m,t)*.

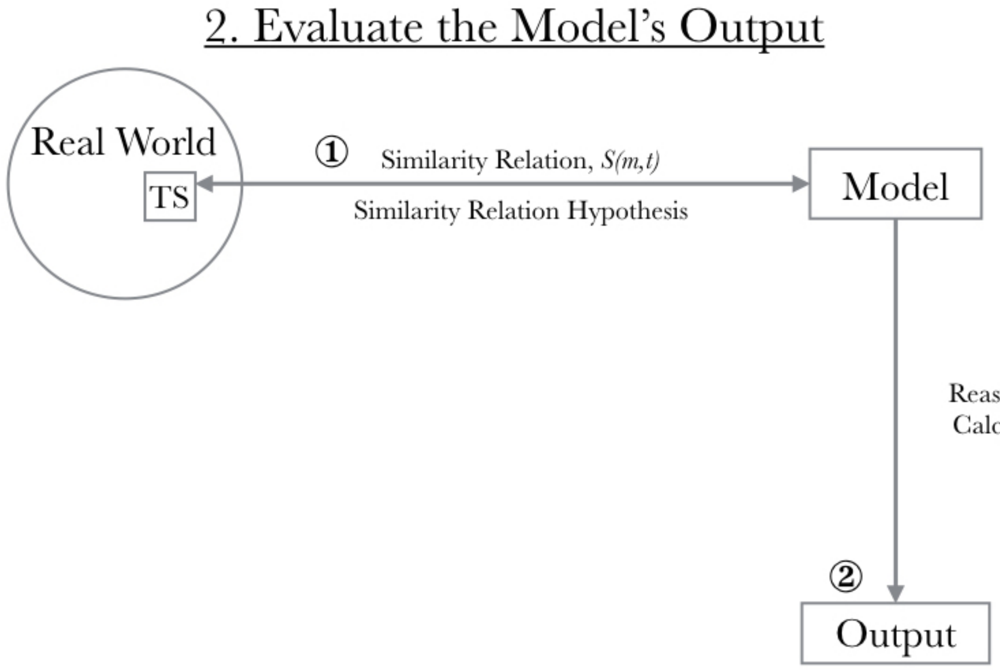
This hypothesis embodies information regarding the ways in which the model has been constructed, and establishes the ways in which it is similar to the target system. The evaluation of the hypothesis proceeds by specifying what are the model, target system, scientific purpose, and domain of applicability, and how the model is similar to the target system, embodied by *S(m,t)*.

In sum, there is a clear, distinct first component in the evaluation process, which is best characterized by the modified similarity-relation hypothesis that contains the weighted feature-matching equation. The similarity-relation hypothesis establishes what the connection between the model and the target system is. It also establishes the sense in which the modeller understands aspects of the model and the target system to be similar, since this part contains the modeller's construal, which determines how we evaluate the first component of fit. For these reasons, it is clear that this first component of evaluation has its own means of establishing and evaluating an element of model fit. There are two components to characterizing model fit, and it is critical that we separate the two conceptually. The two characterizations aim to do two different things. And conflating the two, or failing to separate the two can lead to confusion about what in a model is being evaluated, and how that evaluation proceeds.

3.6.2 *Component 2: Evaluate the Model's Output*

Returning to the GBM picture, a model next provides us with a prediction. The model and the prediction it provides are related by reasoning or calculation in light of experimental design (2006, 30). For the evaluation that takes place at this step, GBM states we are to "identify a prediction, based on the model and experimental setup identified, that says what data should be obtained if the model actually provides a good fit to the real world" (2006, 35). I modify *prediction* to instead be characterized as the *output* of the model. An output is purpose dependent, and can vary based on what question the model is used to answer. This can include, in addition to questions about predictions, questions related to interrelations of the structures in the model itself, or questions related to what the model represents. While this may seem like a strange usage, the reason I am doing this is so that I can talk about the different ways in which we employ models, using the same terminology. Needless to say, broadening prediction to include other purposes of the model, such as simulating aspects of the past, or providing information or an explanation about the causes of phenomena of interest, will better capture the complexities of what is the possible output of the model.

Figure 4: Component 2 – Evaluate the Model's Output


2. Evaluate the Model's Output
Real World
TS
①
Similarity Relation, S(m,t)
Similarity Relation Hypothesis
Model
Reasoning / Calculation
②
Output

Component 2 of evaluating model fit generates and evaluates the model's output.

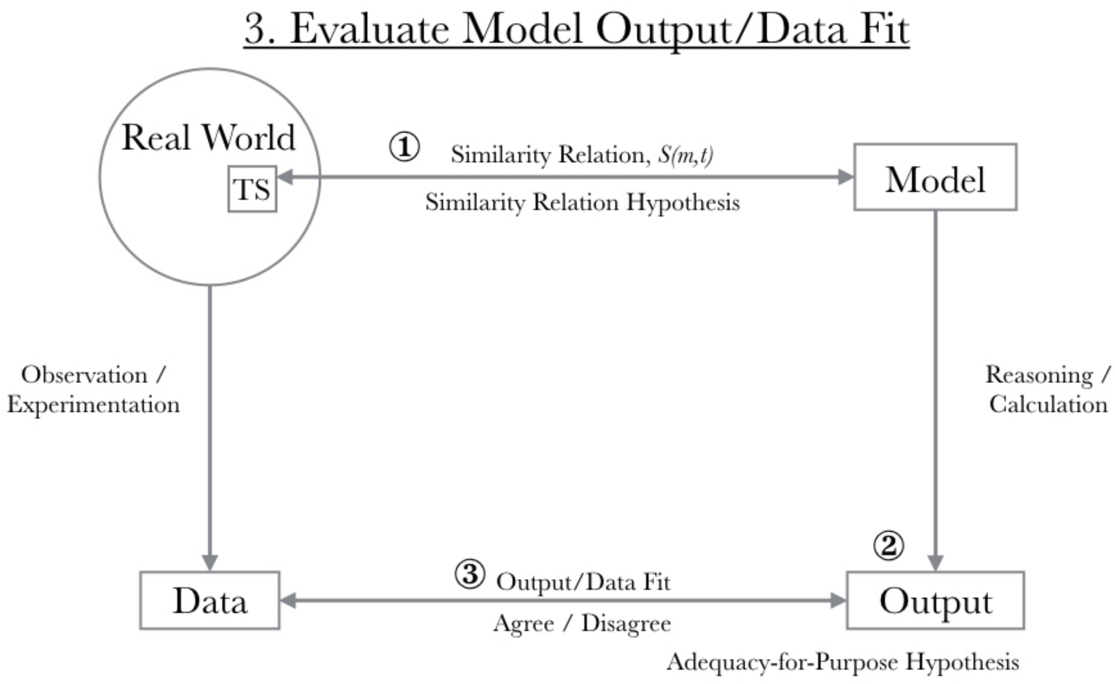
Broadening the prediction component of the GBM view to output will also allow for capturing what is of interest in the first part (i) of evaluation of adequacy for purpose. The first task in evaluating model fit relative to adequacy for purpose involves determining what one is likely to observe if it is true that the model is adequate for its purpose. My second component should be seen as equivalent to Parker's step (i). Before comparing the model's output to data, we need to, in some sense, make an evaluation of whether the model we have constructed is giving us a reasonable output. Of course, as Parker points out, this may not always be possible (2009, 2010). We may be able to generate the output, without being able to determine what will be observed if the model is indeed adequate for the purpose to which it is being put. We may not always have the capabilities to determine what would constitute an adequate model, even though an output can be generated. Regardless, it is worth at least attempting to determine what would follow from an adequate model and would be a reasonable output. In the event it is not possible to determine what an adequate output of the model would be, there will be other strategies in components 3 and 4 that may be useful.

In component 2 of model evaluation, the output from the model is obtained. By asking what it would mean for the model to be adequate, the modeller also starts to reason back about the target system and gains knowledge about where and what to look for back in the world. Such information can help direct research, and guide evaluation connected to components 3 and 4. Finally, component 2 also allows for an initial first check-point to determine if the model's construction is drastically off-track. If a model does not adequately reproduce essential data, then the similarity relation can be re-examined to determine what might be missing (or included) properties of the target system that should be included (or not included) in the model.

3.6.3 *Component 3: Evaluate Model Output/Data Fit*

The third component involves comparing the output of the model to data. These two are related by a "physical interaction that involves observation or experimentation" (Geire et al. 2006, 30).

Figure 5: Component 3 – Evaluate Model Output/Data Fit


3. Evaluate Model Output/Data Fit
Real World
TS
① Similarity Relation, S(m,t)
Similarity Relation Hypothesis
Model
Observation / Experimentation
Reasoning / Calculation
②
Data
③ Output/Data Fit
Agree / Disagree
Output
Adequacy-for-Purpose Hypothesis

Component 3 of evaluating model fit involves comparing the model output to the comparable output from the world.

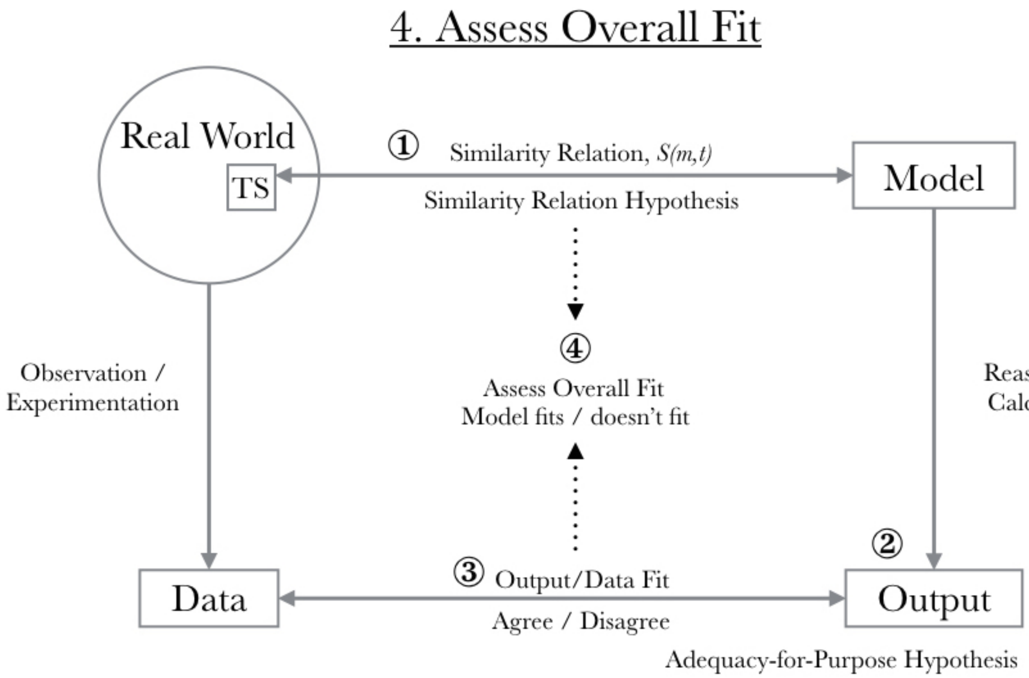
Parker suggests in her evaluative step (ii) that we now check how well what is actually observed fits with what we are likely to observe if the model is adequate. In this component, the model output is compared to the data output through a means of assessing how well these match given the intended purpose of the model. This provides an evaluation of whether the model is adequate. Component 3 is also a comparison of similarity again, by establishing that the model output is similar enough to the data, and within a certain degree of accuracy. This evaluation employs the dynamical fidelity criteria.

At this stage the ways in which the data and model output disagree are also identified. Very rarely do the data clearly and exactly agree with the output of the model. This is why it is important to decide to what degree these outputs must be similar given the purpose. There may also be cases in which the modeller is not able to obtain a comparable real world output to compare the model output against. As discussed in detail by Parker, it often is not clear how the features we can observe in the world relate to the question of whether the model is adequate for the purpose the modeller would like to use it for. I will reserve discussion of what is done in these instances and how this affects the framework until chapter 4. However, the quick answer is robustness analysis.

3.6.4 *Component 4: Assess Overall Fit*

Finally, the fourth component to model evaluation assesses the overall fit of a model along two dimensions—similarity and adequacy. This component is the overall assessment in which both the similarity and the adequacy hypotheses that are at play in this framework are evaluated. The adequacy-for-purpose hypothesis informs us about the assessment of whether the model is qualitatively or potentially quantitatively satisfactory, for the purpose at hand. The similarity-relation hypothesis assesses the manner in which the model, overall, is similar to the target system relative to the purpose for which the model was constructed.

Figure 6: Component 4 – Assess Overall Fit



Component 4 evaluates the overall model fit relative to similarity and adequacy.

The adequacy-for-purpose hypothesis obtains its assessment through comparison of the output of what we consider an adequate model with the comparable real world output. It is only with all these components on the table that one can assess if the model is in fact adequate for the purpose at hand. That is to say, if the model as a tool, is successful for the job that it is put to. We are also permitted to assess if the model, though constructed with a similarity relation for one purpose, can be adequate when put to a new purpose.

By understanding the similarity-relation hypothesis as I have suggested, the ways in which the model is similar to the target system are established. The fidelity criteria aid in the evaluation of whether the similarity relation is satisfactory or acceptable and to what extent. The dynamical fidelity criteria are used to evaluate and determine what is acceptable similarity between a model's output and the data of the real world. The representational fidelity criteria are drawn on in the overall assessment of whether the model is similar in the sense that it is providing these outputs for the right reasons. In this way, the fidelity criteria judgements can also be used to support and articulate further details in adequacy-for-purpose judgements of that hypothesis.

Geire et al. (2006) also provide detail on an underlying assumption of this kind of account of the model-world similarity relation. With respect to the relationship between the model's output and obtained data from the real world, Geire et al. state that,

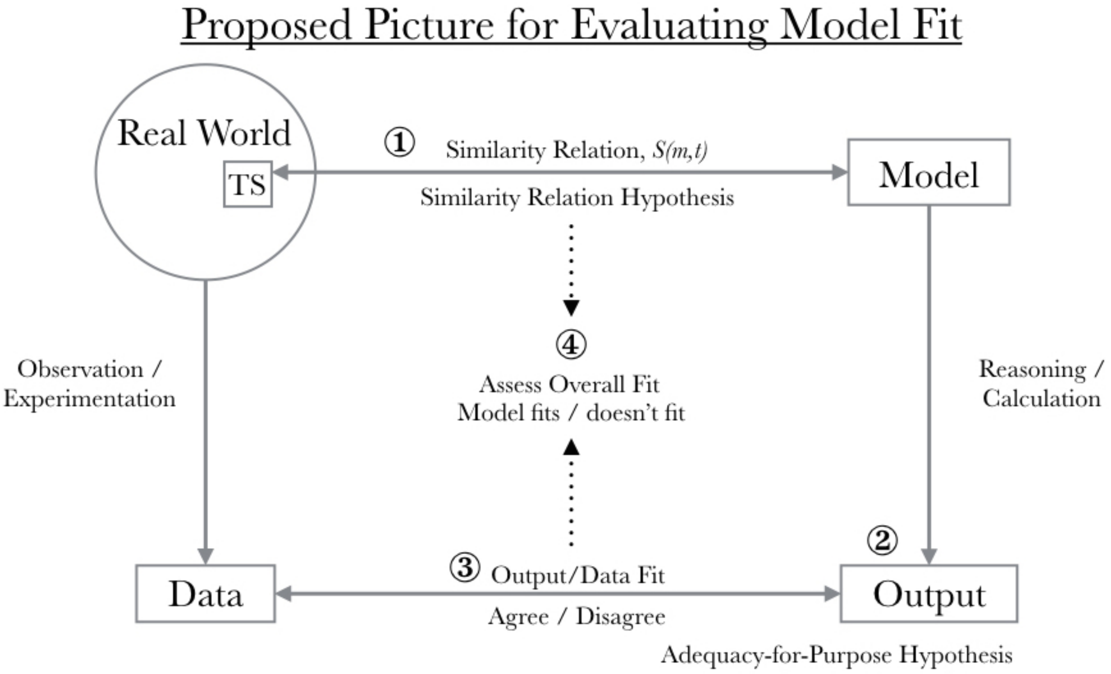
> If what is going on in the real world, including the experimental setup, is similar in structure to the model of the world, including knowledge of the experimental setup, then the data and the prediction should agree. That is, the data should be as described by the prediction. On the other hand, if the real world and the model are not similar in the relevant respects, then the data and prediction may disagree (2006, 30).

The representational fidelity criteria establish our standards for evaluating how well the structure of the model maps onto the target system. If the data and outcome agree to the degree we have established based on our fidelity criteria, we have positive evidence for the model fit as it relates to the similarity hypothesis.

It is this fourth component that also allows for the assessment not only of whether the model is in fact adequate for the purpose, allowing us to determine the extent to which our adequacy-for-purpose hypothesis is true, but also allows for evaluation of the similarity-relation hypothesis as well. It is the evaluation of both of these elements, taken together in this framework, that allows a modeller to say that their model fits, both with respect to both target system, and data outputs[39].

[39] It is important to note the complexities involved in evaluating the strength of evidence with respect to assessments of adequacy. The issue of what counts as evidence for assessments of adequacy is complex, and as Parker herself notes, "confident conclusions about what would count as evidence that supports or confirms a model's adequacy for a given purpose sometimes remain out of reach" (2010, 10). While an understanding of the nature of this process is still needed, further work on the role of evidence would fit in my proposed framework by providing further discussion of judgments made by modelers at this stage.

Figure 7: Complete Proposed Picture of Evaluating Model Fit.


Proposed Picture for Evaluating Model Fit
Real World
TS
① Similarity Relation, S(m,t)
Similarity Relation Hypothesis
Model
Observation / Experimentation
④ Assess Overall Fit
Model fits / doesn't fit
Reasoning / Calculation
②
Data
③ Output/Data Fit
Agree / Disagree
Output
Adequacy-for-Purpose Hypothesis

3.7 Conclusion.

Model evaluation in this framework involves the following components: 1) a similarly relation hypothesis is developed based on the purpose for which the model is constructed. This hypothesis contains *S(m,t)*, which is a specification of all of the ways in which the model is similar to the target system, along with purpose *p* and domain of application *d*. The model's similarity is evaluated at the construction stage, but the hypothesis itself is not evaluated until later, when our adequacy for purpose hypothesis is evaluated as well. 2) The modeller obtains an output from the model, and if able, evaluates if that output is reasonable. 3) The modeller formulates the adequacy-for-purpose hypothesis, and compares the output of the model to the data from the real world. 4) The model-data fit informs our evaluation of both hypothesis statements, and we synthesize these judgements into an overall assessment of model fit.

There are two important considerations to keep in mind. First, the purpose for which the model is developed, which is encoded in *S(m,t)*, does not need to be the same purpose relative to which the adequacy-for-purpose hypothesis is assessed. Second, although there are four components that make up my framework, this framework is not committed to the idea that the components proceed in discrete steps. In actual practice, outcomes at later stages can influence and provide feedback on considerations made at earlier stages. Furthermore, there can be trade-offs between similarity and adequacy given that there also needs to be a certain level of tractability. A model could be evaluated as fitting with high similarity, in that the model includes all features the modeller considers relevant for providing a certain prediction, but this does not necessarily mean that model will be adequate for providing a prediction. Having too many features might make it challenging for the model to actually generate an output—through it may have a high similarity, it will not be adequate for proving the prediction.

Three main questions have driven the analysis over the last two chapters. With respect to the first question—what is the target of evaluation in model fit—I have argued that it is a matter of evaluating hypothesis about the model rather than evaluating the truth of the model itself. With respect to the question of what form the hypotheses should take, I have detailed two forms—adequacy-for-purpose hypotheses and similarity-relation hypotheses—and argued they actually embody two different components of model evaluation. The answer to the third question, how do we evaluate the hypotheses, is that the evaluation of the two hypotheses are part of an overall picture of model fit evaluation. I suggest therefore that this account of model fit provides a more complete framework than others currently offered in the literature.

This leads me to a final question: How do we get from these claims about adequacy and similarity of the model to claims about the real world? What justifies us in making inferences from our models to knowledge claims about the real world? Is there more to making these types of claims than appealing to these established relations of similarity and adequacy? One might argue that this is enough to justify our inferences, but is this really the case? Additionally, it is now important to turn to a question of what to do when there is not as much theory to guide judgements about the weighting of the similarity relation, as well as what to do when a modeller cannot even determine what the expected prediction or outcome of a model might be. What does one do if they are not able to generate the outcome, and perform the complete account of model fit as detailed above? These questions will be the focus of the next chapter.

Chapter 4

4 Making Inferences from Models

4.1 Introduction

In the preceding chapters I have established that evaluations of model fit should be understood as a matter of evaluating hypotheses about the model, rather than evaluating the truth of the model itself. I examined two ways to formulate these hypotheses and argued that we need both a hypothesis about a model's similarity relation relative to a purpose, as well as a hypothesis about adequacy relative to a purpose. These two hypotheses have different targets of evaluation and work together to provide an overall evaluation of model fit. This chapter will offer support for this framework as being a meaningful way to capture the scientific practice of evaluating model fit, as I finally turn to examining the role purpose plays in model evaluation.

The argument in this chapter will consist of two parts. The first part is about how to understand model assessment. It will apply the framework I developed in the previous chapter to show that model assessment must be understood as relative to a purpose. I will develop several examples from astrophysics in detail to support this argument. Part of the argument requires me to establish there are at least three general kinds of purpose. The particular purposes to which any given model is put can be quite specific. However, I will argue that they fall into three general kinds: description, prediction, and explanation. The difference is related to the kind of output obtained from the model when attempting to use it for a particular purpose. The details will also become clear through the examples presented.

The second part to the argument in this chapter is related to how inferences from the models about the world are justified. I argue that these justifications are always grounded in similarity between the model and the target system. My framework allows for the connections among similarity, adequacy, fit, and justification for inferences about the world to be disentangled. I begin with some preliminary arguments to set the stage.

4.2 Preliminary Arguments

Claims about a model's fit with a target system, or the success of a model, can be meaningfully analyzed using the framework I have developed. First, the framework provides a means by which to disentangle claims and the background reasoning that supports claims about a model's success. Second, it provides the needed understanding and analysis of the source of a model's success. This is critical for seeing how the model can be extended, and understanding what grounds justifications for inferences made from the model.

Thus far, I have argued that assessment of model fit has two parts, assessment of the similarity relation and assessment of adequacy for purpose. This claim will be further supported in this chapter as I deploy the framework to analyze examples of models from astrophysics. The assessments of both similarity and adequacy are always made relative to a purpose. I argue that there are, in general, three kinds of purpose to which a model can be put. By using the framework to analyze a model's similarity and adequacy relative to a purpose, I argue we gain a better understanding of how claims about model success can be understood.

We also gain grounds for making inferences from the model back to the actual system, as well as the justification for why a model can be extended. There are three ways in which a modeller may want to extend the model, either 1) to a new purpose (for example, from using the model to predict then using it to explain why that feature is part of the system; 2) to a new domain of application (for example, if we think a model gives successful prediction on short time scales of 10 years, can it give us successful predictions over the next 100 years?); and 3) to real world claims (for example, if our model predicts x occurring, then we are justified in thinking that x will be the case in the real system as well). In the end, I argue that the understanding of this source of success of a model and the justification for seeing how a model can be extended is grounded in the model's similarity relation.

In this section, I will first provide an argument for there being three kinds of purpose. I then provide further details on how one can understand how the inferences from models inheres in the similarity relation. With these final parts of my argument on the table, I will then demonstrate how the framework can be deployed as a tool to gain insight into evaluations and claims about models.

4.2.1 *Three Kinds of Purpose*

My framework, as developed in the previous chapter, is based on the idea there are three kinds of purpose that are relevant for model assessment: description, prediction, and explanation. One might object to this as an arbitrary distinction and claim that there is no reason to break purpose up in this manner, or that there are more than three kinds of purpose. However, I think these kinds of purpose correspond to the major goals of science. For example, Andersen and Hepburn (2015) characterize the aims of science in their article, "The Scientific Method" as: "the basic aim and method of inquiry identified here can be seen as a theme running throughout the next two millennia of reflection on the correct way to seek after knowledge: carefully observe nature and then seek rules or principles which explain or predict its operation". Models are representational tools used in scientific inquiry for these ends. Therefore, the three-part division of kinds of purpose into description, prediction, and explanation is a natural way to carve up the kinds of questions models are employed to help answer. So, I shall invoke this classification, and, as will become apparent as I examine real cases of modeling, will speak in favour of the utility of this way of understanding purpose.

