1 Introduction

Scientists construct quantitative models to explain observations about natural systems in a coherent and rigorous manner. These models can be characterized by their scope and validity. The scope of a model is the set of observable quantities that the model can generate predictions about, and the validity of a model is the extent to which these predictions agree with empirical observations of those quantities. Today, as the number of models and the quantity of empirical data increases, scientists face a grand challenge: efficiently discovering relevant models and characterizing their validity against a continually growing body of available evidence. A scientist typically proposes a new model by publishing a description of how it works along with an argument justifying its utility to some targeted scientific community. This argument is made in words, accompanied by a few relevant equations and figures, and judged by the process of scientific peer review. Ideally, reviewers determine whether the models predictions are consistent with all available and relevant data and compare their accuracy against those produced by previously published models. This is increasingly difficult; although authors are expected to facilitate this process by referring to relevant experimental data and conducting a literature review, these citations are likely to be both incomplete and biased, in that authors will dedicate the most space in their publications to data that their model explains well relative to a subset of previously developed models, or to an implausibly simple null hypothesis.

If each modeling paper contained figures a) highlighting the models accordance with all related experiments and b) comparing its performance to every related model, then publications would be encyclopedic and their main points would become obscured. The strength of model publication as it stands is the focused description of how a model works and its conceptual and technical advances. A weakness, however, is that evaluating the scope and validity of a published model is intractable using the contents of a publication alone. In other words, publications tell us clearly how a model works but provide only incomplete outlines for telling us which goals it hopes to achieve and how well it achieves them. This problem is only exacerbated as more data is gathered in the years following the original publication. Although the validity of a model may change in light of new experimental observations, there is no systematic process in biology today for re-evaluating existing models. Although the new data and its most important theoretical implications may propagate informally through a scientific community or appear in periodic reviews, the original publications, which serve as a resource of first resort for new scientists and others outside of the community, will be cited as-is in perpetuity.

We can now distill the central problem discussed in this proposal: the process by which models are validated is not sufficiently rigorous, comprehensive or ongoing. This makes it difficult for scientists to identify which models best predict quantities of interest, to compare proposed models against one another, and to precisely identify problems that have not yet been solved in their research area. To overcome these problems, we propose formalizing the model validation process by creating a collection of software and associated cyberinfrastructure dedicated to making scientific model validation more systematic. This validation framework will exist in parallel to the publication system, allowing the latter to focus on answering how, while referring to the former for a comprehensive answer to how well.

1.1 Existing Efforts

While there have been several facilities developed for data and model sharing in biology, there have been few attempts to facilitate evaluation of models against data. For example, the Collaborative Research In Computational Neuroscience (CRCNS) data-sharing website (www.crcns.org), and the Open Source Brain repository (www.opensourcebrain.org) are separate facilities for data and model sharing in neuroscience, respectively. The CRCNS website is specifically focused on data sharing for the benefit of computational modelers, and has benefited from community contribution of several excellent data sets. The latter resource, the Open Source Brain Initiative, is focused on model description and execution, and is an emerging example of the power of standards in informatics. However, it lacks a means to test models, because it lacks the means to compare model outputs against data. The work proposed here aims to bridge these two kinds of resources, strengthening each in turn.

There are related efforts in the machine learning community to develop models and validate them against publicly-available datasets. The thrust of Kaggle (www.kaggle.com), for example, is to drive model development by organizing competitions where developers receive a set of training data and submit competing algorithms, compared automatically by their cross-validated accuracy on an unrevealed test set. The success of Kaggle [[ref]] shows that open competitions can be highly effective, and that this paradigm of modeling as a competition draws in large numbers of data scientists across traditional discipline boundaries. However, the models developed in this way must fit into a few very general use cases in machine learning: classification, regression, and clustering. The validation criteria are relatively straightforward in these domains. Discipline-specific competitions in biology have also resulted in technical advances for specific problems. For example, the quantitative single neuron modeling competition [ref] has helped us understand the complexity-accuracy tradeoff among reduced models of excitable membranes, and identified models with the best predictive power for spike trains. Another competition, the Hopfield challenge [ref], famously illustrated through its difficulty the challenges facing computational neuroscience. [[Other examples from biology still needed.]] Our challenge is to develop a general framework in support of such distributed data-driven model validation workflows, where the validation criteria may be complex and discipline-specific. We will be particularly focusing on validation challenges in the biological sciences.

2 Outcomes and Products

Our framework is centered around simple validation tests that compute the agreement between a result from a model execution and a single experimental observation. The overall validity of a model can be identified with the collection of tests that the model passes. This methodology is inspired directly by the nearly ubiquitous practice of unit testing in software engineering. A unit test evaluates whether a portion of a computer program (often a single function) meets one simple correctness criterion. A suite of such tests that cover the desired behaviors of a program validates the overall functionality of the program, and helps isolate specific causes of failures. Developers often write unit tests before writing the program itself, following a methodology known as test-driven development (TDD) [**?**]. TDD is based on the idea that a suite