The reason it is important to consider these different kinds of purpose is because they correspond to different kinds of output that modellers attempt to obtain. In the case of a descriptive purpose, the modeller obtains from the model an output that somehow represents the features present in the target system[40]. In the case of a predictive purpose,

[40] These can include unknown descriptive features, which are distinct from predictions. To illustrate this, consider the following analogy of constructing a mental map, or model, of an old house. One walks through the house, representing each room in a mental map. Imagine that, if in walking through and developing the mental representation of the house, we realize that there must be a hidden room between the library and the study. What has happened is that in order for the model to accurately represent the rooms we have observed it must be the case that there is an unobserved feature of the house. We have not generated a prediction. Rather, it is a condition on the adequacy of our model as an accurate description of the house. What could

the modeller obtains from the model an output corresponding to a future or past state of affairs about the target system that is not originally built into the model. In the case of an explanatory purpose, the modeller obtains from the model an output that can serve as an explanans in an explanation of some phenomenon[41].

Furthermore, with respect to constructing a model, the purpose to which a model is to be put will affect what features (mechanisms and attributes) the model needs to include. For example, if a model is being constructed to describe the current state of a target system, then the model must be constructed such that it includes the relevant features about the target system in the present state. If the modeller wants to then use that model to predict a future state of the system, then the model must include the features that are considered relevant; for example, including relevant mechanisms would be essential. Of course, any model can be constructed with more than one purpose in mind, as well as serve more than one purpose simultaneously. However, this should be understood as a composite of these individual assessments made relative to the kinds of purpose included in my framework. A model may not always provide a satisfactory prediction of every aspect of the target system. Furthermore, a model that makes a satisfactory prediction might not provide a satisfactory explanation for why what was predicted will be the case. If a model is expected to give a good description and good prediction, those two elements must be assessed individually, relative to the standards of that purpose.

be said to be a prediction is the interaction between a measuring device and the target system (in this case perhaps a sledgehammer breaking through a wall). However, this is an expanded model.

Another analogy to make this point clear is the following: consider what Christopher Columbus is apocryphally said to have "predicted" in 1492. It might be natural to say that he predicted the world was round. However, that the world was round was not a prediction. What he in fact predicted was that if he sailed west (i.e. performed a certain test), he would find himself in Asia, thereby validating the accuracy of his model of the structure of the earth. His model of the earth as a sphere was a descriptive model of a possible way the earth was. "That the earth is round" is not an output of the model, and therefore not a prediction in the paradigm sense of prediction.

[41] This account of explanation is taken from Alisa Bokulich (2011) and will be further detailed later.

4.2.2 *Extending Models and Making Inferences.*

To summarize the point about purpose: The original intent during the construction of a model[42] does not make it the case that the model can be used for only that purpose. A modeller can attempt to use any model for any purpose, and in most cases, models are put to many purposes simultaneously. However, a model's success relative to one purpose does not entail that it will be successful for everything. My framework is a tool for teasing apart these distinct threads of model evaluation. Insofar as a model makes predictions, it is assessed relative to different criteria than it is insofar as it gives explanations. And the assessment of a model relative to its descriptive fit is again different. This is due to the difference in the relevant output of a model for these kinds of purpose.

However, the success for many models depends very heavily on their success along a single dimension of assessment. In §4.3, I will give several examples of how these assessments are to be analyzed in my framework. In general, this analysis proceeds as follows: The model's primary purpose is identified and the corresponding kind of output from the model is assessed relative to its real world analogue. If it is not possible to obtain the real world analogue output, then robustness analysis is deployed. Robustness analysis is a means to analyze the output of models, which allows us to determine the extent to which the model's output might depend on particular idealizing and simplifying assumptions. The model output, and real world analogue output components are used to make an assessment of adequacy and similarity relative to the relevant purpose. If the model was explicitly constructed with a particular purpose-relative *S(m,t)* in mind, then the similarity assessment will be relatively easy. If not, some reconstruction will be necessary. This purpose-relative analysis of model fit along both

[42] Of course, it is not the case that every model is explicitly constructed with a particular purpose in mind. Often models are inherited from other areas in science and are applied to new cases. Those that do not work are discarded, and those that work are kept. What my framework offers is a tool for analyzing why these models are successful. I will show that it is due to a high degree of similarity. Even though similarity was not taken into consideration when the model was imported into its new domain of application, my framework offers guidance about how to reconstruct the kind of similarity assessment that grounds the fit of the model.

similarity and adequacy dimensions provides for a more meaningful understanding of the model.

As mentioned above, a primary goal in using models as representational tools in science is to learn more about the system the model represents and answer a range of questions, such as: What is the best way to make sense of the observations and data that are being acquired? What will the future state of the system be? Why does the system behave the way it does? What is causing the system to behave in that way? As already discussed, the nature of models presents a certain challenge in making claims about the world based the model. Therefore, the fundamental question that needs to be answered is what justifies making inferences from our models to knowledge claims about the real world. Ultimately, this justification is important because we must have confidence that the results, or outputs of the model, are not simply artifacts of the particular means used in constructing the model. What prevents one from directly making inferences from the model to the real world is the need to ensure that the features of the model are not just artifacts, or accidents from the representation and idealizations.

One might think that an assessment of the model's adequacy for the purpose might be enough—the model provides an output that seems reasonable when checked against an output from the real world and is evaluated as matching well enough; so we should be able to extend it. Yet I argue similarity is also a piece in the overall assessment of model fit. What is the relevance of a similarity assessment when we have an assessment of the model's adequacy? What does an assessment of the similarity-relation offer that assessment of adequacy for purpose does not? The reason similarity plays an important role in my framework is because it grounds the inferences that can be made from the model. A model having a high degree of similarity relative to a purpose-dependent *S(m,t)* is what grounds the inferences that can be made about the world from the model. It is because we can identify how the model is similar to the target system in the relevant ways that we can be confident in drawing conclusions about the target system that go beyond the information that was built into the model in the first place.

Wendy Parker also points out this problem. She uses an analogy of a model being like a tool that is known to be useful in one case. She asks what would be needed to extend this tool other cases?

> If a tool is suitable for removing this nail from this wall, will it also be suitable for removing that other nail from that other wall? The answer is impossible to determine … unless one considers further information about the tool and about the original nail and wall. … Similarly, except in special cases, it is impossible to determine what should be observed in a given test situation if a model M is adequate-for-purpose, even with the help of true auxiliary assumptions about the conditions of the test situation, unless further information about M and about the target system is available (2010, 9).

Elsewhere, she says,

> …from the assumption that a model M is adequate for a purpose P, little else follows about either the target system or M, leaving it unclear what scientists should find in various test situations if M is adequate for P, unless they have additional information about M and the target system … Perhaps the best that scientists can do is to draw on what led them to think that the model might be adequate in the first place (2010, 8).

Similarity plays this role in my framework. It is an explicit catalogue of the reasons and ways in which the modeller took the model to accurately represent the target system in the first place. As such, it fulfills the role of providing "further information about M and about the target system". This allows for the extension of the model to new cases where it is known that the model is relevantly similar to the new target system, a new domain, or for a new purpose.

What an assessment of adequacy alone fails to do is provide information about why the modeller considered the model to be adequate in the first place. Without knowing what underlying structures, or features (attributes and mechanisms) made the model a good representation of the target system, it is impossible to know why extension of the model is justified. But perhaps more importantly, because it lacks this feature, adequacy does not provide grounding for inferring from something being true of the model to that thing also being the case in the target system the model represents.

My framework provides an account of the grounding for these inferences in the similarity relation. While similarity may not be the only way to extend an adequacy account to provide grounding for these kinds of inferences, it is included in my framework because it is a general, useful, and comprehensive way to think about the process of encoding the underlying structures, attributes, and mechanisms that connect the model to the target system. Furthermore, the specification of features in the weighted feature-matching equation is what preserves the information about what has been included in the model and why. From the following examples, I will show that the similarity relation is what allows a modeller to track what relevant features are uncovered through model exploration processes such as robustness analysis.

Adequacy-for-purpose assessments alone are not enough to understand the use of models in science. The repeated successful applications of a model justifies further extension only if there is reason to believe that the model is representing features that actually exist in the world. Otherwise, it could be luck that the model works in all of those cases. Those who think assessments of adequacy provide information about the underlying causal structures are making an assumption about the power of the account that is not supported by the existing literature. As characterized by Parker, adequacy does not include information beyond what the model is successful in doing. For these reasons, similarity is the basis of the model-world relation in this framework. It is the best choice because using the weighted feature-matching equation as a bookkeeping device explicitly solves the problem Parker pointed out in the above quoted passages.

The nature of the similarity-relation hypothesis I have proposed provides the basis for knowing how the model can be justifiably extended. This can either be a desire to apply a model to a new question for the same purpose (if model M is adequate for predicting x, will it be adequate for predicting y?), to different purposes (if model M predicted x, can it explain why x is case?) as well as beyond the original domain of application (if the model adequately predicts x in a certain domain, can it also predict x in a different domain?). It is in these cases that the importance of the similarity relation as part of the assessment is demonstrated.

In the end, there are various ways in which we want to be able to make inferences about the world from models. One is that causal interactions captured in the model represent actual causal interactions in nature. Second, we might also want the outputs of a model to be applied to new situations, or to stand in for evidence in instances where we cannot do experiments directly in the world. That is to say, a model might be created for a certain context or purpose, but the hope is to apply it to a new or unobservable situation, or a different purpose altogether. Third, we want to be able to say that we can trust a model that makes predictions about the future. The nature of the similarity relation also allows for the justification for inferences from claims about the model back to claims about the target system and real world.

4.2.3 *A Framework for Understanding Success Claims*

The framework I have proposed is to be a tool to use to gain a better understanding of the success of models. Thus far, I have analyzed hypothesis statements that take particular forms. How does the framework I have developed account for more realistic success claims that scientists actually make? I argue that these sorts of claims can be analyzed into more basic components that correspond to the similarity and adequacy hypotheses. Take, for example, the following evaluative claim:

> My structure formation model successfully describes the evolution of large-scale structure because it yields a two-point correlation function for the galaxy distribution that is tolerably close to what is observed.

While this statement does not take the form of the hypotheses on which I have based my framework, I argue that a claim like this is actually a composite that can be analyzed and broken down into the basic evaluative components and hypotheses I have proposed.

In this example, the target system has been identified as the large-scale galaxy distribution, and the model is a specific structure formation model. There is an adequacy claim that the model is adequate for the purpose of predicting the distribution of galaxies. I am analyzing this in terms of prediction rather than description, as was stated in the original success claim, because of the nature of the output of the model, the two-point correlation function. The success claim also includes information about why the model has been assessed as adequate. This information is relevant to the second and third components of model fit, namely, obtaining an output from the model, comparing it to the analogue output from the world, and determining whether the model is adequate. The output of the model is the two-point correlation function, and the real world observations to which it is compared are the observable small-scale distribution of galaxies. There is also a statement about error tolerance, which can be understood as coming from the dynamical fidelity criteria.

In my framework, this is a statement about the model-data fit of component 3, and corresponding assessment of adequacy of the model. So, the adequacy-for-purpose hypothesis looks like,

> This structure formation model is adequate for the purpose of predicting the distribution of galaxies.

There is nothing in this success claim about the similarity relation. This is because there is nothing that implies any assessment of the representational features of the model or its extendibility. Therefore, this success claim is not a claim of overall model fit. This is not uncommon for real world success claims made by scientists. Often what is most important to the scientists is that their model does what they want it to do, that it is adequate for the purpose. Questions about the model's representational success, and how they are justified in extending the model often are not reflected in the published work. This is the kind of reasoning that goes on behind the scenes and does not necessarily count as a result in way relevant for publication. Nonetheless, I argue that it is a crucial part of understanding how a model is successful overall.

In what follows, I analyze several models from astrophysics in order to detail how my framework can be used to understand assessment of model fit considering the three purposes a model can serve and the kind of output that accompanies each purpose. In each example, I will consider example success claims and highlight how the justification for extending a model beyond its original purpose or domain depends on success relative to similarity relative to a purpose. Likewise, any inference from a model, to claims about the world rely on the similarity relation in the same way.

4.3 Examples from Astrophysics.

While it is possible for one model, constructed for a certain purpose, to then be evaluated for different purposes, I intend to argue that it is a mistake to evaluate different purposes with the same standards. In what follows I shall discuss three possible purposes for which a model can be intended: the purpose of providing a description, prediction, or explanation. I shall also examine the output related to each purpose. Understanding how models provide a description, or prediction will be relatively straightforward. Understanding how a model explains is much more complicated. I use Alisa Bokulich's account (2011) of how a model explains, as it provides the best understanding of the relationship between a set of explanans that includes the output of a model on one hand, and the explanandum on the other.

Different models are designed and constructed with different purposes in mind. The purposes under consideration in the model construction determines the similarity relation, and what is prioritized as elements, or feature sets in the model, as well as their relative weightings. What is critical, but overlooked, is that we must distinguish the purpose considerations at construction and the purpose considerations at the assessment of the model's adequacy. A single model can serve different purposes but the evaluation of its overall fit must be constituted of its fit with respect to each of those separate dimensions of purpose. Evaluation of model fit must therefore be relativized to a purpose. The justification for the inferences we make about the world from the model follows from the model's success along those dimensions.

In this section, I will look at cases in which a model is constructed and evaluated relative to the same purpose. The cases I have chosen are cases in which I am evaluating the model fit relative to one and only one purpose. The claims made about model fit will apply only to the one purpose—description, prediction, or explanation—that is under consideration. However, this does not mean that in general every model serves only one purpose, or was constructed with one purpose in mind. It is possible to construct a model that is intended to predict and describe. However, what is important to realize is that in the weighted feature-matching formulation of the similarity relation the modeller needs to acknowledge *why* they are including the elements they have chosen. It is also possible that a model constructed for a descriptive purpose only would be the same as one constructed for a predictive purpose. Again, what is important is the acknowledgement of what purpose was under consideration in the construction stage.

Finally, it is important to keep in mind that after the initial assessment of model fit relative to one purpose (such as descriptive purpose), a model can then be assessed for its overall fit relative to a different purpose (such as a predictive purpose). What this requires is to assess the model constructed for a descriptive purpose for its adequacy relative to the new purpose. If it is determined not to be adequate for the new purpose, then one must, in revising the construction of the model, make note of what must be changed in the similarity relation such that the model can also serve a descriptive as well as a predictive purpose.

4.3.1 *How Models Describe*

Perhaps the simplest possible purpose that a model can serve is to provide a description of a target system. A model that serves the purpose of providing a description will be concerned with matching the information or empirical data about the target system as closely as possible. A model that aims to describe can be considered to be merely phenomenological, in that it seeks only to save the phenomena. In this sense, if the purpose of the model is to describe, the modeller's intentions are to accommodate the data points or features of the target as closely as possible given that the model needs to describe what is occurring in the target system over a certain domain of application.

4.3.1.1 Concrete Model of the Moon

The paradigm case of a model serving a descriptive purpose is a concrete model—a real, physical object that is intended to stand in a resemblance relationship to its target system. Take, for example, a scaled physical 3D model of Earth's moon at the present day. This model can serve a descriptive purpose, given that it represents the moon and some of its features to a very high degree of accuracy. The features of the target system that are represented in the model include the topography of the craters on the moon, relative distances between features, and coloration. This 3D model of the moon can be evaluated relative to the purpose of accurately describing these features at a particular scale. A success claim, then, might look like, “The scale 3D moon model is successful in that it accurately describes the positions of the Kepler and Copernicus craters relative to each other”.

Such a claim can undergo evaluation of model fit in my framework relative to a descriptive purpose in the following way: Component 1[43] is concerned with constructing a model such that it obtains a very strong similarity relation, given that the intended purpose of the construction of the model is to describe the system to a certain degree of detail. As such, it is important to capture the relevant attributes in the target system. In formulating the similarity relation for the model, the modeller explicitly notes what, in the target system of the Earth’s moon, they want to include in the model, and what, given the purpose of description, they are not concerned with. The modeller may care not only about what the craters on the moon look like, but also about their relative depths. The 3D model of the moon will be similar to the target system of the real moon for the purpose of describing features of the real moon. These features and their importance are captured in establishing our similarity relation *S(m,t)*, which details the weighted list of attributes and mechanisms.

Additionally, given that the model is being evaluated for the purpose of providing a description in the present, the domain of applicability will be only the domain for which the modeller already has the information they seek to describe. That is to say, given that the purpose of the model is only a current description, they do not need be concerned with possible future states of the system. In this example of the moon scale model, a concern is not what the topography of the moon may have been thousands of years ago, nor its future states.

With respect to evaluative component 2, the model’s output is some fact about the features of the model that serves as a description of the target system. For example, if the modeller wanted to know how many craters there are on the moon above a certain size,

[43] Recall that component 1 involves establishing the similarity relation between the target system and the model via the weighted feature-matching account of similarity.

they could consult the model, and count the craters. Component 2 involves determining what is to be observed if it is true that the model is adequate for its purpose as a description. In this case, it is expected that the model does in fact accurately describe the target system given that the selected features are relevant. There should be a relatively straightforward check in this example; the modeller would want to make sure the desired attributes and mechanisms are present in the concrete model. The constructed concrete model is expected to resemble the actual moon in the relevant ways on a smaller scale. In this example, we would expect the model to contain the craters of interest[44] such that we can count them.

Component 3 involves comparing the output of the model to the comparable output of the world. In this case, it is the obtained data from which the model was constructed. In the case of evaluating the model for a descriptive purpose, the modeller simply checks that all of the features or data they wanted the model to contain are, in fact, present in the model. Component 4 involves evaluating the model fit on the two dimensions—similarity and adequacy. The similarity-relation hypothesis assesses the manner in which the model is similar to the target system for the purpose of providing a description. Models being evaluated for a descriptive purpose have a heavy weighting of the overall model fit tied to the similarity relation. What are thought of as the important features to be described were identified in the *S(m,t)*, and then it is a matter of ensuring that those features, attributes and mechanisms, are in the model. The adequacy-for-purpose hypothesis informs us about the assessment of whether the model is qualitatively or quantitatively satisfactory for the purpose of providing a description at hand. In the context of a description, we simply need to determine that the overall data fit is present to the desired degree. For models that serve the purpose of description this can be understood as simply a matter of fitting, or accommodating, the model to the target system. The descriptive purpose is often simply a means of communicating or conveying scientific facts about the target system.

[44] Of course, smaller craters might not be represented accurately. If the modeller were interested in the model as an adequate model for describing those craters, then the similarity would need to be high enough.

From a model with high descriptive fit, such as the one in this example, not much can be inferred beyond what was explicitly included in the model during the construction. A user of the model can reliably extract descriptive facts about the target system, the moon, without having to perform actual observations of the moon. However, the model does not include the kinds of mechanisms that would allow one to make interesting predictions or give explanations. For example, if the modeller wanted to explain how the moon's craters came to be, this 3D model would not be adequate. Likewise, if the modeller wanted to predict what the future landscape of the craters would be on the moon if a meteoroid hit, this model would not be extendable. The only way to know *why* the model cannot offer such predictions, and such explanations, lies in evaluating the similarity relation relative to these purposes. With respect to prediction, it might be that the moon model is not similar with respect to the type of rock the actual moon is made of. For an explanation, the model has not included relevant past information about the moon as it has evolved over time. In general, the kinds of features included in a model constructed with just a descriptive purpose in mind do not include the mechanisms that are so crucial for providing predictions and explanations.

Consider again the success claim, "The scale 3D moon model is successful in that it accurately describes the positions of the Kepler and Copernicus craters relative to each other". This claim straightforwardly is about the adequacy of the description. The model is adequate for the purpose of informing the model-user of the distance between the large craters to a certain degree of accuracy. Given that it was constructed with a similarity relation aimed at a descriptive purpose, it has a high degree of similarity. The model-user is justified in inferring from the model that these descriptive features are true of the target system as well. It is through imposing my framework that these differences between assessing the adequacy and assessing the similarity relation are brought to the forefront.

4.3.1.2 Light Curves

A simple concrete physical model, such as the moon model, is one example of a model designed to serve a descriptive purpose. Another example of a model constructed with the main purpose of providing a description, seen frequently in scientific research contexts, is a model that is based on very limited data about its target system. In this case, the target system is empirically accessible only via observations that output a discrete data set. The model seeks to represent the target system by incorporating the data points appropriately. The purpose of the model is simply to provide a description of the target system based closely on the data points. The construction of the model may take the form of some kind of curve fitting.

In cases from observational astronomy, frequently the modeller will have access to only a very small, finite set of data that does not contain much information. For example, astronomers may obtain only an image and spectra[45] of a galaxy that cannot be interacted with directly. Often it is considered a huge success simply to describe the system, or some of its features, to some degree of accuracy. Models of this kind try to incorporate the data as closely as possible, but often have to balance that with theoretical considerations. As such, most models do include a certain amount of background theory, even if it is just about how to understand what the data represent. Recall that such background theory, in my framework, plays a part in informing the weightings in the similarity relation, *S(m,t)*.

An example of this kind of modeling from astronomy is the generation of light curves for celestial objects as a means for describing and cataloguing different types of objects. A light curve model graphs the light intensity at different wavelengths of a celestial object as a function of time. This can be used to model features of the system's rotation, its interaction with another system, the evolution of a system over time, or even the existence of a system unable to be directly observed, as is the case in exoplanet research. A success claim in this context might be, "this specific light-curve best fits the data points we have about this star system, and so we should consider it to be the best description of the system".

[45] Spectroscopy is the analysis of the spectrum of light emitted from a source. Different kinds of matter can be identified by their unique absorption and emission lines. Because of this, scientists can identify the kind of matter present in systems. This is particularly useful in astronomy because it can be used to identify the material constituents of stars, galaxies, interstellar gas, etc.

Looking at an example, consider an investigation of MT Ser, the central binary star system of the planetary nebula Abell 41 (Bruch et al. 2001). This will be the target system. Astronomers on this project were interested in determining what the period of the binary system might be, to provide a description of the system. Given this target system and purpose, the astronomers attempted to construct a model with the purpose of providing a description of the system over short timescales, which would include its period. The known attributes of the system are the obtained observations or data points about the luminosity[46]. Constructing a model such that these attributes are closely accounted for in the model will receive priority, and as such, receive a high weighting in our similarity relation *S(m,t)*.

There is some ambiguity in this case, as a light curve is often described as a model of the data. In my framework, a data set itself is never the target system. This is a case where the small data set represents everything known about the target system, in this case the binary star system. However, this real world system itself is the target of the model. That is, the model represents features of the target system. In this case, all the features that are available to be accounted for in the model are represented by the data about luminosity. In cases like these, the most complete descriptive model of the target system is coincident with the best data model of the data set.

Component 1 involves the construction of the model by incorporating the empirical data about the target system that has been obtained and attempting to address the known error that exists in obtaining the data. In the context of a light curve, each measurement of the luminosity of the object will contain error bars, which reflect known error such as interference of the light with the earth's atmosphere. For example, one procedure used to account for this error is a Gaussian regression process (Faraway et al. 2016; Spencer & Reese 2014; Chatzopoulos et al. 2012). This process uses a Gaussian curve under each data point to determine what the most useful curve will be, based on the location of the subsequent data point. There are theoretical motivations for adopting any

[46] A B filter was used throughout their observations in order to inhibit the contamination of the measurements by the strongest nebular emission lines (such as Hα) (Bruch 2001, 900).

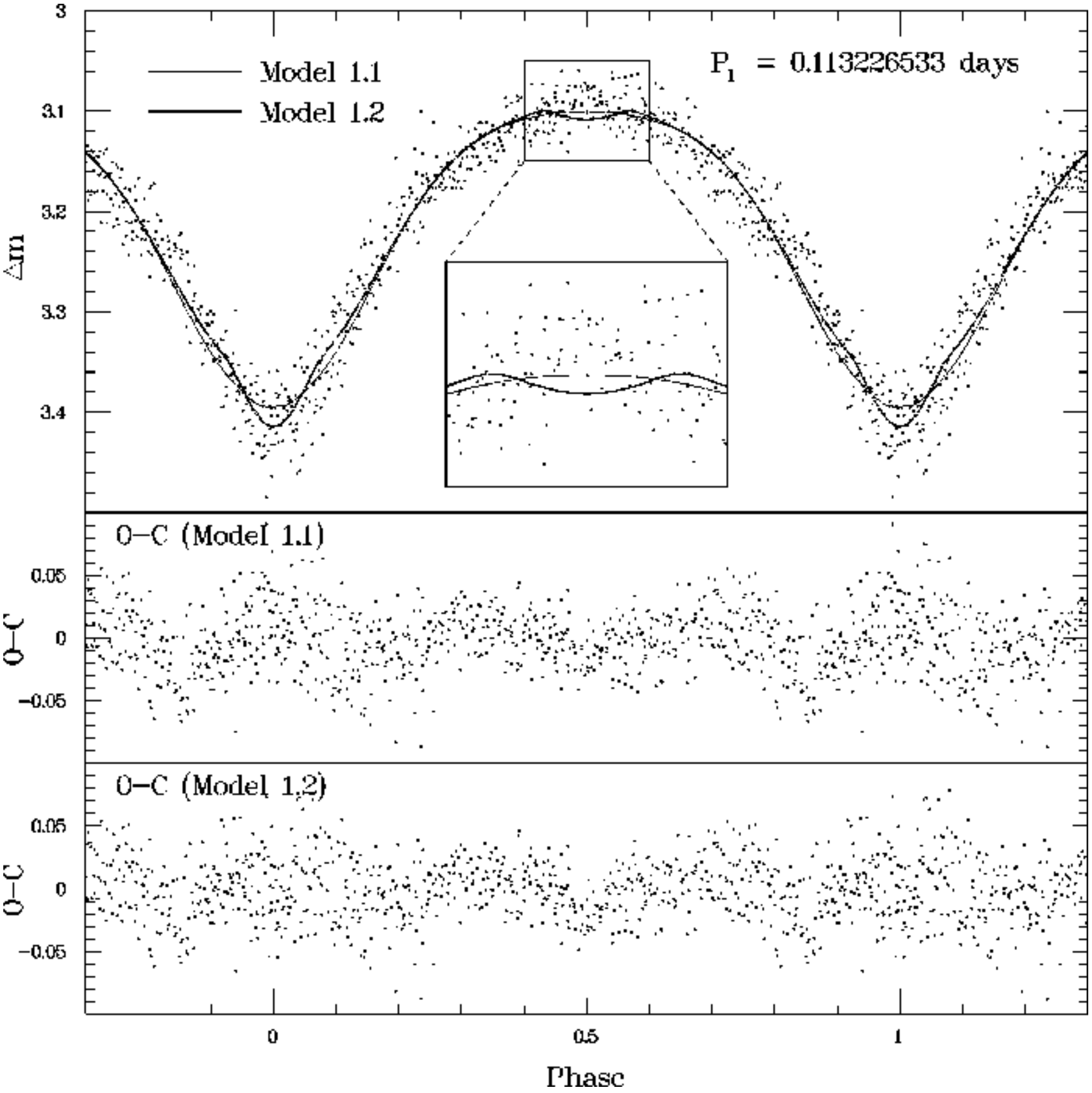
approach for accommodating error; these procedures, however, are always used to make the model as descriptively accurate as possible. That is, given substantive assumptions about the nature of the sources of error, one chooses different statistical algorithms for determining the “best fit”.

A very interesting feature of this example is that it is unknown what kind of stars are in the binary system. As a result of this, the modellers obtained two possible models, with the period dependent on what kind of stars are assumed to be in the system. That is to say, what is unknown in this case is a feature of the target system. In order to address this unknown, two models are constructed, each establishing a similarity relation relative to the two possible target systems. One possibility is a low temperature component orbiting around a hot sub-dwarf star, the second is two hot sub-dwarfs of similar temperature and luminosity (Bruch et al 2001). The modellers considered themselves to be constrained to these two possibilities by drawing on larger background theory related to star formation and various stellar properties considered to be well established. Each model represents one of the two possible real systems.

> We have analyzed the light curve of MT Ser, the binary central star of the planetary nebula Abell 41, within the framework of two quite different models. An unambiguous decision between the two models appears impossible in view of the current knowledge about the system. (Bruch et al. p. 909).

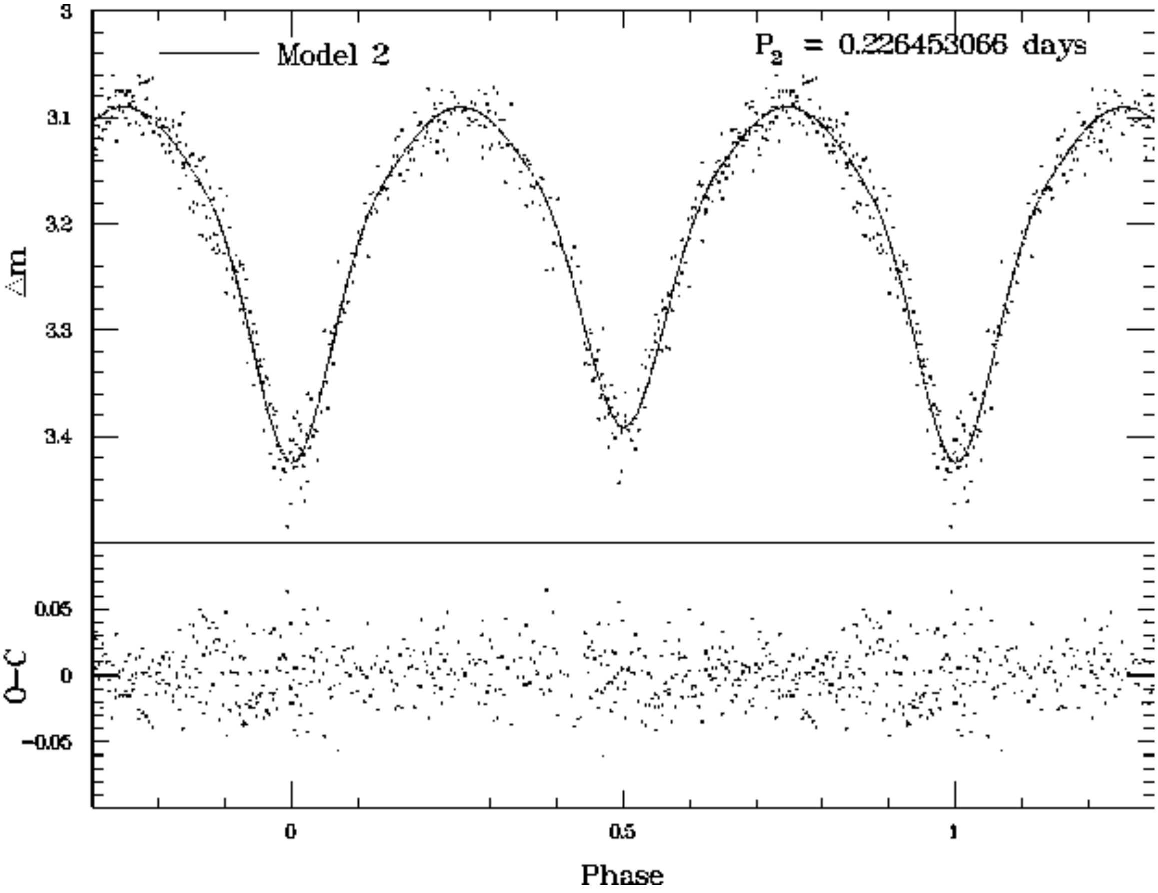
Figure 8 (below) includes the observational data points and the first model used to describe the possible target system if it is a low temperature component orbiting around a hot sub-dwarf star. Model 1.1 and model 1.2 are two possible descriptions using the same assumptions that differ only in minor respects. The period on this model is *P*=0.113 days. Figure 9 is the second model, of two hot sub-dwarfs of similar temperature and luminosity. The period for this case is *P*=0.226 days. These models are constructed by accommodating the same observational data.

Figure 8: Low Temperature Component Orbiting Hot Sub-dwarf Star.



Model 1. Top: Light curve data of MT Ser, and two different model light curves (solid and dashed lines). Center and bottom: O–C curve, or the difference between the observations and the model light curves (from Bruch et al. 2001, 901).

Figure 9: Two Sub-dwarfs, Same Temperature and Luminosity.



Model 2. Top: Light curve data of MT Ser, and model light curve. Bottom: O–C curve, or the difference between the observations and the model light curve (from Bruch et al. 2001, 901).

The output in component 2 is the curve that best accounts for the data points. In this example, there are two models that describe the period of the system. If the model is adequate for describing this target system, then we would expect to see a model that closely accounts for our data points. For the time being, I am not going to discuss which of these two models should be chosen over the other; this evaluation comes in at a later point.

For component 3, one would expect that the curve accounts for all the data points as closely as possible. What is likely to be observed, if the model is indeed adequate, is that any future observations of this system would match, and be described using one of these models. At this point, an assessment about whether these models are similar enough to the target system is determined by assessing the amount of error permissible. As seen in the bottom graphs of Figure 8 and 9, there is a numerical value of just how similar the line is to each data point; this relation is called the O-C curve.

Component 4 involves the assessment of the overall fit of the model. In this case, a descriptively adequate model will account for the data points. Both models, in this case, seem to be adequate given their respective target systems. What the astronomers are unable to say, given the current state of observations, is which one is the right description of the actual target system. In order to be able to assess this, they would need to obtain further observations to help narrow down their description of the system to know what the binary system is composed of. If they were to get definitive evidence that one of these two target systems is the real target system, that would verify the descriptive adequacy and similarity of one of the models. It is not a prediction of the kind of stars that need to be in the target system that is verified, but a description.

This is a case where the modellers are unsure which model accurately represents the actual target system. The first model (in Figure 8) includes, among its attributes in its weighted feature-matching equation *S(m,t)*, a low temperature component orbiting around a hot sub-dwarf star. The second model (in Figure 9) includes, among its attributes, two hot sub-dwarfs of similar temperature and luminosity. Since the models are based on different sets of attributes, they are different models.

What are the modellers justified in inferring about the system, beyond the observational data? Even if it is assumed that each of the two models has a very high fit, it is still uncertain which one accurately represents the real world system. Which model is actually similar to the target system is unknown. Therefore, the modellers cannot, with certainty, extend their model to make predictions about what would be observed through other means. All they can do is make conditional inferences, assuming one or the other of the models accurately represents the target system. If the astronomers were to discover which model is actually similar to the target system, then they could make inferences that go beyond the model.

Consider again the success claim, “this specific light-curve best fits the data points we have about this star system, and so we should consider it to be the best description of the system”. This should be understood as claiming that the best we can do, given the uncertainty about the actual target system, is provide a model that matches everything known from observational data. The model is adequate for the description of the data.

Analyzing this by breaking up our evaluation into an assessment of adequacy, and an assessment of similarity allows for this to be clear. Both models are adequate for the purpose of describing what is known about the target system. Depending on the actual nature of the target system, either model could turn out to have high similarity to it. However, since that nature is unknown, because it is underdetermined by the observational data, the overall situation is that there is not enough information to properly assess the similarity relation. Due to this state of epistemic uncertainty, there is not high similarity between either of the models and the actual target system. Since similarity is underdetermined, modellers are not justified in making inferences about the actual target system from this descriptive model.

One might question whether this model is also adequate for various predictive purposes. For example, if it were known that the target system was two hot sub-dwarfs, could this model predict future luminosity observations and the future period? There is a sense in which these future states of the period, and possible future luminosity data points, can be called a prediction. They amount to saying that the model, the description of the system, will still be adequate in the future. These claims gain their justification entirely from the legitimacy of the model’s descriptive adequacy and similarity. What these sorts of claims amount to is taking the model—which has been constructed relative to a descriptive purpose and assessed as adequate relative to a descriptive purpose—and extending it and applying it to a predictive purpose.

In general, it is possible to take a model that was created to serve a descriptive purpose and evaluate its adequacy relative to a different purpose, such as predictive adequacy. The similarity relation is held constant; the model was established to have a similarity relation relative to a descriptive purpose (component 1). However, we can then evaluate whether the descriptive model is adequate for a certain predictive purpose (component 3): for example, predicting future states of the target system. However, there might be cases in which it would be concluded that the descriptive model is not adequate for a predictive purpose, given that the model—during construction and formation of the similarity relation—was not made to include features that would be needed for making such a prediction.

To this point, models with similarity relations generated relative to a descriptive purpose can be accused of overfitting as a result of the attempts to accommodate a set of data. Such a model may run a risk of not being able to provide accurate predictions about future states of the system. However, this sort of critique fails to acknowledge is exactly the point that I wish to make, that prediction and accommodation aim at different purposes. While we would want a predictive model to avoid overfitting, the standard for a simply descriptive model is to fit the data appropriately. It is a mistake to evaluate a model for which the intended purpose is to describe by using the same standards that are designed to evaluate models that seek to predict.

Above I have detailed how to understand the evaluation of a model fit relative to the purpose of description. Based on this form of model fit evaluation, what is it that licences inferences back to the real world? The purpose of a descriptive model is to accurately represent the desired features of the target system and reproduce the relevant empirical data. Based on how we evaluate these models, we can see that our justification for making inferences from descriptive models is based on the similarity relation. Providing the details of the similarity relation is what allows us to identify in what ways the model has successfully represented the target system. Because the model is similar, it is a successful description.

4.3.2 *How Models Predict*

Very rarely is the only intended purpose of a model to describe the target system. Scientists often intend models to do more, so that they can learn about the world rather than simply systematize the chosen relevant empirical data. Often models are constructed to provide a prediction as an output. Ideally, models intended to serve the purpose of providing prediction seek to make novel predictions. This typically means more than simply reproducing the observations that were already used to generate or test the model's initial construction. The challenge for predictive models is that we cannot determine if their predictions are in fact correct until we obtain data against which to compare those predictions.

4.3.2.1 Binary Black Hole Merger

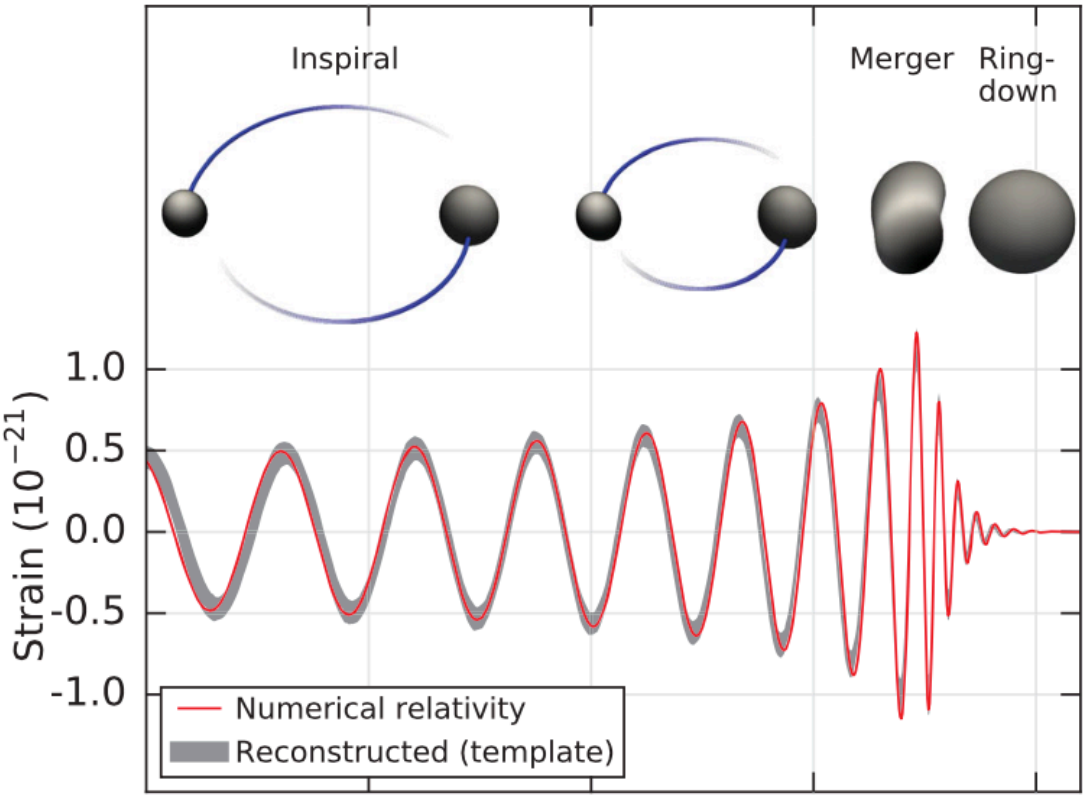
For an example of a model serving a predictive purpose, let us look again at models of black holes. In February 2016, the Laser Interferometer Gravitational-Wave Observatory (LIGO) scientific collaboration reported the first direct detection of gravitational waves, and the first observation of a binary black hole merger (Abbott et al. 2016, 061102-1). This research group, in their endeavour to detect gravitational waves, used a model of a binary black hole merger to predict what would be observed at the LIGO detectors. A success claim from this discovery might be, "The detected waveform matches the predictions of the waveform models for the spiral and merger of a pair of black holes, and the ringdown of the resulting single black hole"[47].

In this example, the target system is a binary black hole merger, and the model is evaluated relative to the intended purpose of predicting what happens when these two black holes merge. Evaluation at component 1 involves identifying what aspects of two merging black holes the model needs to include. Similar to the example in Chapter 2, the model intends to capture attributes such as the mass and spin of the black holes and the underlying mechanism of gravity as described by general relativity. The domain of applicability is to account for the time in which the two black holes merge. However, the model is not intended to capture an extended period of time before the merger, or afterwards (that is, the evolution of the black hole after the merge). This information establishes the similarity relation relative to the purpose of providing such a prediction of the merging black holes.

[47] While there is no explicit statement of the success of the model, this success claim is adapted from the results reported in Abbott et al. 2016.

Component 2 involves obtaining an output, which in this case is a prediction of what will happen during the merger. When the model was run, it provided a prediction of what the gravitational waves would look like when interacting with the LIGO detector.

Figure 10: Black Hole Merger

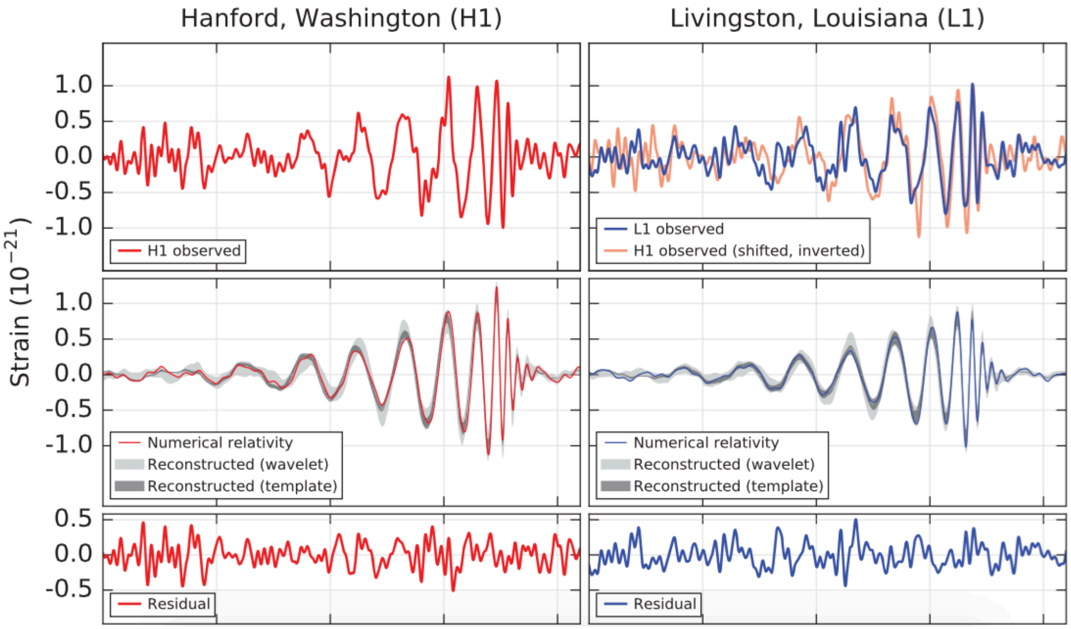


Computer model predictive output for a binary black hole merger output on detectors (Abbott et al. 2016, 3).

This was judged by the modellers to be a reasonable expected output, if the model were indeed adequate. This initial judgement was made based on reasoning related to the theory (what should be expected in general relativity) and assumptions made in the model construction process.

In September 2015, LIGO recorded observational data from their detectors of a gravitational-wave event. These detected strain readings were compared to their model's prediction. Component 3 involves determining how the model's prediction compares to the comparable real world output observed. In this case, the scientists were able to make a comparison to data, checking how well what is actually observed fits with what we are likely to observe if the model is adequate.

Figure 11: LIGO Detections



Top graph overlays observational data from LIGO detectors located in Livingston, Louisiana (L1) and Hanford, Washington (H1) of observed binary black hole merger. Bottom graph is the model output for a binary black hole merger (Abbott et al. 2016, 3).

The top graph in Figure 11 is the data from the two LIGO detectors. The bottom graph is the model output (same as Figure 10). Evaluative component 4 involves assessing similarity and adequacy. Based on error tolerance, modellers evaluate whether the model output matches the observed data closely enough to validate the prediction of the observable consequences of merging black holes. In this case, LIGO scientists were able to judge their model as being adequate for the purpose of providing a prediction.

This, in turn, provides justification for the similarity relation used during the construction phase for their model to serve a predictive purpose. These two assessments taken together constitute a good model fit. Some support for a fit was gained in components 1 and 2, and further support was gained once data was obtained at component 3. This allowed for a stronger justification to be made at component 4 relative to both dimensions, adequacy and similarity, of model fit. Should the observations have been different from the predictions of the model, the modellers would have known the model was not adequate for predicting the observable consequences of the black hole merger. This would have indicated that the model needed to be altered.

In this case, the similarity relation heavily weighted the theoretical consideration from general relativity. This is because the only reason for believing in the existence of the target system and the possibility for detection of gravitational waves is that these are consequences of the theory. A high degree of similarity in this case is due to the fact that the model is constructed from the background theory, rather than being constructed from direct observations of the target system. A consequence of this is that if the model had been shown to not be adequate for its purpose, then reassessing the similarity of the model to the target system could have had implications for the confidence in the theory itself. This is an example of a case where a model serves as an intermediary between the observations and the theory. And as I will argue in the next section, similarity is what justifies making other inferences. For example, the high similarity of this model justifies using it in giving explanations. Again, it is important to note that this similarity is grounded in our confidence in the theory. In cases like this, it is because the model is strongly theoretically grounded that it can offer explanations.

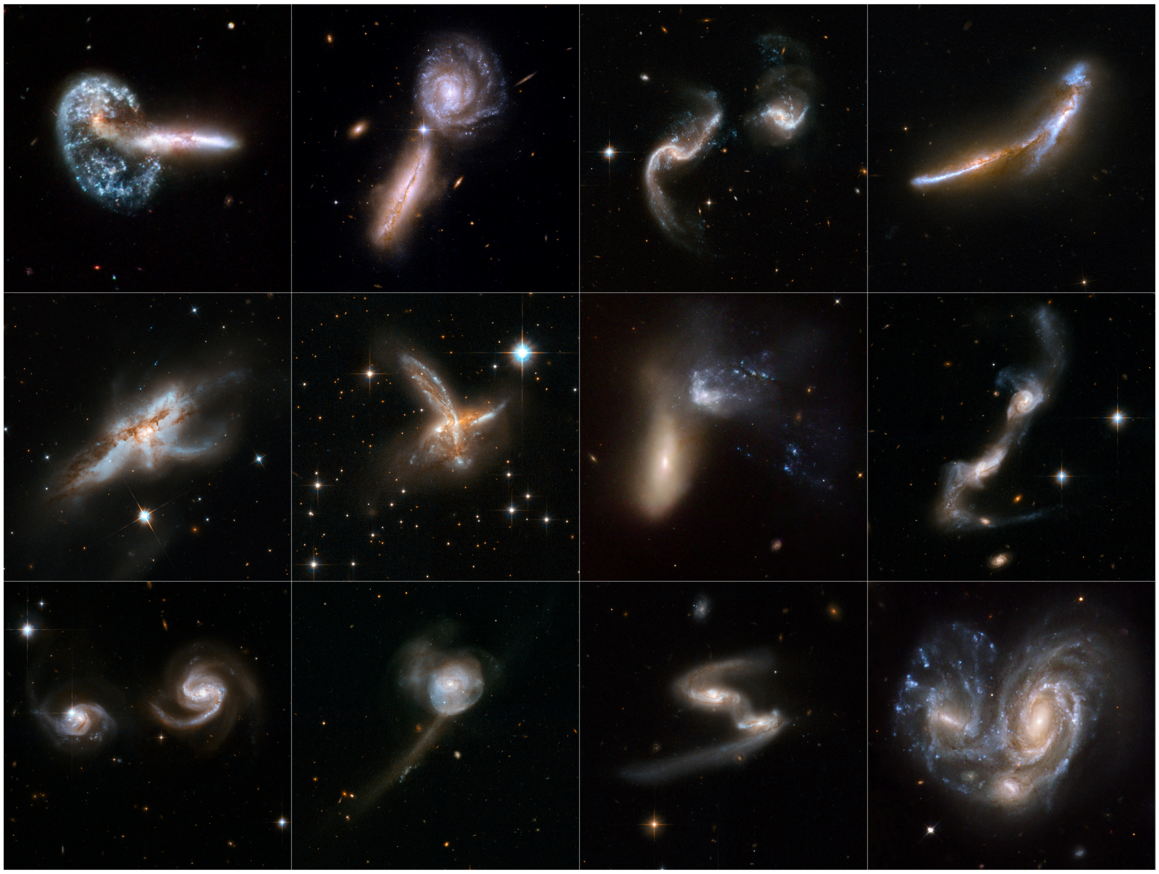
Returning to what can be learned from using my framework to evaluate success claims about models, such as, “The detected waveform matches the predictions of the waveform models for the spiral and merger of a pair of black holes, and the ringdown of the resulting single black hole”. This contains a component corresponding to an adequacy-for-purpose hypothesis: “the black hole merger model is adequate for the purpose of providing a prediction of the observable output of merging black holes”. However, the claim does not include a similarity component. This can be seen by the fact that it makes no claim about the extendibility of the model or its representational similarity to the system. However, in this case, as discussed above, the similarity is due in large part to the fact that the model is highly informed by theory. So inferences about the target system can be drawn from this model, and scientists should expect it to apply to other similar target systems.

When a modeller is able to provide reasoning relevant to the four components in my framework, this what I will call a full account. By providing reasoning involved at each component, there is stronger justification for claims about the model fit. However, in some sciences such as astrophysics, things do not always work out so well; more commonly scientists are not able to compare a model's prediction to data from the real world. And more importantly, when constructing a model that will be used to make claims about future states, modellers often know there is not the luxury of comparing the model's prediction to data. Yet the goal is to make claims about model fit even in the absence of additional observational data. We need to understand how to gain justification for claims about a model's fit in these cases as well. This will be the focus of the next example.

4.3.2.2 The Collision of Milky Way and Andromeda

In astrophysics, many of the target systems and events of interest take place over timescales much larger than we will ever have the opportunity to observe, and astronomers are only able to obtain a snapshot of an event. One such example is galaxy interactions—galaxies whose gravitational fields result in a disturbance of one another.

Figure 12: Colliding Galaxies



Snapshots of various galaxy interactions (image credit: NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University)

Astrophysicists are not able to directly observe the complete interaction of two galaxies colliding over time. However, they can investigate snapshots of interactions, such as those in Figure 12, based on obtainable observational data (e.g. current directional velocity of each galaxy) and constructed models in order to learn about these interacting systems.

The Local Group is a gravitationally bound collection of galaxies, of which our own home galaxy, the Milky Way, is a member. Observations indicate that Andromeda, a nearby galaxy, is moving towards our Milky Way at 110 kilometres per second (Cox et al. 2008). Astronomers want to predict if Andromeda will collide with the Milky Way and if so, predict how substantial the collision would be, as well as when this collision would happen. A means by which such a prediction can be made is through construction a model of this target system.

In what follows, I will present a model of the Andromeda and Milky Way (from Cox et al. 2008). My intention is to provide a case where a model's primary goal is to provide a prediction, but there is not the option of comparing the model against real world data. So no evaluation can be done with respect to component 3. I use this as an example of a partial account of model fit and for discussion of how claims about the world based on models, where comparison to the real world is not possible, are justified. Component 1 involves model construction and establishment of the similarity relation between the model and the target system. The target system in this example is the Milky Way and Andromeda galaxies. The prediction is of their trajectories and time of collision. The domain of application with respect to time is through the collision and on very large scales so that it is appropriate to abstract away from features other than the center of mass. Full discussion of the construction and the establishment of similarity relation is outside of the scope of this paper. What I have included is just a small subset of the details that are particularly relevant to my discussion[48].

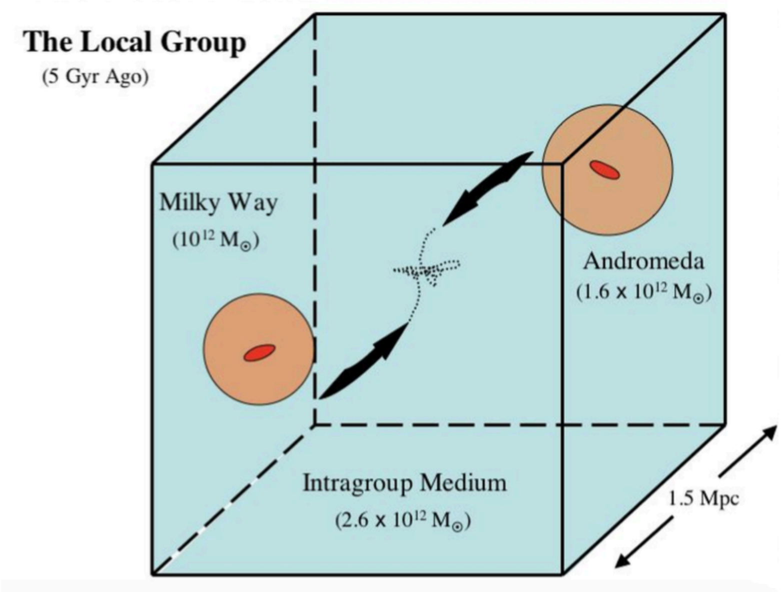
In constructing a model of two possibly interacting galaxies, determining the mass of these galaxies is critical to providing a prediction of the interaction. So is critical to include mass in the model. The modellers based their model on previous models of single galaxies, in which the baryonic matter (visible matter) is contained entirely within the rotationally supported exponential disc and central bulge, and is surrounded by a massive dark matter halo, which has nearly 20 times the mass of the baryons (Cox et al. 2008, 462).

[48] What is particularly interesting in this example is that the Cox et al. 2008 paper, from which this example is taken, contains an entire section in which the modellers detail what about the target system they see as essential, and what may not be as relevant or important to capture in their model for the purpose of providing the prediction they are interested in. What they have done is provided the explicit establishment of a similarity relation that takes place in relation to component 1. Should this model compete against a different model, divergences will be quickly identifiable.

There are also a number of observations that have already been obtained about these two systems, and so a model must satisfy all of these observational constraints. These observational constraints include the current distance of separation of the two galaxies, of 780kpc, their radial speed of 120 km s^{-1}, local circular speed (i.e. rotation of the galaxies) of 220 km s^{-1} (Cox 2008, 462). These observations are known within margins of error of less than 5%, and so the radial speed is considered to have an extremely high weighting of importance in the model. However, there are also features the modellers consider less well constrained, such as the present estimate for the transverse velocity of Andromeda, and spin orientation of the two galaxies with respect to the orbital plane of the merger. Cox et al. note that while such details are necessary if one wants to provide a model of the entire Local Group, such details are not as critical for influencing the merger between just the Milky Way and Andromeda.

While it is true that general relativity is our best theory of gravity governing the target system, the modellers have chosen, for the sake of calculability, to make the justifiable approximation that the galaxies are a simple two-body problem governed by Kepler's equations. This means that each galaxy is treated as a point-mass located at the galaxy's center of mass, and Kepler's two-body equations govern their mutual attraction. While the target system's underlying mechanism is gravity as described by general relativity, the model is governed by Kepler's laws. This is an appropriate approximation in this regime, and the predictions won't deviate significantly from those derived from general relativity. Nevertheless, this is one way in which the model and target system are dissimilar, which would likely be assigned a moderate penalty in a similarity relation. In the end, the system is idealized so that the initial configuration of our Local Group model consists of the Milky Way and Andromeda as a two-body system embedded in a 1.5 Mpc across cube containing a diffuse, constant-density intragroup medium of equivalent mass.

Figure 13: Milky Way and Andromeda Interaction



Visual representation of initial configuration the model (image from Cox et al. 2008, 463).

Having constructed the model, in component 2 we obtain the output, the prediction. In the context of this model, the output is obtained by evolving the model using a self-consistent two-body simulation. The output obtained from this calculation is generic properties of the merger, including the merger time-scale, the possible evolution of our Solar system, and properties of the merger remnant (Cox 2008, 462). The prediction is that the first close passage of Andromeda by the Milky Way will occur in approximately 2 Gyr, with the final coalescence occurring in less than 4 Gyr (4 billion years). For the evaluation of component 2, the modeller considers what is likely to be observed if the model is adequate for the purpose of predicting when the galaxies will interact. This process involves analysis of what sort of observations would indicate fit, such as actually observing a close passage of Andromeda in approximately 2 Gyrs—or, on a shorter time scale of 1 million years, Andromeda remaining on the trajectory the model has predicted.

However, in this example, it is not possible to obtain the comparison for component 3. There is no equivalent observational output from the real world to compare the model's output against. Without observing when the two galaxies in the target system interact, it is not possible to obtain the matching data to compare to the model's prediction. In this case, it is not possible to check how well what is actually observed fits with what we are likely to observe if the model is adequate. Yet it is assessments of model fit in these contexts, cases in which we do not have real world observations against which to compare the model output, that are most important. Predictive models are most frequently constructed in order to make claims about real world events before those events come to pass. Moreover, many sciences use predictive models as a means by which to determine what sort of hypothetical interactions with the system might result in a change of the predicted future outcome.

How do we assess the adequacy-for-purpose and similarity-relation hypotheses, and determine the model's fit in these cases? I will argue this is the role of robustness analysis and will provide further detail on this point below in §4.5. Briefly, robustness analysis involves comparing the output of a model to the outputs of other models in attempts to separate genuine predictions from outputs that are accidents of the construction of the model. Robustness analysis can be used in aid of supporting the output of the model involved at component 2, but also by supporting assumptions made in the construction of the model in component 1. The modellers in this example performed a robustness analysis by generating alternative models by altering the values of certain parameters.

> We have performed 20 additional runs in order to test the sensitivity of our results to various assumptions of our model – mainly involving the initial orbit of the MW and Andromeda. These runs yield similar estimates for the merger time-scale as well as for the possible locations of the Sun in the future, provided that the intragroup medium is indeed similar to our fiducial case. While this gives us some confidence that our results are robust, an even larger suite of models, that spans a much wider set of model assumptions, will provide better statistics on these results (473).

As seen in this example, the matter content of the two galaxies was based on other models about other galaxies. By gaining further justification for the elements in component 1 and 2, modellers can gain confidence in the output for these cases in which they cannot compare the model output directly to the future state of the system.

Moving on to component 4, for cases in which there is only a partial account, there is, to some extent, a weaker claim about model fit as compared to full accounts.

Nevertheless, the similarity-relation hypothesis can be partially assessed through the aid of the dynamical fidelity criteria, which help to determine if that prediction seems within a desired error tolerance. In the present example, this takes the form of assessing error with respect to the timescale of collision predicted. The representational fidelity criteria can be used in evaluating whether it is reasonable to think the model is making the right predictions for the right reasons. This part involves reflective assessment on justification for establishing the similarity relation in the way it was and on whether the level of similarity of the model to the target system still holds upon examination of model output.

With respect to the assessment of the adequacy hypothesis, the model can be considered adequate as long as it is consistent with other data obtained about the evolution of local galaxies over long timescales. An interesting feature of this example is that there is not any meaningful application of this prediction, given the extremely large timescale over which it takes place. Moreover, there is no real way we could interact with the system to change this outcome. Other cases of partial assessment of overall model fit may not have this feature—with climate models for example, modellers want to be able to see what sort of interventions could affect the output. In the case of the colliding galaxies, while a more refined prediction that captures more of the details would be more similar to the actual target system, a more precise prediction may not be necessary.

Returning to assessing the success of the model in my framework, the authors of the paper from which this example is taken state their achievement this way: "we quantitatively predict when the interaction and merger of the MW and Andromeda will likely occur" (462). Again, this claim is primarily an assessment of adequacy. The justification for the confidence in the accuracy of the prediction comes in part from robustness analysis (as seen above) and in part, from the similarity of the model to the target system for its purpose. It is because the model based on Kepler's laws is similar enough to the target system for the purpose of predicting the trajectories the galaxies will follow that there is confidence in the predicted timeframe for the collision. As the modellers note, this approximation is not appropriate at longer timescales, and therefore the model is not similar enough to the target system to extend the model and make predictions at significantly earlier times, "Note that extending the simulation to significantly earlier times is not adequate since stellar ages imply that the two Galactic discs (and presumably their haloes) have not been fully assembled at $z \gtrsim 2$" (466).

Through these two examples (LIGO and the Milky Way-Andromeda collision), I have detailed how to understand the evaluation of model fit in the context of a predictive purpose. In the cases in which there is a full account (LIGO being able to compare model output to real world output in component 3), what justifies the inferences made from the model to the real world system is partly tied to the assessment of adequacy for purpose. It is the fact that the model's output matches the real world, and therefore is supported as being adequate for predicting the features of the target system. However, if the model had not matched the observations, then the modellers would have known that they had not captured some important feature of the target system in the model, and would re-evaluate the features that made up the similarity relation. In this sense, the justification is also tied to the similarity relation. The similarity relation establishes that the relevant features of the real world system are represented in the model. If the model's prediction is then found in the real world, the model has the relevant similar feature located somewhere in it. It is in this way that a justification for inferences about the world from the model is established in the case of having a full account.

In the cases in which the modeller is only able to obtain a partial account (unable to compare the model output to the real world), we need to approach the justification for making inferences from the model's prediction slightly differently. If it is not possible to compare the model's prediction to data from the real world to see if the prediction is correct, then there needs to be an alternative way to build confidence in the prediction[49]. That is to say, when it is not possible to complete the component of reasoning that is related to assessing adequacy for purpose through comparing output to data, the goal

[49] Of course, we may want to build confidence in the prediction in the case of a full account as well. So this is not to say that robustness analysis is deployed only in the context of a partial account. Rather, evaluation in the context of a partial account relies heavily on robustness analysis to evaluate the obtained output in component 2.

should be to make the reasoning about the output as strong as possible. In §4.4, I will argue that one of the strongest ways to do this is through robustness analysis.

4.3.3 *How Models Explain*

Of the three purposes, the justification for explanations from models proves to be the most challenging to analyze. Within the philosophy of science literature, there are over four decades of discussion of what constitutes a scientific explanation (Salmon 1989). In the philosophy of modeling literature, extending discussion to how a model can provide an explanation has been a challenge as well. For my purposes, I propose to import Alisa Bokulich's (2011) account of how models can be understood as explanatory. Bokulich has provided a persuasive and detailed account of how models explain, which I describe below. I take her account to be the most general account for models, and so I use that account to detail an example of evaluating model fit in the context of an explanatory purpose.

4.3.3.1 Bokulich's account of model explanation.

Alisa Bokulich (2011) provides an account of the core features of a *model explanation*, the conditions under which it is reasonable to take models to be genuinely explanatory (2011, 38). She does this by critically analyzing three proposals in the philosophy literature for how models can explain—Craver's "mechanistic model explanations" (2006), Elgin and Sober's "covering-law model explanations" (2002), and what she calls McMullin's "causal model explanations" (1978, 1985)[50]. In light of her analysis of these three accounts, Bokulich provides what she considers to be a general framework of what these accounts all have in common, and therefore what should be the features or conditions of a model explanation:

1. The explanans must make essential reference to a scientific model, and that scientific model involves a certain degree of idealization and/or fictionalization.

[50] "Causal model explanations" is Bokulich's term for McMullin's view. However, his term for his own view is "hypothetico-structural".

2. The model explains the explanandum by showing how the elements of the model correctly capture the patterns of counterfactual dependence of the target system.

3. There must be a "justificatory step", in which we specify what the domain of applicability of the model is and show that the phenomenon in the real world to be explained falls within that domain.

With respect to her first feature, part of the goal in Bokulich's paper is to provide an account consistent with understanding models as fictions. In what follows I am not aiming to defend an account of models as fictions but rather to discuss the issues at play in a way that is not committed to this specific understanding of the ontology of models. I will treat the first feature as models involving some degree of idealization, but without the stronger commitment to a false model being a fiction.

Bokulich's second feature stems from an account of scientific explanation offered by James Woodward (2003). Woodward understands an explanation as providing information about a pattern of counterfactual dependences between explanans and explanandum, where counterfactual dependences can be understood as "What-if-things-had-been-different questions" (Woodward 2003, 11). Bokulich takes on board Woodward's account but tempers it. She thinks model explanations should seek a pattern of counterfactual dependence, but without the requirement of causal manipulation of the system. She considers it "a mistake to construe all scientific explanation as a species of causal explanation, … and it is certainly not the case that all model explanations should be understood as causal explanations" (Bokulich, 39). Elaborating on feature two, she explains that the elements of the model can be said to reproduce the relevant features of the explanandum phenomenon. The model should also "be able to give information about how the target system would behave, if the elements described in the model were changed in various ways" (Bokulich, 39).

The component of this account most relevant to my framework is the third condition. Bokulich take the "justificatory step" as specifying the model's domain of applicability. This step is intended to:

> draw explicit attention to the detailed empirical or theoretical process of demonstrating the domain of applicability of the model. In other words, it involves showing that it is a good model, able to adequately capture the relevant features of the world (39).

She thinks justification for the model's domain of applicability can proceed in two ways. The justification can proceed "top-down" from theory, in which an overarching theory specifies "where and to what extent the model can be trusted to be an adequate representation of the world" (30). However much more commonly, it proceeds bottom up, through various empirical investigations. The main result of this step is to distinguish between models that are genuinely explanatory and models that are merely phenomenological (39-40).

My reason for selecting Bokulich's account is because she argues that we can now recognize Craver, Elgin and Sober, and McMullin's accounts as "subspecies" of model explanations, each differing from the others with respect to where the "origin" of the counterfactual dependency lies. In addition to mechanistic explanations (where the model is the mechanistic parts which make up the explanandum-style whole), covering law explanations (where the explanandum is a consequence of the laws cited in the model), and causal explanations (where the model causally produces the explanandum), she suggests at least a fourth subspecies, structural model explanations (p. 40). In a structural model explanation, the explanandum exhibits a pattern of dependence on the elements of the model cited in the explanans, but this dependence is also a consequence of the structural features of the theory(s) employed.

I consider Bokulich's account as easily complementing my framework as the way to understand how models can explain and how this evaluation takes place. The first element to Bokulich's account involves the explanans making reference to an idealized or fictional model. I take my framework to be concerned already with models of this kind. However, it is not the case that the explanans make essential reference to a scientific model. Instead it makes reference to *the output of* a scientific model. These outputs will be relative to the context or type of model explanation; sometimes the output is a model structure, sometimes it is a prediction. It is these elements that will feature in the explanans of an explanation. Second, the model is to "explain the explanandum by showing the counterfactual structure of the model is isomorphic (in relevant respects) to the counterfactual structure of the phenomenon" (43). In the footnote to this point, Bokulich says she is using the notion of isomorphism loosely. I propose to read this as similarity. In order to make claims about the structure of the model, we need to have already detailed that structure. Furthermore, the model's ability to answer a wide range of "what-if-things-had-been-different" questions requires the ability to know what exactly was chosen to include in the model. Finally, Bokulich requires us to then justify that domain of applicability is an adequate guide to the domain of the phenomenon. Bokulich also sees the domain of applicability as central to providing the boundaries for which the model is to be applied. I have already indicated that these boundaries must be specified in the establishment of the similarity-relation hypothesis. Therefore, I will take Bokulich's account, with a slight modification to her first feature, to provide the details for evaluation of model fit in the case of models being used for the purpose of providing an explanation.

4.3.3.2 Precession of Mercury's Perihelion

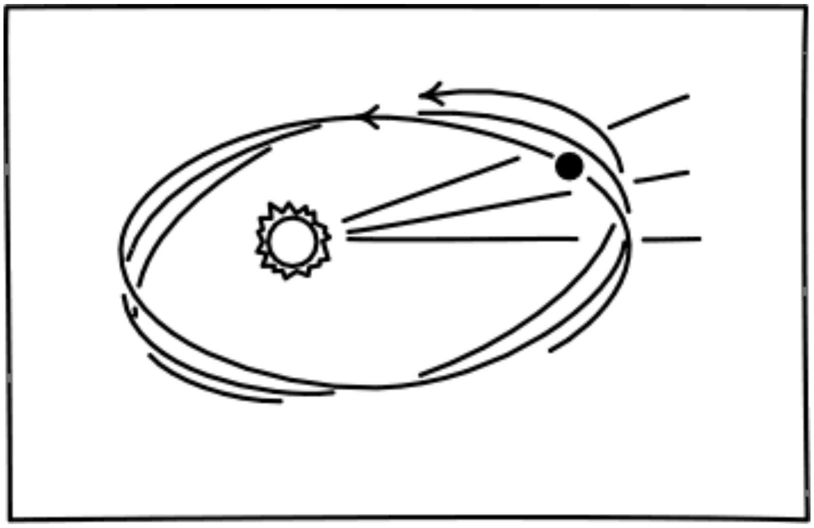
Turning to an example, I will examine an evaluation of model fit relative to an explanatory purpose for the precession of Mercury's perihelion. Under Newtonian gravitational theory, when a smaller mass orbits a larger mass, it will follow a circular or elliptical path. However, as Mercury orbits the Sun, it does not retrace the same elliptical orbital path each time, but rather it will rotate, or *precess*, over time. Astronomers say that the perihelion—the point on its orbit when the planet is closest to the Sun—advances. There are a number of effects in the solar system that might cause the perihelion of planets to precess around the sun, such as the gravitational attraction from other planets. However, the predicted precession, based on Newtonian mechanics and the influences of all other known planets in the solar system did not match the observed precession. Therefore, astronomers in the 20th century were interested in seeing if changing one of the underlying mechanisms, part of the gravitational theory, might produce models that explain the observed precession[51].

The target system to be modeled is the solar system, with a particular interest in Mercury and the Sun, for the purpose of explaining why there is a precession of Mercury's orbit. The salient features, attributes and mechanisms, that the model should capture for the purpose of explaining the orbit of Mercury include the mass of sun and planets, as well as the trajectories and properties of orbits. While the model is based on the Newtonian gravitational model, the modellers are interested in seeing if inclusion of a general relativistic mechanism might accurately explain the precession. The selection of these features makes up our component 1 and establishment of the similarity relation to the target system. The generated model is the set of mathematical equations that provide the planet's orbit. This mathematical model can be run over time, given that the domain of application of the model is over an extended period.

Component 2 evaluates the output of the model and considers what would be the case if the model were adequate. Figure 14 gives an illustration of what a precessing perihelion looks like, but it should be noted that the orbit is exaggerated.

[51] The other option was to maintain Newtonian dynamics, but posit an additional planet, Vulcan, between the Sun and Mercury. However, observations did not indicate such a planet's existence (Bertone et al. 2005).

Figure 14: Mercury's Perihelion



Precession of Mercury's perihelion (image from Norton, 2013).

The explanation for Mercury's perihelion precession is due to the dynamic nature of spacetime in general relativity. This is included as an element in the mathematical model, which is derived from theoretical considerations. The general relativistic mechanism for gravity ensures geodesic motion through a dynamic spacetime[52]. Because the curvature changes with the presence of the mass-energy of Mercury and the sun, the geodesic that the planet follows will precess on each orbit (and differ from the precession predicted in the Newtonian model).

The reason this model can serve an explanatory purpose is detailed by Bokulich's account of how models explain. This model is explanatory in that the explanans makes reference to the output of the model. The model explains the explanandum by showing how the elements of the model correctly capture the patterns of counterfactual dependence of the target system. There has also been a specification of the domain over which the model applies.

[52] Recall that for the similarity relation, the dynamical considerations in a model are captured in its mechanisms. One intuitive way to think about how this applies to general relativity is to consider the "shape" of spacetime to be the cause of gravitational effects. However, this is not essential and should not be seen as an endorsement of a substantivalist view of spacetime or a commitment to the idea that spacetime has causal powers. General relativity as the "mechanism" for gravity could be redescribed as an attribute.

Comparing the model output to real observations of Mercury in component 3 focuses on comparing the structure of the explanation to the elements of the target system. For example, if the explanandum admits of a causal explanation, the modeller searches for those causal dependencies in the target system. Identifying those dependencies in the real world shows that the explanation is adequate for the explanandum. Likewise, if the output of the model fits into a structural explanation, one searches for counterfactual dependencies or relations in the target system. In the case of Mercury's perihelion, we see that in the model the structure of spacetime, and its dependencies on the local matter content, exist in the real world target system.

Component 4 involves assessing the fit of the model for the purpose of providing an explanation. What it means for there to be a model fit is that the model produces certain outputs that feature in a model explanation in Bokulich's sense. The model is adequate for the purpose of providing an explanation in that we see similar dependencies in the real world target system. As seen in the case of prediction, the similarity relation also plays a role in evaluating the fit of the model as a model that explains, in that it ensures that we include relevant structures in the model in order to give explanations that fit. This is another case in which we have a full account.

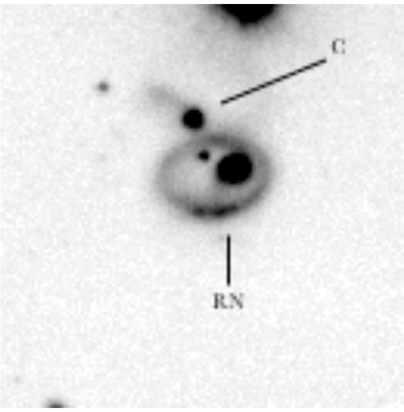
An example success claim in this case is, "The general relativistic model explains why Mercury's perihelion precesses at the observed rate". My framework allows for this to be analyzed generally as an adequacy claim, in that general relativity can provide an adequate explanation for the precession of Mercury. However, should one want to extend the explanation in this model, for example, to explaining why there is a precession in a different case, the modeller must rely on the assessment of different similarity relations. It is only if one thinks that the mechanism for gravity, and other relevant attributes composing *S(m,t)* for the case of Mercury's perihelion also hold for the new case, that one is justified in extending the model. For example, to explain the perihelion motion for another planet, it would be necessary to include the same mechanism for gravity as in the model of Mercury's perihelion motion. It is because the model has been assessed as having a corresponding similarity relation that the model can be extended in the right way.

4.3.3.3 Ring Galaxy Formation

However, as seen with predictive models, there are cases in which one can provide only a partial account, as it is not possible to compare the explanation to its observable consequences in the real world. In astrophysics, this happens frequently given that it is only possible to obtain a snapshot of an event through telescope observations. One such example is the case of galaxy collisions. This will be the focus of my example of evaluating a partial account with respect to a model's purpose of providing an explanation.

Most galaxies have relatively standard shapes, such as spiral or elliptical, though not all galaxies will fall into these categories (Mo et al 2010). These other galaxy shapes are often referred to as *peculiar* galaxies. One such galaxy is II Hz 4:

Figure 15: Ring Galaxy



Optical v-band image of II Hz 4. RN indicates the ring galaxy, and C the companion (image modified from Ramono et al 2008).

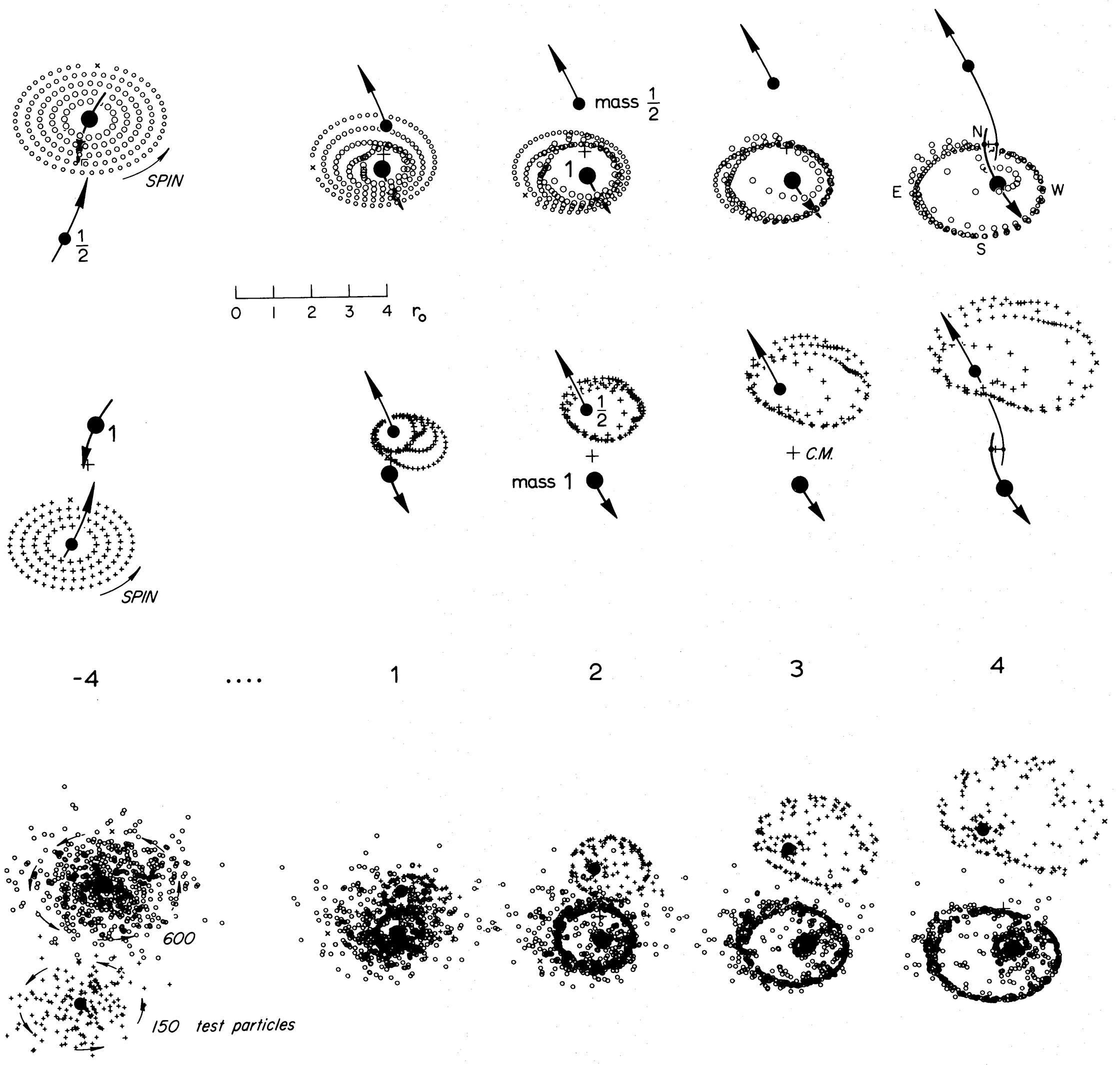
What is 'peculiar' in the image is that the main galaxy of interest seems to have a clear center component, as well as ring shape around it. Galaxies like II Hz 4 are now categorized as collisional *ring galaxies*. Galaxies of this shape are somewhat rare, and so astronomers want to figure out why these galaxies are shaped this way.

One possibility is that this was a spiral galaxy, but the companion object (above II Hz 4 in the image) collided with it. As a result of this collision, the galaxy somehow formed its ring shape. The problem, however, is that astronomers do not have access to past observations to determine if these galaxies did look different at an earlier time. Furthermore, galaxy collisions take place over significantly long time scales; so it is not possible to directly observe the different states of a collision. The only means by which to investigate this target system is to construct a model in order to determine what likely happened.

For Component 1, a model of the target system, II Hz 4 and its companion, is constructed, and the similarity relation established. From slit spectroscopy observations of the galaxy and its companion, astronomers obtain spectra of both of the galaxies' nuclei. From the observed spectral lines, they determine the radial velocities of the two galaxies at the present time. These observations provide a starting basis for what a reasonable similarity between the model and the target system will be.

A quick account of the similarity relation is as follows: For constructing a dynamical model of these galaxies, one needs a model that has similar orientation, and spin to the target system. There is however no need to capture the exact trajectory of every particle in the model; one must simply provide a general how-possibly account. The model then, does not need to have the same number of particles as the target system; so this dissimilarity will not be penalized. The target system galaxies are idealized in the model so that they are composed of only a few particles. While it is possible to estimate the actual mass of the galaxies in the target system, the model allows the masses of the galaxies to vary as a means of exploring how the two galaxies might interact, thus allowing for a kind of robustness analysis. Allowing the mass in the model to vary will help in developing an explanatory model. As such, the model failing to share the attribute of the mass with the target system will not receive a heavy penalty. The underlying gravitational mechanism for this model will be softened short-range gravity (Lynds and Toomre 1976, 387*)*. Softened short-range gravity is an approximation of general relativity that has been shown to apply in this domain (Lynds and Toomre 1976). While this is not similar to the target system, given that the purpose of the model is to provide an explanation, such gravitational representation will be a sufficient approximation.

Figure 16: Particle Models of Ring Galaxies



II Hz 4 and its companion idealized as a small galaxy hitting a simplified disc galaxy head on. Top: collision in which companion is ½ mass of main galaxy, Middle: two galaxies have equal mass, Bottom: interaction as 4:1 ratio of test particles (image from Lynds and Toomre 1976).

At component 2, the model's output is obtained (Figure 16). The purpose of the model is to provide a how-possibly explanation of why these galaxies are shaped the way they are. As seen in Figure 16, in all three situations a ring galaxy is formed from the collision. The explanation then is a simple dynamical mechanism for ring galaxy formation. The companion galaxy falls through the disk galaxy head on, and the interaction results in a ring-like density wave. The output of the model is among our explanantia. If our model is indeed adequate for the purpose of providing this explanation, one would expect to see exactly what is seen here: given the constraints, a ring galaxy forms.

In evaluating component 3, one compares the model output to the real world. However, astronomers only have a snapshot of the end of the interaction in the real world. In cases in which one cannot compare the model's explanation to the real world, and have only a partial account, ideally one would need to check that the model's explanation applies to all relevantly similar target systems. For these cases, we will again see in §4.4 that robustness analysis will lend support. As noted above, varying the masses of the two components of the model allows for a limited kind of robustness analysis to be achieved very easily.

Component 4 is the assessment of the overall model fit. In assessing the model's adequacy, one can make a partial comparison of the model output, to the output of the real world phenomena in that there are similar radial velocities of the system. Considering that in this case the aim is only exploring a how-possibly explanation, the model's adequacy is assessed relative to being able to provide just that. Since this is a how-possibly model, it is much more permissive with respect to what can count as a good explanation. All that explanations of this kind are meant to show is one physically plausible evolution that leads to the desired final state. The similarity relation can gain some support through this comparison of a feature of the model output to similar features in the real world as well.

With respect to assessing the similarity relation, the claim is weaker, given that we do not have enough information from the real world against which to compare our model's explanation. However, through support from the similarity relation, we can judge the model to be providing an adequate explanation. It provides an explanation in that the formation of the ring galaxy is sensitive to the relative masses, orientation, and spin of the galaxies involved in the collision. If these parameters were changed, the final output stage would be different. This shows the counterfactual dependency between the assumptions of the model and the final state of the system.

An example success claim for this case is, “this model provides an explanation for why galaxies, such as those seen in Figure 15, have the peculiar shape that they do”. My framework allows for a claim such as this to be understood firstly as an adequacy claim. The model has been developed so that it provides an account of what happens when two galaxies with certain masses collide head-on, which in turn provides an adequate how-possibly explanation. A model that provides an explanation often is easily extended to provide adequate predictions and descriptions, given that by the nature of its construction, it captures many of the same features in the *S(m,t)* that would need to be present for other purposes as well. In the end, however, if the claims in the model are to be extended to claims about the real world, then the modeller must consider the similarity relation of the model in *S(m,t)* to hold close enough to the target system to justify this extension.

More generally, in evaluating model fit relative to the purpose of providing an explanation, what is it that justifies us in inferring that the explanation from our model is true of our target systems? The answer is similar to the case of prediction, in that what justifies the inferences made from the model to the real world system is tied to the assessment of the similarity relation. The similarity relation establishes that the relevant features of the real world system are represented in the model, and if the model’s ability to provide an explanation also explains the real world phenomenon, the implicit assumption is that the model has the relevant similar feature built in. It is only in this way that one can establish a justification for inferences about the world from the model.

To summarize the lesson of these examples: Models serving a descriptive purpose seek to account for, or accommodate, empirical data and provide an output that somehow represents the features present in the target system. Models serving a predictive purpose seek to provide novel predictions about the system, and produce an output corresponding to a future or past state of affairs about the target system that is not originally built into the model. Models serving an explanatory purpose seek to establish underlying mechanisms and structural dependencies, and produce an output that can serve as an explanans in an explanation of some phenomenon.

I have provided details about how to evaluate models relative to their intended purpose to describe, predict, or explain within my framework. A model is rarely assessed relative to only one purpose. Many models are intended to describe our empirical data at the same time as providing predictions, or explanations. Ideally, a model would be adequate to some degree for all purposes—description, prediction, and explanation. However, it is important to assess the model's successes relative to the different dimensions separately. Models can be extremely successful for one purpose, and therefore scientifically useful, even if they fail with respect to a different purpose. One of the most important features of modeling is that is allows for a distribution of cognitive labor. When investigating a target system, one model can provide extremely effective predictions yet fail to provide an adequate explanation. And a different model can provide an effective explanation, yet fail to provide an acceptable prediction[53].

Finally, I return to a problem identified in my examples: In cases in which it is not possible to obtain a straightforward 'full account'[54], how else can a modeller gain confidence in the model's purpose-relative output when there are not acceptable analogue real world outputs to compare against? Some (Lloyd 2010; Weisberg & Reisman 2008; Weisberg 2006; 2013) have argued that robustness analysis bears the weight of further justifying our models. I examine this claim and how robustness analysis might be able to help strengthen inferences from the model to the world. While robustness analysis can build confidence in the model's output, in the end, it is not the definitive solution. Rather, the similarity relation again helps by identifying what is being modified and prioritized through robustness analysis and the construction of various models. Namely, the similarity relation identifies what the modellers consider to be similar common features identified by robustness analysis.

[53] For a discussion of cognitive labor as it connects to models, see Bokulich 2013, Muldoon and Weisberg 2011.

[54] This is also relevant for cases in which, though a modeller can compare their model output to an analogue real world output, there may be reasons they still want to gain further confidence in the model's output.

4.4 The Role of Robustness Analysis

Many philosophers and scientists have appealed to robustness analysis as a method for determining whether a model's output is the result of something essential to the target system or an accident of the assumptions and idealizations made in constructing the model. Biologist Richard Levins, for example, characterizes the process of robustness analysis this way:

> [W]e attempt to treat the same problem with several alternative models each with different simplifications … . Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies. (Levins 1966, 20).

Likewise, philosopher William Wimsatt understands robustness analysis to achieve the following:

> [A]ll the variants and uses of robustness have a common theme in the distinguishing of the real from the illusory; the reliable from the unreliable; the objective from the subjective; the object of focus from artifacts of perspective; and, in general, that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting (Wimsatt 1981, 128).

The main idea in conducting robustness analysis is to compare several different models, each built using somewhat similar, yet distinct assumptions. If these models all have a similar enough prediction as their output, or if all identify a similar common feature, then that prediction or feature is considered well supported, or *robust*. Given that several models with different assumptions or features all provide the same output, one should consider this to be evidence that the feature is likely to be a feature of the real world as well.

Weisberg (2006, 2013) has argued for the importance of robustness analysis in how we learn through models. To date, Weisberg has provided the most detailed account of robustness analysis, and thus his is the account I will examine. In the next section, I first detail his account of how to use robustness analysis in the search for robust theorems. I then discuss the role robustness analysis and robust theorems play in providing grounds for inferences from models to claims about the real world. I examine the ways in which robustness analysis can be helpful for cases in which we do not have all components, a full account, of model evaluation. Robustness analysis in some cases can give reason to think model-outputted predictions or explanations may be adequate, that is, when there are not data from the real world to compare against. In these instances, robustness analysis may provide further evidence that the output of the model can be used in the similarity assessment, and that the robust properties stand in the similarity relation to properties of the target system.

4.4.1 *Weisberg's Account of Robustness Analysis*

For Weisberg (2006, 2013) the aim of robustness analysis is to separate the scientifically important elements and predictions of the models from those that are accidents of the representation. He provides a four-step account of how robustness analysis is embodied in the search for robust theorems (Weisberg 2006; 2013).

> Step 1: Examine a group of models to determine if they all predict a common result, a *robust property*.
>
> Step 2: Analyze the models for the *common structure* which "generates" the robust property.
>
> Step 3: Combine steps 1 and 2 to formulate the *robust theorem*. Robust theorems take the form of a conditional statement,
>
> Robust Theorem: *Ceteris paribus* if [common structure] obtains, then [robust property] will obtain.
>
> Step 4: Conduct stability analysis of the robust theorem to determine what condition will defeat the link between the common structure and the robust property.

With respect to step one, it is important to compare a set of models that are similar, yet distinct. It is important, Weisberg argues, that there be a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the models analyzed (2006, 737; 2013, 158). In order to generate these varied models we can, for example, construct related models of the same target system but vary the idealizations made. Another option could be comparing models of similar target systems. Thinking of the San Francisco Bay – Delta model, if one wanted to see how the bridge might respond to certain weather conditions, one could compare the model to other models of other bridges.

The first step is followed by, or conducted in parallel with, the second step. Step two, Weisberg argues, involves finding the core structure in this set of models that generates, or gives rise to the robust property. This step can happen in a range of ways, depending on the models at hand. He considers the simplest or most straightforward cases to be those in which the common structure will be the models having the same physical, mathematical, or computational structure. Harder cases, however, involve "models that are not developed in the same mathematical or computational frameworks, or may represent a similar casual structure in different ways or different levels of abstraction" (2013, 158). These cases will rely heavily on the theorist's ability to judge relevantly similar structures.

> In the most rigorous cases, theorists can demonstrate that each token of the common structure gives rise to the robust behavior and that the tokens of the common structure contain important mathematical similarities, not just intuitive qualitative similarities. However, there are occasions in which theorists rely on judgment and experience, not mathematics or simulation, to make such determinations. (Weisberg 2006, 738)

Weisberg considers step three to be where we obtain an empirical description from our first two stages, which contain only formal or mathematical information. A fundamental aim in modeling is to move from information about the model, to empirical claims about the world. Weisberg considers step three to be where this takes place: "the third step of robustness analysis involves interpreting the mathematical structures as descriptions of empirical phenomena" (2006, 728). It is this analysis that produces our robust theorem. Finally, step four determines the extent or limits of the theorem's robustness. There will be conditions under which a model will no longer generate the robust property. Modellers determine what these limits are through existing data and further empirical investigation[55].

For Weisberg, the reasoning process that gives us good grounds to believe predictions and explanations of robust theorems involves determining: 1) how frequently the common structure is instantiated in the relevant kind of system, and 2) what defeats the core structure giving rise to the robust property (2013, 159-60). Weisberg considers the first question to be best settled empirically. However, it can also be answered using techniques associated with robustness analysis. With respect to the second question, he thinks there are three kinds of robustness analysis that allow us to investigate different ways that the core structure can be defeated. *Parameter robustness* involves examining what happens when the values for the model description's parameters are varied. In this case, the modeller intends to examine to what extent the change of parameters changes the behaviour. *Structural robustness analysis* involves adding new mechanistic features to the model in order to examine how parts of the causal structure represented in the model produce different behaviours or properties in the model (2013, 162). *Representational robustness* involves representing the mechanistic features of the model in a new representational framework. While parameter and structural robustness aim to analyze how variation in mechanistic attributes affect the model, representational robustness aims at investigating if the way in which the attributes are represented affects the production of a property of interest.

4.4.2 *Discussion*

There are three points to be made with respect to how robustness analysis fits within my framework, as well as the extent and limits to which it might aid inferences made from models. The first relates to how I think one should understand the "similar structures" that are identified by robust theorems and their relation to the *S(m,t)* similarity relation. The second relates to the extendibility of robustness analysis to cases where the purpose of a model is to provide explanations rather than predictions. The third relates to whether

[55] The inclusion of and attention drawn to step four is what Weisberg considers Sober and Orzak (1994) to have missed in their criticisms of robustness analysis.

robust results have any epistemic significance, and acknowledging there may also be two different targets of evolution with respect to robustness analysis.

With respect to my first point, step two of robustness analysis directly appeals to similarity between models or judgements about “relatively” similar structure. What needs to be clarified is that Weisberg is concerned with identifying similar structure as similarity relates only to the features that have been included in the model. Step two also highlights the importance of establishing the similarity relation between the model and the target as a component in the process of evaluating model fit. In detailing the pragmatics of model construction, modellers are explicit about what structures were chosen and incorporated in the model. By knowing what structures are in a model they can more easily compare and determine what structures are common across many models. This allows for a clear understanding of what sort of robustness analysis is being completed (be it parameter, structural, or representational). In this sense, robustness analysis is useful in that it helps the modeller learn about what features, attributes and mechanisms, might be in the target system, and so what features might have an effect on the output the model generates.

With respect to my second point, the way in which Weisberg presents the process of obtaining a robust theorem discusses only the cases in which the model’s purpose is providing a prediction, and as presented, does not work for explanatory models. In his account, in order to obtain a robust theorem, it seems that we must have a model that provides a prediction, and we must be able to identify common structures. Recall that structure in this case is just the formal mathematics, computation, or concrete structure. A robust theorem is what allows us to connect predictions of a model to its structure, and ultimately provide explanation for the prediction. This is not the same as a model that serves an explanatory purpose. However, Weisberg’s process for obtaining a robust theorem can be extended to models serving an explanatory purpose in the sense developed above. In the case of a model serving an explanatory purpose the output is used as an explanans. Therefore, in the case of obtaining a robust theorem, we are concerned with the robust property being is included in the explanans. Weisberg’s step three then is concerned with establishing how the explanans is connected to the structure and providing an explanation for the way in which the model features in the original explanation.

Robustness analysis identifies similar structures in models and shows that the models all produce a similar output. While this can help one gain confidence in the fact that the output of the model is being produced for the right reason, it does not necessarily permit making inferences from the models to the real world. This leads to my third point related to whether robust results or outputs of the model have any epistemic significance. Wendy Parker (2011) also examines this by asking under what conditions an inference from robustness, to likely truth of the output, is justified. She argues that while there are conditions under which robust results may have special epistemic significance, it is not always the case that they do—it will depend on the ensemble of models and hypotheses (Parker 2011, 584).

In order to determine if a robust results has special epistemic significance, Parker argues that one must be able to provide the following argument (2011, 584):

1. ***Likely Adequacy Condition***: it is likely that one simulation[56] in this collection is indicating correctly regarding hypothesis *H*.
2. Each of the simulations in this collection indicates the truth of *H*.

∴ It is likely that *H*.

The most challenging aspect to making this argument, Parker thinks, is determining if there is good evidence that the *likely adequacy condition* has been met. There are at least two approaches for making the argument that this condition is met. The first focuses on the ensemble construction. On this approach one must argue that, “an ensemble of models samples so much of the current scientific uncertainty about how to represent the system (for the purpose at hand) that it is likely that at least one model produced is indicating correctly regarding *H*”[57]. That is to say, if the ensemble of models is generated

[56] In this context, Parker uses “simulation”, however in the context of my discussion this is equivalent to a model.

[57] Parker notes that this is similar to Michael Weisberg’s claim: “The key comes in ensuring that a sufficiently heterogeneous set of situations is covered in the set of models subjected to robustness analysis” (2006, 739).

with the concern being how to construct and represent the system of study for a particular purpose, with a large enough ensemble, it is probable that one of the models produced is likely be an adequate representation (largely due to the goal being constructing a sufficiently heterogeneous set of models).

The second approach focuses on ensemble performance. On this approach, an ensemble of models is viewed as a tool for indicating truth or falsity of a hypothesis, in which past performance of the ensemble is cited as evidence that it is likely one of the models in the ensemble is correctly indicating *H*. That is to say, if an ensemble of models is used to provide a certain output, such as a predicted future value or state of the system, then it is likely that one of those models provides the correct value, and the rest provide an output within some specified distance from that value.

What I take from the analysis offered by Parker is that there is a distinction in robustness analysis that could benefit from the distinction I have made with respect to target of evaluation being a similarity relation or adequacy-for-purpose. I take it that, for Parker, arguments that focus on ensemble construction are comparable to a target of similarity, in that what is being evaluated is representation of the system by the models. Ensemble performance, on the other hand, seems to be concerned with further assessing adequacy-for-purpose, in that what is being evaluated is the ensemble of models serving as a tool. Therefore, it is important to specify whether robustness analysis is being deployed to determine whether the model accurately represents the target system, verses robustness analysis concerned with determining whether a model will be adequate for a particular purpose.

Furthermore, there is an important difference between the goal of robustness analysis in the way Weisberg discusses it and the way Parker discusses it. In order for robustness analysis to proceed, Parker must assume, following Orzack & Sober (1993) and Woodward (2006), that if the ensemble of models covers the possibility space of ways to represent the target system, then one of the models in the ensemble will be the correct representation. For her, robustness analysis is a way of dealing with uncertainty about which model in the ensemble is the best fit. On the other hand, Weisberg does not make this assumption. He talks about robustness analysis as a way of discovering structures that ought to appear in the correct representation of the target system. However, no model in the ensemble, as constituted at any time, must accurately represent the target system. For Weisberg, robustness analysis supports the conclusion that the modeller is identifying a causally salient feature of the target system for the purposes of making a certain kind of prediction, or explaining the output of a model.

Robustness analysis can be of use to modellers in my framework in the following two ways: in cases in which there is uncertainty with respect to the constitution of the target system, but the modeller does have access to data to which she can compare the output of the model, identifying a robust feature of a class of models can give her confidence that the feature is present in the target system. In cases where the constitution of the target system is available (a similarity relation has been established), but the output of the model cannot be compared to the relevant aspects of the target system, then a robust theorem can give the modeller confidence in the prediction (or explanation) of the model, because it aids in identifying causally salient features of a relevantly similar class of models. In the former case, the feature of the target system is robust across an ensemble of models that produce the same output. In the latter case, the output of the model is robust across an ensemble of models that are all appropriately similar to the target system.

Robustness analysis and the identification of robust properties provide information only about the models themselves. In order to infer that the identified structures are real, an additional justificatory step is needed. That is to say, robust theorems identify the relationship between a property that holds across a class of models, and the desired output. This is the distinction Parker gestures towards—assessments of the adequacy are much more complex. If this output helps support an adequacy claim, then one can say that the presence of the robust property is part of the reason that the model is adequate for its purpose. However, as argued above, there is still no licence to make inferences about structures in the real world from this adequacy claim. Those kinds of inferences follow only if one believes that the robust property represents something in the target system. Again, this is a similarity claim.

As a final point, this is not to say robustness analysis is useful *only* in instances in which we do not have real world data to which one can compare a model's output. Robustness analysis can also further help a modeller understand what features impact the model's output, in cases in which there *is* real world data to compare the output to. In both cases robustness analysis can aid in understanding how the model works, and how the model might relate to the target system. However, in cases in which we do not have real world data to compare to the output of the model, robustness analysis can be particularly helpful in that the modeller learns more about why the output from the model is obtained.

4.5 Conclusion

Above I have detailed how evaluation of model fit is conducted when a model's intended purpose is to provide a description, prediction, or explanation. Justifying claims about the model's fit is tied to the similarity relation in that it makes explicit how the modeller constructed the model to be similar to the target system given a certain purpose. It also justifies our evaluation of model fit based on adequacy for purpose, in that we directly compare the model output to the comparable real world output. These two elements combine to give us the full evaluation of the fit of a model relative to a purpose. What follows from this is that our justification for inferences about the real world from a model must be relative to the purpose of the model as well.

Yet if one has a model with a certain similarity relation relative to a certain purpose, be it descriptive, predictive, or explanatory, and it is evaluated as adequate along that same dimension of purpose, this does not mean that it is necessarily going to succeed along the other dimensions. Therefore, we should not think that it is appropriate to make inferences about anything other than what we have determined to be appropriate to that dimension of purpose. To evaluate a model is to evaluate it relative to a purpose, and for the model to fit, or be successful, is to fit or be successful relative to that purpose. What we can do with models is make inferences from the models about the real world, but this will be constrained to a way in which the model reflects certain features of reality and not others through the established similarity relation at construction.

The similarity relation is critical because it provides justification that allows for determining when a model should not be extended or if the model must be modified in order to serve a different purpose. The importance of capturing the pragmatics of the modeller's choices in the similarity relation is that it allows for examining what features are relevant for the evaluation relative to one purpose. It also allows for comparisons with features to be included in the model with respect to alternative purposes. If it were judged that the model would have to include other attributes and mechanisms to serve the second purpose, then we would know it would be inappropriate to apply that model without making those changes.

Chapter 5

5 Tracing the Path of Justification for ΛCDM and MOND

5.1 Introduction

In contemporary cosmology and astrophysics, the Lambda Cold Dark Matter (ΛCDM) model is considered to be the current best model of large-scale structure formation that is in general agreement with observed phenomena. None of the parameter values in the model are fixed by our current best theory; instead their values are determined based our observational evidence. One of the most notable features of the ΛCDM model is the inclusion of dark matter[58]. According to the model, over 84% of all matter in the universe is dark matter —matter that is currently unobservable at any electromagnetic wavelength (Komatsu 2011; Bertone 2005). Part of the reason astronomers believe that there must be so much unobserved matter, is because galaxies rotate much faster than they would if all the matter they included was that which we can see. That is to say, the predicted rotation speeds given the mutual gravitational attraction of all the luminous matter in a given galaxy are much slower than what is observed. There must be more matter than we can see.

However, some astrophysicists have seriously questioned whether positing that 84% of the mass of the universe is made of matter we have never seen is a justifiable move. Some of these critics have proposed alternative models, which solve the problem of the galaxy rotations by MOdifying Newtonian Dynamics so that the missing mass is not required. The MOND approach is viewed as contentious, as many astrophysicists consider general relativity to be well established as the theory of gravity (Dodelson 2011). They therefore regard the adoption of modified Newtonian dynamics as unjustifiable. However, advocates of MOND claim that their models have as good of a fit as ΛCDM for describing observed galaxy dynamics (Milgrom 1983; Famaey &

[58] Of course, dark energy is another unusual feature of the model. However, in this chapter I will be focusing discussion on dark matter.

McGaugh 2012; McGaugh 2014; 2009)[59]. Part of the goal in this chapter is to analyze this claim. I will argue that the claim is accurate as long as it is understood as being relativized to a specific purpose and specific domain. The most important question is whether there is genuinely a conflict between the two models. My framework offers the tools to analyze this question and specify where the conflict arises.

This is a case in which there are two models that include different elements, and even differ fundamentally in terms of the theories on which they are based. Yet, both models have been evaluated by scientists to be models that successfully describe the observations, make adequate predictions, and even offer explanations. They are both said to be high-fit models. But how can that be if they disagree about fundamental physics? How can we make sense of a claim like this? How should we deal with situations in which we have two models that seem to contradict one another, yet are both evaluated as having a good fit?

At first glance, this may seem like a simple fundamental disagreement and that it is a mistake to evaluate MOND models as having a good fit, given that they do not use general relativity, our current best theory of gravity. However, the framework I have developed will allow us to see how both ΛCDM and MOND can be considered well-justified, high-fit models—given different choices about what to prioritize. While some scientists regard this as a case of Kuhnian incommensurability or believe that there is a purely subjective choice to be made (McGaugh 2014), I argue that they are not seeing the debate as it should be understood—as one primarily about choosing the purpose of models and then assessing whether they are useful for that purpose.

In this chapter, I compare the justification for ΛCDM to justification for MOND. I show how both of these models are well justified, high-fit models with respect to their adequacy for certain purposes. I will show how MOND models have high fit with respect to descriptive and predictive adequacy. However, they do not have as good a fit with

[59] Proponents of MOND view their models as is equivalent or superior in some respects to ΛCDM. For the sake of argument I am granting the advocates of MOND their success claims.

respect to explanation, as a result of their inclusion of a different mechanism for gravity. While high similarity of mechanisms is not necessary for description and prediction, I will argue it is necessary for explanation.

My evaluation is a result of paying careful attention to the role theory plays in the construction of the two models. My framework offers a toolkit for analyzing these sorts of debates so that an understanding of the disagreement can be reached. I conclude that both models can be evaluated as having good fit, when considering their fit within their respective domains of application. I claim that the apparent conflict between the two models arises due to extending both models past the domain where they are successful. In attempting to extend the model to new domains, the modeller relies heavily on a model's explanatory fit. And extending claims of explanatory fit relies on strong commitments to the similarity relation established, particularly with respect to the way the model represents theoretical commitments.

Returning to the overall goals of this dissertation, this case study will show how my framework deals with complicated cases in which there are multiple purposes involved in assessment of both similarity and adequacy. This is a case where the way in which theoretical considerations come into play during model construction is critical. In the end, this case study will show the benefit of understanding model assessment and justification in the way I have described. It will also emphasize how this framework can be an effective tool for discussions about fit between seemingly conflicting models.

5.2 Construction of the ΛCDM Model

The study of exact solutions of Einstein's field equations is one area of research in astrophysics, and the ΛCDM model is one exact solution[60]. While in this dissertation I

[60] The astrophysics literature refers to FLRW models as a class of *exact solutions* (rather than class of models) to Einstein's field equations specified by the FLRW metric. That is, any specification of parameter values, such as curvature (Ω_k) and mass-energy density (Ω_m), that is consistent with the field equations is referred to as a solution to the model (Hamilton 2014). In this context, the ΛCDM model is a parameterization of the perturbed FLRW models. However, the common usage of terms differs in philosophy. The reader should understand the FLRW solution as a set of models; any specification of parameters yields a particular model, such as the ΛCDM model.

have been referring to ΛCDM in the singular, it is actually a class of models. The particular parameterization of ΛCDM that incorporates all of the best empirical data is the model that I have called and will continue to call “the ΛCDM model”. Presently however, I will briefly discuss the class of ΛCDM models in general, which includes this special parameterization.

The ΛCDM models are derived from the class of perturbed FLRW models. The standard FLRW models make the assumption of homogeneity. This entails that there will be no structure present in the FLRW models. The perturbed FLRW models do not make the assumption of homogeneity. Rather, they allow for the presence of inhomogeneous regions of higher matter density. The ΛCDM models incorporate this feature and show how it can evolve through gravitational attraction into large-scale structure. The ΛCDM models differ from the FLRW models in another important way: they include further specification of the kinds of matter that exist in the model. This includes baryonic matter, radiation, cold dark matter, and a cosmological constant Λ.

The particular parameterization of ΛCDM based on empirical data includes the best estimations of the overall curvature (Ω_k) of the universe, mass-energy density (Ω_m), and other parameter values. This is considered the best model of the universe, and will be discussed in further detail below. Its ability to account for large-scale structure formation is considered one of its successes. However, as will be seen below, it has a challenger in MOND when it comes to accounting for behavior on the scale of individual galaxies. In this section, I will detail the general construction of the ΛCDM model, highlighting the main assumptions and idealizations, as well as discuss the domain and purpose for which it was constructed.

5.2.1 *Einstein Field Equations to FLRW*

Our current best theory of gravitation is general relativity. General relativity provides a unified description of gravity as curvature in spacetime, in which the curvature of spacetime is related to the energy and momentum of the mass and radiation present. This relation is specified by the Einstein field equations (EFE), a system of partial differential equations. The field equations are ten equations that describe gravitational interaction as a result of spacetime being curved by matter and energy. By providing conditions, or assumptions, for the global properties of the spacetime within general relativity theory, one can reduce these ten equations to tractable, usable coupled equations (Weinberg 1972; Misner-1973; Hamilton 2014).

General relativity theory is extremely well established. Any model that required the abandonment of general relativity would be a radical departure from the current best understanding of the structure of the universe. General relativity theory has been extremely successful in predicting gravitational phenomena at certain length scales. Astrophysicists have considered it justified to extrapolate beyond the observable success of general relativity and to posit that general relativity is the correct theory of gravity on cosmological scales as well.

Astrophysicists are interested in applying the field equations in order to find some way to model the structure of our entire universe. Since gravity is expected to be the dominant force on large length scales, models of the evolution of the universe at this scale are based on general relativity. The target system in the construction of such a model is the entire universe, but the domain over which the model is to apply involves only very large distance scales. The model would not be required to properly capture the inclusion of smaller features. That is to say, the astrophysicists are not a concerned with capturing details such as stellar populations within galaxies; rather the model is to be restricted to regularities above the length scale of galaxies, but below that of the Hubble radius (Hamilton 2014, 72). The intended purpose of such a model is to provide a mathematical description of the evolution of the universe, and the growth of large-scale structures over extremely long periods of time. The model is to serve a descriptive purpose, but also embody some predictive purpose as well, given that astrophysicists want to make claims about the possible evolution, as well as past and future states of the universe (Hamilton 2014, 83).

In order to develop tractable mathematical models for describing the global characteristics of the entire universe from the field equations, assumptions need to be made. Of course, these must also be justifiable idealizations or approximations of the target system according to the accepted construal for astrophysical models. Two assumptions, which render the simplest class of models from the field equations, are that the universe, on very large scales (i.e., > 100 Mpc), is homogeneous and isotropic (Ryden 2003). The assumption of homogeneity states that, regardless of location in the universe, the mass-energy distribution is uniform on large scales, and so spacetime has a uniform geometry. Spacetime can be foliated into three-dimensional hypersurfaces of constant time, so that any point on the surface has the same spacetime geometry. The assumption of isotropy says that there is no preferred direction in space. This means that on large scales observers will have similar observational evidence, regardless of the direction they look. These two idealizations are jointly referred to as "the cosmological principle", and are considered to be supported by a variety of observational data (Hamilton 2014, 71; also see Lahav 2001; Hansen 2004; Beisbart 2010; Maartens 2011).

These assumptions provide the large-scale smooth metric, the Friedmann–Lemaître–Robertson–Walker (FLRW) metric. The metric provides a means by which distances can be measured. It is also assumed that spacetime can be treated as a perfect fluid[61] (Wald 1984; Rugh & Zinkernagel 2011; Melia 2015). The motion of points through this fluid is used to represent objects such as galaxies. By applying these idealizations and approximations to the field equations, and taking into account symmetries in the equations, we arrive a class of models referred to as the Friedmann–Lemaître–Robertson–Walker (FLRW) models[62].

5.2.2 *FLRW to ΛCDM*

The FLRW models describe a universe with two unknowns (Hamilton 2014; Ryden 2003). The first is a global scale factor, *a(t)*, for the universe, and the second is the constant curvature of the universe. The scale factor is not directly observable, and its value must be determined indirectly from observations, similarly for the curvature. Many

[61] This is the assumption that the stress-energy tensor is that of a perfect fluid.

[62] The class of FLRW models includes any model related to the FLRW models, including the perturbed FLRW models.

contemporary astrophysicists see their work as efforts to determine what the values of these unknown variables are. There are various methodologies, observations, and supporting models and simulations that go into attempts to determine the values for these parameters (Spergel 2015; Hamilton 2014). To this extent, this work aims to determine features of our target system in an effort to further refine the model used to represent the target system. From the observable part of our universe, our observational data from various independent sources (such as WMAP, BOOMERanG and Planck) indicates that the curvature of the universe is almost perfectly flat (Giannantonio 2010; Komatsu et al. 2011; Ryden 2003). However, the global scale factor for the universe has more uncertainty.

A global scale factor is a function of time and represents the relative expansion of the universe. However, the differential equations for the scale factor *a(t)* depend on the content of the universe, which is parameterized by various cosmological parameters: Ω_m, the average matter density (including matter of particles and dark matter) that undergoes dilution with the scale factor; $\Omega_\Lambda = \Lambda/3H^2$, where $H=\dot{a}/a$ is the Hubble expansion rate, and Λ is the cosmological constant, a constant that does not dilute with the scale factor; and Ω_k, the average curvature of the universe (Hamilton 2014; Ryden 2003). These parameters are related through $\Omega_k = 1 - \Omega_m - \Omega_\Lambda$. A negative value corresponds to an infinite hyperbolic universe, a positive value corresponds to a closed spherical universe, and zero is an infinite flat Euclidean universe.

Since the goal of ΛCDM is to describe structure formation, it requires a slight alteration to the standard FLRW matter-density assumptions. FLRW has a uniform matter density. However, with a uniform distribution of matter gravitational effects would be acting equally from all directions on all points. This is a uniform model that lacks structure. In order to generate large-scale structure, ΛCDM needs to assume that there are random perturbations in the matter density distribution at very early times in the evolution of the universe. The regions with higher than average matter density will attract matter from surrounding lower density regions, leading to a concentration of matter that will become the large-scale structure of the universe. This evolution is what is governed by the linearized field equations in this model. The use of these equations is justified because, on the scales relevant for ΛCDM, gravity is weak, meaning that the spacetime metric is nearly flat (Wald 1984, 74). The standard Big-Bang model of cosmology does not give rise to the necessary density perturbations. ΛCDM requires matter density perturbations, which are inconsistent with the assumptions of the Big-Bang model. One popular way to generate these inhomogeneities in the model is by including an assumption about an inflationary period during the early universe to generate the density perturbations that allow for the formation of large-scale structures (Hamilton 2014, 72).

Through obtaining observational data about our universe, astrophysicists are able to narrow down the values for these cosmological parameters. The model with parameter values that are in agreement with those observations and proposed attributes in our target system is called the ΛCDM model of cosmology. The name, ΛCDM, comes from the two attributes that have been included in the model (and as such determine a certain range of permissible parameter values). 'Lambda' refers to the inclusion of Λ as dark energy, a form of energy that is thought to permeate all of spacetime and accommodates our observations that indicate the expansion of the universe at an accelerated rate. 'Cold Dark Matter' refers to the inclusion of cold dark matter in the model in order to accommodate our observational data about matter content, and structure formation. Further discussion of these parameters will be reserved for the following sections.

5.3 The Evaluation and Justification of the ΛCDM Model

Returning to my framework, I first want to consider the similarity relation established in the construction of the ΛCDM model. The underlying mechanism for the model is gravity as described by general relativity, and is included by virtue of using the field equations as the model's base. In order to construct a model from this, various idealizations and approximations have been made about the target system: the structure of the entire universe, over an extremely large-scale domain. In addition it is assumed that the matter density distribution deviates from uniformity, such that there are regions of higher density that act as "seeds" for structure formation. The weak field approximation states that the linearized field equations can be used in cases where gravity is not overwhelmingly strong. This approximation is appropriate at this scale; so the linearized field equations are used to model the evolution of these seeds, which over time give rise to large-scale structure.

By examining the later determination of parameter values, one feature becomes quite clear: gravity as a mechanism in the model has been given an extremely high weighting. As astrophysicists attempted to learn more about the target system, in order to provide possible parameter values in the model, many situations were encountered in which the modellers were required to include new attributes about the target system, such as the existence of dark matter and dark energy, in order to maintain the use of general relativity as the fundamental underlying mechanism[63] in the model.

The values for the parameters, as well as the idealizations made along the way, were adopted in order preserve a descriptive similarity relation in the model in which the highest weighted feature for the similarity relation is general relativity as the mechanism for gravity. What this illustrates is that the model is based on the theory of general relativity. The ΛCDM model will be assessed as having a high degree of similarity to the target system under this particular similarity relation weighting. This is due to the fact that the component of the weighted feature-matching similarity equation corresponding to the mechanism for gravity will have a very high weighting. Other dissimilarities between the model and the target system will not lead to a high penalty. Over the course of the construction of the ΛCDM model, astrophysicists have been continually obtaining new data and using it as a comparison point with observed phenomena to further refine the model, and make the model fit better with observations of the target system.

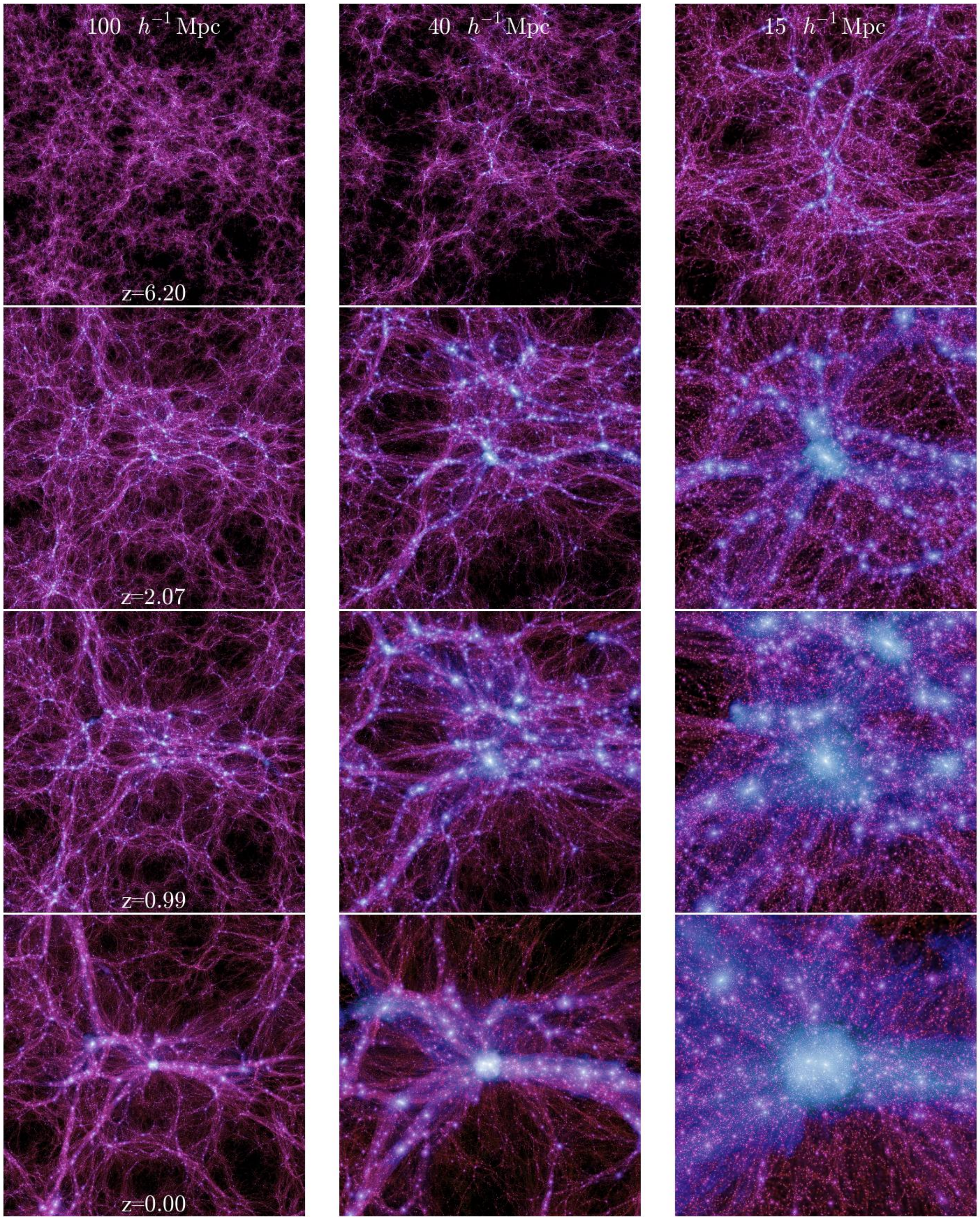
Having constructed the ΛCDM model for a descriptive purpose, we can use it to obtain a descriptive output. Recall, that in order to obtain an output, reasoning or computation related to the model must occur. One such attempt to obtain an output from

[63] As a reminder, “mechanism” in this context refers to the dynamical considerations in a model. One intuitive way to think about how this applies to general relativity is to consider the “shape” of spacetime to be the cause of gravitational effects. However, this is not essential, and should not be seen as an endorsement of a substantivalist view of spacetime or commitment to the idea that spacetime has causal powers. General relativity as the “mechanism” for gravity could be re-described as an attribute.

the ΛCDM model is the Millennium Run I and II[64]. The Millennium Runs are the largest ever computer simulations of the formation of large-scale structure as described by the ΛCDM model. The Millennium Run offered a visualization of the ΛCDM large-scale structure description of the matter content of the universe from 13 Gyrs ago through to the present, as well as its future evolution. The Millennium II simulation used the ΛCDM model with parameter values in best agreement with observational evidence at the time: $\Omega_{\mathrm{tot}} = 1.0$; $\Omega_{\mathrm{m}} = 0.25$; $\Omega_{\mathrm{b}} = 0.045$; $\Omega_{\Lambda} = 0.75$; $h = 0.73$; $\sigma_8 = 0.9$; $n_{\mathrm{s}} = 1$, where h is the Hubble constant at redshift zero in units of 100 km s^{-1} Mpc^{-1}, σ_8 is the rms amplitude of linear mass fluctuations in 8 h^{-1} Mpc spheres at z = 0 and *n*s is the spectral index of the primordial power spectrum (Boylan-Kolchin et al 2009, 1151). Running the simulation involved following 21603 particles within a cubic simulation box of side length Lbox = 100 h^{-1} Mpc. Each simulation particle has mass of 6.885 × 106 h^{-1} M⊙ , and particles were allowed to have individual adaptive time steps. The goal in the simulation is to evolve these particles in accord with the ΛCDM parameter values and the linearized field equations, to represent the evolution of the regions of higher mass density (or “seeds”), in order to show how the ΛCDM model describes structure formation evolution over time.

[64] Millennium Run II and Millennium Run I both have the same cosmological parameters and particles, but Millennium Run II has a box that is smaller by a factor of 5 than Millennium Run I and thus has 125 times better mass resolution.

Figure 17: Millennium Run II



This set of 12 images shows the evolutionary growth of a massive halo over cosmic time From top to bottom, the regions are plotted at redshift 6, 2, 1, and 0. The 3 columns from left to right show the evolution on different length scales. 100 x 100 Mpc/h, the center column is 40 x 40 Mpc/h, and the right is 15 x 15 Mpc/h (in comoving units) (Boylan-Kolchin et al. 2009).

In considering the similarity of the model to the analogous output from our target system, however, this is a case in which there is not full access to observing the entire large-scale structure of the universe nor the timescales on which to observe such things. While astrophysicists cannot make a direct comparison of the output structure from the model to the universe, they can, however, compare the model output to smaller scale systems, such as galaxy clusters, in order to see if these structures are consistent with the output of the model. Some large-scale features can be observed, such as galaxy distributions, and these are similar to the output of the model. However, the output the modellers are most interested in is the description over time of large-scale structure. Since this cannot be directly observed, indirect observations, such as those of smaller scale structures, must suffice.

In general, when a model is constructed so that it includes a strong theoretical component, then it is very likely that the model will be universally applicable. Since highly confirmed fundamental physical theories are to apply at all times, then a model closely based on them should also apply at all times. This suggests that a model with a strong theoretical component will likely be successful at making predictions. Even if the model was constructed with a descriptive purpose in mind, it will allow one to make claims about future and past states about the system. While a fundamental theory is *valid* in any domain, it likely will not be universally *useful*. Approximations and idealizations are made when constructing a model for a domain that differs from the standard domain of the theory. These approximations and idealizations may build in a domain dependence for the model that may not exist for the theory itself[65]. An example of this is the Oppenheimer-Snyder black hole model seen in chapter 2. While the model was based on general relativity, Oppenheimer-Snyder idealized away features of the target system that are relevant on longer timescales (such as spin). This allowed them to create a useful model for their purposes, but it was not applicable outside of the domain that was determined by their idealizations and simplifications about the target system.

[65] This is especially true for cases of non-fundamental theories that apply only in a limited domain.

Turning to assessment of adequacy of the ΛCDM model for various purposes, having a strong theoretical component in the establishment of the similarity relation for the construction of a model serving a descriptive purpose means that the description should apply at all times. Therefore, the ΛCDM model is also (in a sense) a predictively adequate model, as providing a description of a future or past unobserved state is a kind of prediction for which adequacy can be assessed. Furthermore, a heavy theoretical component in the similarity relation also means that the model should be able to be applied easily to unobserved cases, yielding new predictions in this sense as well.

Yet the ΛCDM model is an incomplete case, as we cannot compare the full structure output to the real world, even though we can compare some features of the model output to the real world, such as smaller scale distribution of galaxies. Astrophysicists do not have direct observational access to the large-scale structure of our universe, nor the timescales relevant to the predictions. They can, however, compare the model output of structure formations to smaller scale objects, such as galaxy clusters, and see if these models seem to be similar. On this basis, astrophysicists have judged the ΛCDM model to be predictively adequate as well. It has successfully predicted a variety of empirical consequences ranging from galaxy distributions on large scales, to lensing phenomena caused by dark matter halos surrounding galaxies (Mo et al 2010).

Most importantly, such a heavy inclusion of theory in the similarity relation, and in the construction of the model means the model fares well with respect to its explanatory adequacy. Because the ΛCDM model is based so closely on general relativity, it has already included the structures that are similar to the causal dependencies that exist in the target system. It has already included these explanatory elements, rather than needing to use the model to discover them. When a model is based closely on theory, in general, it will be explanatorily adequate because we take theories to offer good explanations[66].

[66] When a model is not based closely on theory, robustness analysis is an indispensable tool to aid in discovering these dependencies.

Even though the ΛCDM model is constructed for a descriptive purpose, in that the parameter values were chosen to accommodate observational data about the target system, it is adequate with respect to prediction and explanation as well. The reason for this is its close reliance on a well-confirmed theory in establishment of the similarity relation in the construction. In general, a model closely based on a theory will predict and explain well. Within the domain of the ΛCDM model—application to extremely large scale structure of the universe—the model can be evaluated as fitting well, with respect to its similarity to the target system for the purpose of providing a description, and it is adequate with respect to providing that description, as well as being able to provide predictions and retrodictions about the structure, and explaining why the structure is as it is.

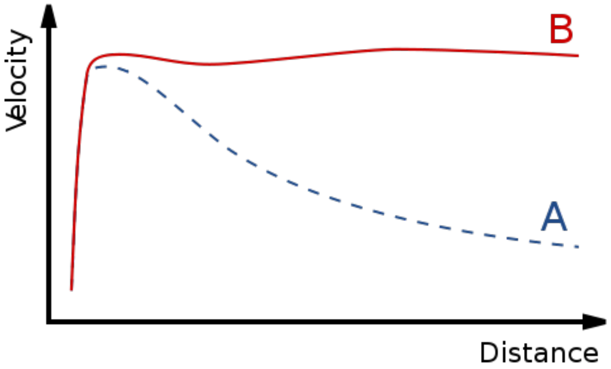
5.4 The Extension of Domain and Rise of Conflict

Since astrophysicists take ΛCDM to be an adequate explanatory model, it is quite natural to want to extend the model to other domains. If a model is explanatorily adequate, and we consider the explanatory mechanisms to be not just adequate, but fundamental to the similarity relation in construction of the model, and actually similar to the target system, then the model should have some explanatory adequacy with respect to other domains.

For this reason, while the ΛCDM model was constructed to apply on a large-scale domain, it could be extended to a different domain, such as the small-scale structure of a single galaxy. For the domain of small-scale structure, the field equations, on which the model lays its foundation, reduce to Newton's law of gravity by using both the weak-field approximation and the slow-motion approximation, which are considered justified approximations for single galaxies or clusters.

This allows for the ΛCDM model to be extended to model galaxy rotation structures. And in fact, it was single galaxy rotation curves that partly led to the inclusion of dark matter as an attribute in the ΛCDM model (Ryden 2003). A galaxy rotation curve is a plot of the orbital speeds of visible stars or gas in a galaxy against their radial distance from that galaxy's center. How much matter is visible in a given galaxy determines a simple curve for rotational speed as a function of the distance from the galactic center. However, the actual observational data did not match the calculated expected curve.

Figure 18: Example Galaxy Rotation Curve



Example galaxy rotation curve. The rotational velocities of stars are plotted against their distance from the center of a galaxy. The dotted line A plots what the expected rotation curve based on visible matter. Solid line B plots what is actually observed (Nesvold 2013).

According to general relativity[67], for the galaxies to rotate in the way the observations indicated, there must be significantly more mass in the galaxy than the mass we are able to see. As such, dark matter was postulated as an attribute of the universe and added to our model of the universe (and thus adding the 'DM' to ΛCDM).

One of the most interesting problems in astrophysics is that modellers often do not even know everything that constitutes the target system that is being modeled. In astrophysics more generally, one of the largest problems in constructing a model of the large scale structure of the universe is modellers are able to directly observationally access only a very small part of the system. They have no direct access to the entire target system of interest, and moreover they are not even totally sure what is in the target system.

[67] In this domain, the weak field approximation applies, meaning that it is actually Newtonian dynamics that is used to model galaxy rotation. However, general relativity is still considered to be the correct theory for this system. Newtonian mechanics is used because it is considered to be a good approximation in this domain.

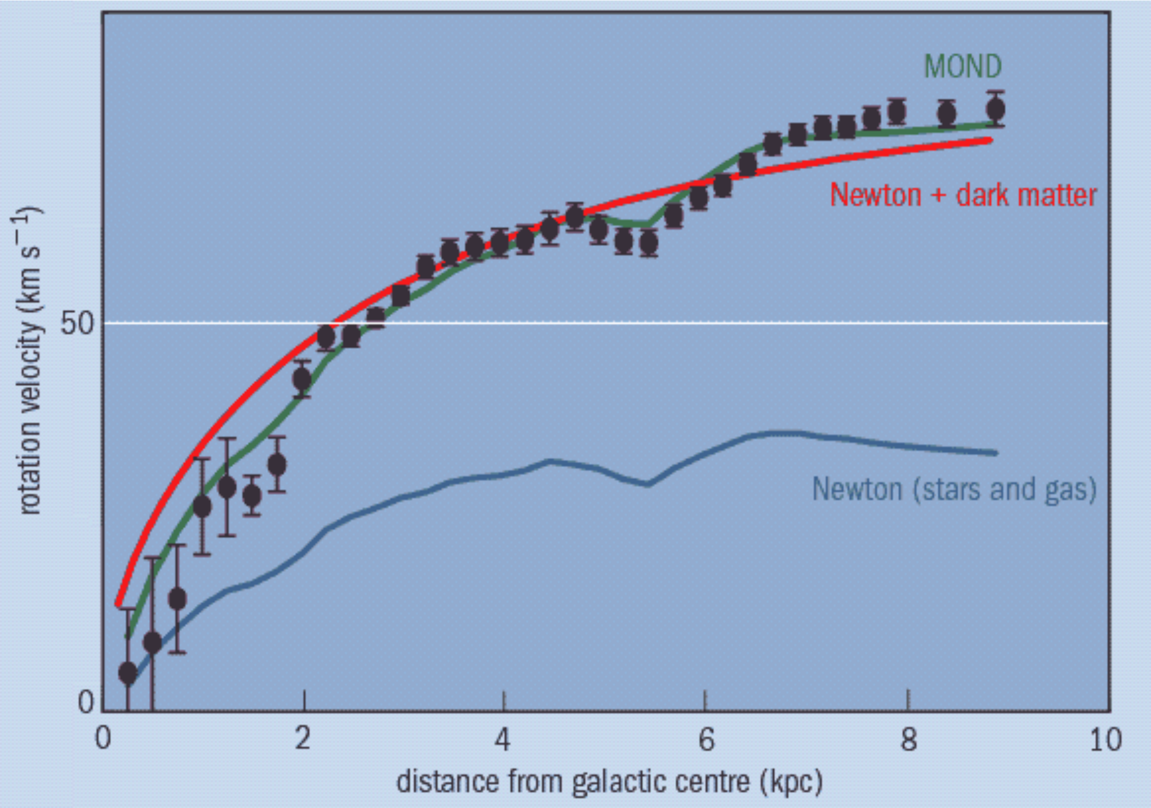
In the case of the ΛCDM model, two extremely controversial attributes are included: dark matter and dark energy. These two features have been postulated to be part of our universe, but almost entirely because astrophysicists have wanted to maintain general relativity as the basis of the model, heavilty weighting it as a mechanism. The inclusion of dark matter is a strange feature, as we currently do not know for certain that it exists in the target system. There is a debate about whether postulating dark matter in order to maintain general relativity is a bad explanatory move. As a result, other astrophysical research programs have developed alternative models that do not posit this strange dark matter. However, the problem is that in order to succeed, they have to abandon general relativity as the mechanism for gravity in the model. While some consider this to be a matter of a subjective choice between incommensurable models, my framework will show that it involves questions about the assessment of different models for different purposes. Both are successful in their domain of application for multiple purposes. However, when the models are extended to a common domain, they are actually in conflict with each other, and can be directly compared.

5.5 MOND as an Alternative Means to Fit the Data

Proponents of Modified Newtonian Dynamics (MOND) (originally proposed by Mordehai Milgrom) consider the inclusion of dark matter in order to maintain consistency with general relativity to be a bad explanatory move. The MOND research program attempts to explain galaxy rotation without dark matter by modifying the underlying laws of physics. By modifying the underlying physics proponents are able to provide a model for galaxy rotation that not only does not posit dark matter but also matches the observed data descriptively, with a higher degree of accuracy than the extended ΛCDM model (McGaugh 2014; Milgrom 1983).

Take, for example, the measured rotation curve of the galaxy NGC 1560 in Figure 19. The observed rotation velocity data points, plotted as a function of distance from the galactic center, are compared to the predictions of three models.

Figure 19: NGC 1560 Rotation Curve



The measured rotation curve of the galaxy NGC1560 shown by the data points. Newtonian curve based on the measured mass distribution (in blue), MOND (in green), and Newtonian + dark matter halo of the type predicted by CDM simulations (in red). (McGaugh (from Milgrom) 2009).

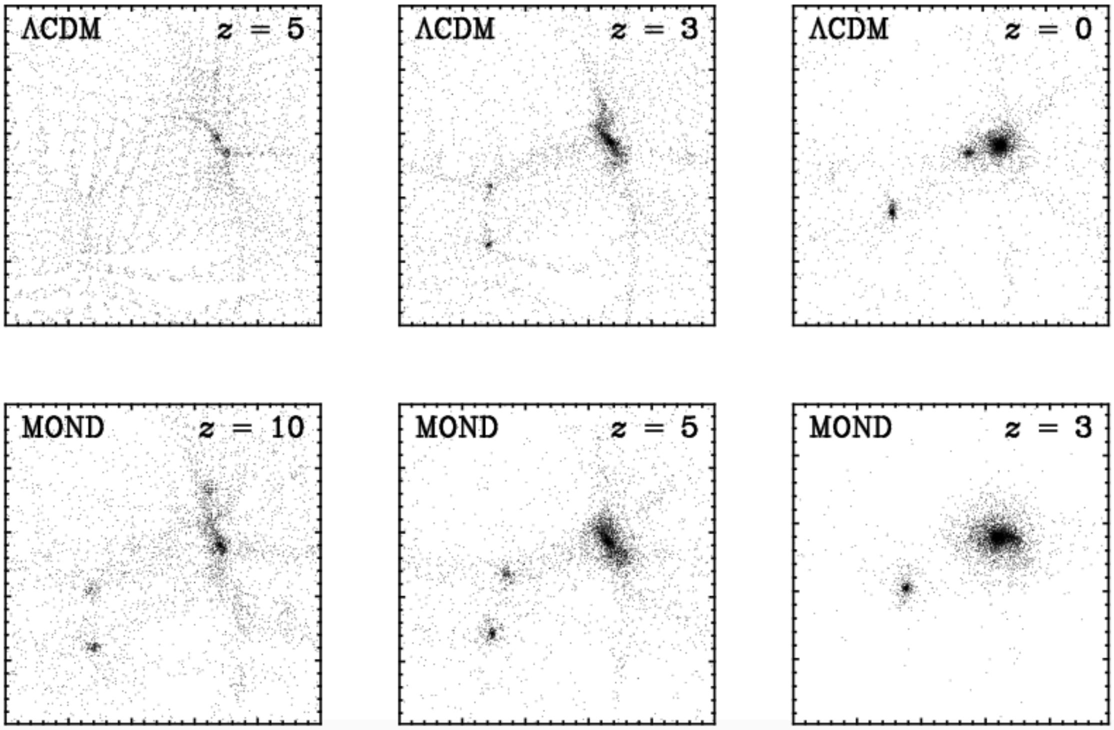
The ΛCDM-based galaxy rotation model (ΛCDM extended by further approximations for its application to a single galaxy as a model with Newtonian gravitation and the inclusion of dark matter) is able to offer an adequate descriptive fit of the rotation curves, in that it generally gets the shape correct. However, if we want a model that, with the smallest amount of deviation, describes the data points, we must say that MOND is the better model. The MOND model for galaxy rotations supports a larger descriptive similarity to the data points. The MOND model also predicts the data points of rotation curves of other galaxies with higher similarity than ΛCDM, and is thus more predictively adequate[68]. The key difference is that the proponents of MOND also want it to serve the purpose of

[68] For extended discussion and further examples of MOND verses ΛCDM galaxy rotation curves, see Randriamampandry & Carignan (2014), “Galaxy Mass Models: MOND verses Dark Matter Halos”.

offering an explanatory model, in that the explanation for these data points is modified Newtonian dynamics and not general relativity. As Milgrom states, MOND “explains almost all aspects of the mass discrepancies in galactic systems with no need to invoke dark matter. This is what MOND claims to achieve” (Milgrom 2009, 5).

Much like their colleagues who support ΛCDM, MOND proponents consider their models for galaxy rotation to be heavily based on theory, and thus believe that their similarity relations are actually capturing a real causal mechanism in the target system. So, MOND proponents have attempted to extend its domain beyond its descriptively well-fitting galaxy rotation curves. However, when moving to large scales, their models fare quite poorly. Take, for example, the extension to large-scale structure formation models.

Figure 20: ΛCDM and MOND Structure Formation



Large scale structure formation model slices for ΛCDM (top) and MOND (bottom). Notice the MOND have shifted redshift z values, for the same ΛCDM model above. (McGaugh 2014, 12)

In Figure 20 the top row is a depiction of the predicted structure formation from the ΛCDM model for the present (*z=0*), as well as several instances in the past (*z=3, z=5*).

These model outputs are considered by astrophysics to be well fitting with respect to what is known about our universe's current structure. The MOND model also achieves similar structures, but in MOND the structures develop too soon (McGaugh 2014, 13). What astrophysicists consider to be the current structure ($z=0$ in ΛCDM) occurs in MOND at $z=3$. Astrophysicists consider the MOND-based model of structure formation not to fit with our observational data-based claims of the large scale structures (McGaugh 2014; Dodelson 2011). Therefore, MOND does not have a good fit when applied to larger domains. However, ΛCDM is evaluated to fit well.

5.6 The Importance of Similarity Relation and Domain of Application

Recall that the puzzle about the ΛCDM and MOND models that motivated my discussion is how they can both be so successful, yet be based on fundamentally different physical theories. How do we make sense of this fact?

Each of these models, in its own domain, fits well with respect to description and prediction, and each includes potentially problematic explanatory structures. ΛCDM fits well with respect to describing and predicting the large-scale structure of the universe. However, in order to maintain general relativity as its fundamental explanatory mechanism, it has to include dark matter as an attribute of the target system. MOND fits well for describing and predicting the small-scale structure of single galaxy rotations. However, in order to provide this accurate description, general relativity as the underlying causal mechanism is abandoned in favor of modified Newtonian dynamics. Both of these models fit well within their domains and for their intended purposes.

What my framework allows us to see is that the models are not in conflict when applied in their domains, and it is only when they are taken to be good explanatory models, and those explanations are extended, that they come into conflict. In so far as ΛCDM is a model of structure formation, it does not apply to galaxies directly and does not directly conflict with MOND. However, its underlying fundamental physical theory is general relativity, which does conflict. It is ΛCDM's similarity with general relativity that makes it highly explanatorily adequate in its domain. However, the attempted extension to new domains, justified by high similarity to general relativity, is what brings it into direct conflict with MOND.

With respect to assessing the overall explanatory fit for these two models, it is important to remember that one needs to separately assess explanatory adequacy and similarity for the purpose of explanation. Both models are adequate for explanations in their respective domains. Recall that adequacy is assessed by attempting to use the output of the model in an explanation. Both models produce outputs that can feature as explanantia in an explanation for phenomena relevant to their domains.

But what is interesting about cases from astrophysics, such as this, is that we do not always know exactly what constitutes the target system. A consequence of having uncertainty with respect to what is in the target system is that there is some uncertainty with respect to our assessment of the similarity relation. The similarity relation establishes the representational relation between the attributes and mechanisms in the model and the attributes and mechanisms in the target system. If we do not know what is in the target system, we will not know what to capture in the similarity relation.

A claim that a model has good explanatory fit means that the model is an adequate explanation, but also that it is similar to the target system in the right kind of way. Recall, it is of special importance for explanatory similarity that the mechanisms be accurately represented in the model. If we are justified in claims that the model has a good explanatory fit, then it should be the case that the model is capturing actual mechanisms in the similarity relation that represent actual causal dependencies in the target system.

This is why we consider it justified to extend a good explanatory model beyond its original domain of application. A good explanation is grounded by the accurate representation of the target system captured in the similarity relation. For something to be a good explanation, it presupposes the idea that the similarity relation is capturing real features of the system. This is especially true with respect to mechanisms because that is how the causal relations in the world are represented in the model. Relatedly, if an element of an explanatorily adequate model is indispensable for giving a good explanation, then one should conclude that what it represents is a part of the target system.

Where the ΛCDM and MOND models run into trouble is when they move beyond their original domains and attempt to extend their claims about the model's explanatory adequacy. Both are adequate explanatory models in their original domains. But what they disagree about is what they consider to be in the target system. That is to say, they disagree about the mechanisms and attributes included in the similarity relation.

What my framework allows us to see is that when we think of a model as having high explanatory fit, because of the way the similarity relation is structured, we are endorsing the idea that the mechanisms in the model represent real causal relations in the world, and that the attributes in the model stand in that relation to one another. If scientists think they have a model with high explanatory fit, what they think they have is a model that has good similarity to the target system, and to the real counterfactual dependencies that exist in the world.

Models with high explanatory fit make the most commitments to what there is in the world, and this is the reason why it is natural to think they can be extended to new domains. If modellers have identified features about the target system within a certain domain, and find that when they extend the model to a new domain those features are still there, it is likely because they really are there in the world. As a result, one can generate new models based on having a good explanatory understanding of a phenomenon. However, this cannot necessarily be done with prediction or description. Explanatory models commit to more, namely, to the existence of real counterfactual relations.

So what does this mean, more specifically, for the case of ΛCDM and MOND? It means that there is not a conflict between the ΛCDM and MOND models when they are applied to their own domains. The ΛCDM model was constructed to describe extremely large-scale structure, and does well within this domain. However, the two models are in conflict when one attempts to extend them. We see this in ΛCDM when it is extended to small-scale structure of single galaxies. While it might be adequate for the purpose of describing the rotation curve, it does not do it as well as the MOND model for galaxy rotation does. What is hiding here in the details, however, is that the real conflict between the two is their explanation for the phenomena.

The question, then, becomes which explanation is better supported? ΛCDM has, as far as our current knowledge of our target system goes, the stronger similarity relation because the mechanism for gravitational attraction in the model comes from our current best theory. The questionable attributes that it includes (namely, dark matter and dark energy) are not *theoretically* problematic according to general relativity. But on the other hand, the questionable mechanism in MOND (namely modified Newtonian dynamics) *is* theoretically problematic, because it is inconsistent with our current best theory of gravity. One might think that the point of models, and therefore the point of MOND, is to help us explore cases in which our current best theory may not be the correct theory. Alternative models, like MOND, can help us formulate and assess new theoretical hypotheses. They are a valuable tool in testing our best theories, and discovering possible alternatives.

However, if the dynamics in the MOND model represents a viable alternative theory, then when that model is applied outside of its original domain of application, it should still do well. The MOND model should fit in the new domain, if it has identified the causal aspects that are actually in the target system. However, we do not see that. MOND, on large scales dramatically misses the mark (McGaugh 2014). This is evidence that the modified Newtonian dynamics is not identifying a real causal relation in the world, and just happens to work at the scale of the original domain of the model. While the ΛCDM model does not do better than MOND for galaxy rotation, the ΛCDM modellers do acknowledge that their model likely does not have the full similarity story, or that some of the idealizations they have made in constructing a model for the large-scale domain (such as homogeneity and isotropy) are not justifiable when examining the smaller scales.

Stacy McGaugh (2014) argues that the difference between the ΛCDM and MOND models is a matter of "mutual incommensurate paradigms", that they are opposing explanations for the observed mass discrepancies in the universe. Each paradigm has different pertinent data, and “where one makes clear predictions, the other tends to be mute. This makes comparison of the two fraught” (McGaugh 2014, 16). However, my framework provides a different means to understand the issue at hand. We should not understand these two models as being incommensurable, but rather as reflecting differences in understanding how to justify the extension of high-fit models, relative to the model’s similarity, and adequacy relative to a purpose and domain of application. My framework allows for comparison between the two, and offers a means of identifying where the models differ and why. My framework allows for a richer progress in the discussion of the conflicts between the models.

5.7 Conclusion.

My framework enables us to see where the ΛCDM and MOND models disagree, and to understand how it could be possible for two models with fundamentally different physics to both be good fits in their domains for their purposes. They conflict only when they are extended, and one is justified in extending them beyond their original domain only if the models are thought to be good explanatory models. And anyone who thinks they are good explanatory models is committed to the attributes and mechanisms in the models standing in some representational relation to attributes and mechanisms in the actual target system.

The point at which the models enter into conflict arises when we attempt to move beyond the models to claims about the real world. Ronald Giere says, “There is no best scientific model of anything; there are only models more or less good for different purposes” (Giere 2001, 1060). I have argued that this is true. However, we must acknowledge what these models commit us to when we attempt to move beyond the model itself. It is when we attempt to make inferences from our models to the real world that we must reflect on what the establishment of the similarity relation commits us to. To learn about the world from a model, the model’s construction and assessment at each stage is of primary importance.

Chapter 6

6 Conclusion

6.1 Concluding Remarks

In this dissertation I have presented a framework that is intended to be used to help disentangle the interwoven threads of evaluation of model success, model extendibility, and the ability to draw ampliative inferences about the world from models. I began by identifying three important questions that guided the development of my framework: What is the target of evaluation in model assessment? How does that evaluation proceed? What licenses us in making inferences about the real world based on the evaluation of our models as successful?

The framework identifies two distinct targets of model evaluation: representational similarity between the model and target system, and the adequacy of the model as a tool to answer questions. Both assessments must be relativized to a purpose, of which there are three general kinds: descriptive, predictive, and explanatory. These purposes differ in the way they inform the similarity relation, which is relevant for the similarity assessment and for the output they produce, which is relevant for the adequacy assessment. Any model can be assessed relative to any purpose, but a model encodes certain decisions made during the model's construction, which affects its ability to be applied to a new purpose or new domain. My framework shows that extending a model, and drawing inferences from it, depends on its representational similarity.

This framework has been successful in that it has allowed me to analyze an important contemporary debate in astrophysics between the proponents of MOND and the more commonly accepted ΛCDM model of structure formation. I have shown that the supposed conflict between the two models can be resolved by showing that it is an artifact of inappropriate extension of the models, when the explanatory similarity is not sufficient for such an extension. This conflict is not properly understood as incommensurability, as is sometimes claimed. Rather, it is a conflict between models designed for specific purposes in specific domains being unjustifiably extended. The framework has proven its value in this case and could easily be applied to analyze similar cases to identify the source of conflict in model disagreement.

However, this framework also has some open ends for further exploration and work. It is possible that there may be more than three kinds of purpose. Indeed, the three general kinds of purposes I have identified may not be exhausted or comprehensive. I do however think they are “primary” in some sense. The addition of another kind of purpose would require reflection as to what one might expect at the output stage of the framework that differs from the outputs already captured by description, prediction, and explanation. However, should there be more than just these three I see no reason the model cannot be extended to account for those.

There may be scientific models that do not neatly fit into the proposed analysis. However, my framework clearly captures large swaths of models actually used in science. Any instances of model evaluation that might not fit in this framework would be a wonderful find, as it would illuminate possible missing aspects that may be specific to a certain science or type of modeling. Additionally, one could also object that the four components of framework presuppose discrete temporal steps in the assessment of models. While this is necessary for dialectical purposes, nothing about how the analysis actually proceeds requires that they be considered as successive steps. Rather, what I have done is to draw conceptual distinctions in the ways in which models are assessed. Furthermore, this framework could allow for tracking the complexities of model evolution is an iterative process, that does not proceed cleanly from component 1 to 2 to 3 to 4. Rather, there may (and in fact often are) revisions to the construction of the model and the similarity relation in light of the output or comparable observational evidence obtained. But what the framework does track is this process, and the various aspect to the decisions made during the evaluation.

The role of evidence is also an open question both in my framework, and philosophical work on modeling more generally[69]. The role of evidence in the discussion of modelling can be quite messy, however what my framework provides is a means for identifying where evidential considerations come into play. Evidence, be it observational, experimental, or raw data can play a role in in determining what the target system will be, in the assessment of the similarity relation, or in assessing the adequacy of the model's output for the purpose. My framework allows for identifying these nodes of entry for data, as well as how that evidence is used in the justificatory process. It may be that certain evidence might support a hypothesis statement, while evidence in favor of the model more generally differ. A closer look at the role of evidence in these contexts may also be illuminating for a topic I have set aside: confirmation. There are at least three ways in which the relation between models and evidence can be discussed, the first relates to whether observations or the model itself confirms our understanding of a target system. Second, whether a variety of evidence can validate the use of a certain model to a higher degree than a single line of evidence. Finally, whether the success of a model can serve as evidence or confirmation for a particular theory. What my framework can provide is a means to identify the various justificatory processes at play in the evaluation of a model, which may be helpful for disentangling what notions of verification, validation, or confirmation are at play in a positive assessment of a model as successful. As I noted, I have attempted to provide a framework absent of discussion of scientific confirmation in order to examine the process of model evaluation. Future work should look to see how the framework developed based on examining the practice of scientific modeling fits within the existing literature on confirmation generally.

Relatedly, this project has focused on instances of modeling in which there is little uncertainty with respect to identifying the target system, in which the data is clear, and in which the models are already well-developed. However, it is often the case in actual scientific practice that the data itself is uncertain. For example, there may be sources of

[69] As a point of reference, discussion of evidence in the context of scientific models is absent from the Stanford Encyclopedia of Philosophy's entry on "Models in Science".

systematic error in certain observations made by astronomers. Most mature sciences have techniques for addressing error in their data, which allows them to create models even in cases of uncertainty. Considerations of this kind have not played a role in the analysis presented in this dissertation, and may represent an interesting complication to my framework. This is a potential area for future research. However, a few brief remarks can be made presently. Uncertainty may enter in before the model evaluation of fit proceeds, in that there may not be agreement about what features should be included in the similarity relation, and weighted-feature matching equation (component 1 of the framework). Part of the advantage to the weighted-feature matching equation is that by specifying the weightings, and why certain features are weighted in the way they are, the modeler is tracking which aspects of model they are less certain about. Error analysis may be a critical part of the assessment of adequacy, or the assessment of the similarity relation during component three or four. In comparing the model to comparable output from the target system, error analysis may help inform a modeller's evaluation. Again, this points to the strength of my framework as a means to identify the instances in which these sorts of uncertainties enter into the evaluative process.

Finally, the analysis I have proposed often relies on saying something about a modeler's intentions or decisions during the model's construction. In some cases, one does not have access to any information of this kind, or a model will be used without having been created with any purpose in mind (e.g. inherited from other unrelated or historical sciences, or parameters set by trial and error). In these cases, a certain amount of reconstruction is needed for the analysis to proceed. If the modeler has not considered why the use of a historical model or model from a different discipline may be appropriate, inferences made from those models to the world may very well *not* be justified. To be permitted in making inferences from a model, the modeler must consider the ways in which the model is similar, and represents, the target system to which it is being applied. Regardless of whether the modeler constructed the model themselves, there is some similarity relation analysis that the modeler conducts such that they think the model represents the target system in a way that is meaningful. I see my framework as being particularly illuminating in these cases, as it gives means by which a modeler can express their rational for adopting a specific existing model for their target system of interest.

Recall the joke I heard in undergraduate physics, which started me thinking about these issues: *What is the difference between a physicist and an astronomer? A physicist needs two data points to get a line of best-fit but an astronomer only needs one*. While there is a hint of truth to the joke, it fails to take into account the purpose for which most astronomical models are developed. Astronomers tend to develop models for describing an unknown target system. They operate with a small set of data, and build the best descriptions of their target system they can. Their models are heavily theoretically influenced, and therefore a significant amount can be learned from a small amount of observational evidence. It is through attending to the purpose and domain a model is intended to be a tool for that we can better understand what success in science looks like. Understanding the world is a difficult task, and models are invaluable tools. But we must understand how they represent their target systems, how they are adequate for the jobs we want to use them for, and how we are justified in drawing inferences from them, if we are to be able to truly learn anything about our world.

Bibliography

Abbott, B., R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, et al. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters 116* (6), 061102.

Andersen, H. and B. Hepburn (2016). Scientific method. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 ed.).

Barker, Gillian Abernathy (1997). *Abstraction, Analogy and Induction: Toward a General Account of Ampliative Inference*. Dissertation, University of California, San Diego

Beisbart, C. (2009). Can we justifiably assume the cosmological principle in order to break model underdetermination in cosmology? *Journal for General Philosophy of Science 40*(2), 175–205.

Bertone, G., D. Hooper, and J. Silk (2005). Particle dark matter: Evidence, candidates and constraints. *Physics Reports 405*(5), 279–390.

Bokulich, A. (2011). How scientific models can explain. *Synthese 180*(1), 33–45.

Bokulich, A. (2013). Explanatory models versus predictive models: reduced complexity modeling in geomorphology. In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, pp. 115–128. Springer.

Boylan-Kolchin, M., J. S. Bullock, and M. Kaplinghat (2011). Too big to fail? the puzzling darkness of massive Milky Way subhaloes. *Monthly Notices of the Royal Astronomical Society: Letters 415*(1), L40–L44.

Boylan-Kolchin, M., J. S. Bullock, and M. Kaplinghat (2012). The Milky Way's bright satellites as an apparent failure of Λ CDM. *Monthly Notices of the Royal Astronomical Society 422*(2), 1203–1218.

Boylan-Kolchin, M., V. Springel, S. D. White, A. Jenkins, and G. Lemson (2009). Resolving cosmic structure formation with the Millennium-II simulation. *Monthly Notices of the Royal Astronomical Society 398*(3), 1150–1164.

Bruch, A., L. Vaz, and M. Diaz (2001). An analysis of the light curve of the post common envelope binary MT Serpentis. *Astronomy & Astrophysics 377*(3), 898–910.

Calder, A., B. Fryxell, T. Plewa, R. Rosner, L. Dursi, V. Weirs, T. Dupont, H. Robey, J. Kane, B. Remington, et al. (2002). On validating an astrophysical simulation code. *The Astrophysical Journal Supplement Series 143*(1), 201.

Carnap, R. (1937). *Logical Syntax of Language*, Volume 4. Psychology Press.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Chatzopoulos, E., J. C. Wheeler, and J. Vinko (2012). Generalized semi-analytical models of supernova light curves. *The Astrophysical Journal 746*(2), 121.

Contessa, G. (2010). Scientific models and fictional objects. *Synthese 172*(2), 215–229.

Conway, J. (1970). The game of life. *Scientific American 223*(4), 4.

Cox, T. and A. Loeb (2008). The collision between the Milky Way and Andromeda. *Monthly Notices of the Royal Astronomical Society 386*(1), 461–474.

Craver, C. F. (2006). When mechanistic models explain. *Synthese 153*(3), 355–376.

Dodelson, S. (2011). The real problem with MOND. *International Journal of Modern Physics D 20*(14), 2749–2753.

Downes, Stephen M. (2009). Models, Pictures, and Unified Accounts of Representation: Lessons from Aesthetics for Philosophy of Science. *Perspectives on Science* 17 (4):417-428.

Downes, S. M. (2011). Scientific models. *Philosophy Compass 6*(11), 757–764.

Earman, J. and J. Mosterín (1999). A critical look at inflationary cosmology. *Philosophy of Science 66*(1), 1–49.

Elgin, M. and E. Sober (2002). Cartwright on explanation and idealization. *Erkenntnis 57*(3), 441–450.

Elliott-Graves, Alkistis, (2014). *The role of target systems in scientific practice.* Dissertations, University of Pennsylvania.

Elliott-Graves, Alkistis & Weisberg, Michael (2014). Idealization. *Philosophy Compass* 9 (3):176-185.

Fabbri, A. and J. Navarro-Salas (2005). *Modeling Black Hole Evaporation.* Imperial College Press.

Famaey, B. and S. McGaugh (2013). Challenges for ΛCDM and MOND. In *Journal of Physics: Conference Series,* Volume 437, pp. 012001. IOP Publishing.

Famaey, B. and S. S. McGaugh (2012). Modified Newtonian dynamics (MOND): observational phenomenology and relativistic extensions. *Living Reviews in Relativity 15*(10), 1–159.

Faraway, J., A. Mahabal, J. Sun, X.-F. Wang, Y. G. Wang, and L. Zhang (2016). Modeling lightcurves for improved classification of astronomical objects. *Statistical Analysis and Data Mining: The ASA Data Science Journal 9*(1), 1–11.

Faraway, J. J. (2014). *Linear Models with R*. CRC Press.

French, Steven (2003). A model-theoretic account of representation (or, I don't know much about art…but I know it involves isomorphism). *Philosophy of Science* 70 (5):1472-1483.

French, Steven & Ladyman, James (1999). Reinflating the semantic approach. *International Studies in the Philosophy of Science* 13 (2):103 – 121.

Frigg, Roman (2006). Scientific representation and the semantic view of theories. *Theoria* 21 (1):49-65.

Frigg, R. (2010). Models and fiction. *Synthese 172*(2), 251–268.

Frigg, R. and S. Hartmann (2012). Models in science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 ed.).

Giannantonio, T., A. Lewis, and R. Crittenden (2010). Some things we know about the universe are probably right. *Astronomy & Geophysics 51*(5), 5–16.

Giere, R. (1988). *Explaining Science: A Cognitive Approach*. University of Chicago Press.

Giere, R. N. (1996). Visual Models and Scientific Judgment. *Picturing knowledge: Historical and philosophical problems concerning the use of art in science*, 269.

Giere, Ronald N. (1999). *Science Without Laws*. University of Chicago Press.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science 71*(5), 742–752.

Giere, R. (2006). *Scientific Perspectivism*. University of Chicago Press.

Giere, R. (2009). Why scientific models should not be regarded as works of fiction. In M. Suárez (Ed.), *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 248–258. Routledge.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese* 172(2), 269–281.

Giere, R., J. Bickle, and R. Mauldin (2006). *Understanding Scientific Reasoning*. Thomson/Wadsworth.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy 21*, 725—740.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies 143*, 101–116.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*. Indianapolis: Bobbs-Merrill.

Hamilton, J.-C. (2014). What have we learned from observational cosmology? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 46*, 70–85.

Hansen, F. K., A. Banday, and K. Górski (2004). Testing the cosmological principle of isotropy: local power-spectrum estimates of the WMAP data. *Monthly Notices of the Royal Astronomical Society 354*(3), 641–665.

Hawking, S. W. (1974). Black hole explosions. *Nature* 248 (5443), 30–31.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Hitchcock, C. and E. Sober (2004). Prediction versus accommodation and the risk of over-fitting. *British Journal for the Philosophy of Science 55*(1), 1–34.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science 64*, S325–S336.

Humphreys, P. (2004). *Extending Ourselves Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Jebeile, J. and A. Barberousse (2016). Empirical agreement in model validation. *Studies in History and Philosophy of Science Part A 56*, 168–174.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusion*. New York: Oxford University Press.

Komatsu, E., K. Smith, J. Dunkley, C. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. Nolta, L. Page, et al. (2011). Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Cosmological interpretation. *The Astrophysical Journal Supplement Series 192*(2), 18.

Krasinski, A. (2006). *Inhomogeneous Cosmological Models*. Cambridge, UK: Cambridge University Press.

Lahav, O. (2001). Observational tests for the cosmological principle and world models. In *Structure Formation in the Universe*, pp. 131–142. Springer.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist 54*(4), 421–431.

Lloyd, E. A. (2009). Varieties of support and confirmation of climate models. In *Aristotelian Society Supplementary Volume*, Volume 83, pp. 213–232. The Oxford University Press.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science 77*(5), 971–984.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A 49*, 58–68.

Lynds, R. and A. Toomre (1976). On the interpretation of ring galaxies: the binary ring system II HZ 4. *The Astrophysical Journal 209*, 382–388.

Maartens, R. (2011). Is the universe homogeneous? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 369*(1957), 5115– 5137.

McGaugh, S. S. (2014). A tale of two paradigms: the mutual incommensurability of ΛCDM and MOND 1. *Canadian Journal of Physics 93*(2), 250–259.

McMullin, E. (1978). Structural explanation. *American Philosophical Quarterly 15*(2), 139–147.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A 16*(3), 247–273.

Melia, F. (2015). The cosmic equation of state. *Astrophysics and Space Science 356*(2), 393–398.

Milgrom, M. (1983). A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *The Astrophysical Journal 270*, 365–370.

Milgrom, M. (2001). MOND–a pedagogical review. *arXiv preprint* astro-ph/0112069.

Milgrom, M. (2002). MOND's theoretical aspects. *New Astronomy Reviews 46*(12), 741–753.

Milgrom, M. (2009). MOND: Time for a change of mind? *arXiv preprint* arXiv:0908.3842 .

Misner, C. W., K. S. Thorne, and J. A. Wheeler (1973). *Gravitation*. Macmillan.

Mo, H., F. Van den Bosch, and S. White (2010). *Galaxy Formation and Evolution*. Cambridge University Press.

Morgan, M. S. and M. Morrison (1999). *Models as Mediators: Perspectives on Natural and Social Science*, Volume 52. Cambridge University Press.

Muldoon, R. and M. Weisberg (2011). Robustness and idealization in models of cognitive labor. *Synthese 183*(2), 161–174.

Nesvold, E. (2013). Do elliptical galaxies have dark matter halos? Website. url = {https://astrobites.org/2013/04/04/do-elliptical-galaxies-have-dark-matter-halos/}, accessed June 2016.

Norton, J. (2012). Approximation and idealization: Why the difference matters. *Philosophy of Science 79*, 207–232.

Norton, J. (2013). *Einstein for everyone.* Nullarbor Press.

Oppenheimer, J. R. and H. Snyder (1939). On continued gravitational contraction. *Physical Review 56*(5), 455.

Oreskes, N., K. Shrader-Frechette, K. Belitz, et al. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science 263*(5147), 641–646.

Page, D. N. and S. Hawking (1976). Gamma rays from primordial black holes. *The Astrophysical Journal* 206, 1–7.

Parker, W. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary Volume 83*(1), 233–249.

Parker, W. (2010). Scientific models and adequacy-for-purpose. *The Modern Schoolman 87*(3/4), 285–293.

Parker, W. (2015). Getting serious about similarity. *Biology and Philosophy 30*(2), 267–276.

Peebles, P. J. E. (1993). *Principles of Physical Cosmology.* Princeton University Press.

Peebles, P. J. E. (2014). Discovery of the hot big bang: What happened in 1948. *The European Physical Journal H 39*(2), 205–223.

Primack, J. R. (2015). Cosmological structure formation. *arXiv preprint* arXiv:1505.02821.

Quine, W. V. O. (1969). *Ontological relativity and other essays*. Columbia University Press.

Randriamampandry, T. H. and C. Carignan (2014). Galaxy mass models: MOND versus dark matter halos. *Monthly Notices of the Royal Astronomical Society 439*(2), 2132–2145.

Riess, A. and et. al. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal 116*, 1009–1038.

Romano, R., Y. Mayya, and E. Vorobyov (2008). Stellar disks of collisional ring galaxies. i. new multiband images, radial intensity and color profiles, and confrontation with n-body simulations. *The Astronomical Journal 136*(3), 1259.

Ryden, B. (2003). *Introduction to Cosmology*. San Francisco: Pearson Addison Wesley.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation*. Minnesota Studies in the Philosophy of Science 13, 3–219.

Schupbach, J. N. (forthcoming 2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science.*

Spergel, D. N. (2015). The dark side of cosmology: Dark matter and dark energy. *Science 347*(6226), 1100–1102.

Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science 17*(3), 225–244.

Suppe, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. Chicago: University of Illinois Press.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. CSLI publications Stanford.

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis 55*(3), 393–415.

Thorne, K. (1994). *Black Holes and Time Warps: Einstein's Outrageous Legacy.* New York: W.W. Norton.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vandergurgh, W. (2003). The dark matter double bind: Astrophysical aspects of the evidential warrant for general relativity. *Philosophy of Science 70*, 812–832.

Wald, R. (1984). *General Relativity.* Chicago: University of Chicago Press.

Weinberg, S. (1972). Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity. New York: Wiley.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science 73*(5), 730–742.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy 104*(12), 639–659.

Weisberg, M. (2012). Getting serious about similarity. *Philosophy of Science 79*(5), 785–794.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World.* New York: Oxford University Press.

Weisberg, M. (2015). Biology and philosophy symposium on simulation and similarity: Using models to understand the world. *Biology & Philosophy 30*(2), 299–310.

Weisberg, M. (forthcoming). Validating idealized models. In B. van Fraassen and I. Peschard (Eds.), *The Experimental Side of Modeling*. The University of Chicago Press.

Weisberg, M. and K. Reisman (2008). The robust Volterra principle. *Philosophy of Science 75*(1), 106–131.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. *Scientific Inquiry and the Social Sciences*, 124–163.

Wimsatt, W. C. (1987). False models as means to truer theories. *Neutral Models in Biology 23*.

Wimsatt, W. C. (2002). Using false models to elaborate constraints on processes: Blending inheritance in organic and cultural evolution. *Philosophy of Science 69*(S3), S12–S24.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Curriculum Vitae

Name:	Melissa Jacquart
Post-secondary Education and Degrees:	University of Wisconsin-Madison Madison, Wisconsin, United States of America 2005-2009 B.S. Astronomy-Physics, Physics, Philosophy The University of Western Ontario London, Ontario, Canada 2011-2012 M.A. Philosophy The University of Western Ontario London, Ontario, Canada 2012-2016 Ph.D. Philosophy
Honours and Awards:	Rotman Institute Doctoral Entrance Scholarship, 2012 Academic Achievement Scholarship, Local 610 GTA Union, 2012 Graduate Research Assistantship, 2012, 2013 Rotman Institute Graduate Research Assistantship Grant, 2014, 2015 Mary Routledge Fellowship, 2015 Graduate Research Scholarship, 2015-2016 Graduate Thesis Research Award, 2016
Related Work Experience:	Teaching Assistant, Philosophy Department, Western University 2011-2012, 2012-2013, 2014-2015 Lead TA, Philosophy Department, Western University 2013-2014 TA Training Program Instructor, UWO Teaching Support Centre 2014-2016

Publications:

Kronz, F. and Jacquart, M. (2011). “The Scientific Method” in *Leadership in Science and Technology: A Reference Handbook*. Edited by William Bainbridge. SAGE Publications, p.183–190. ISBN: 141297688X.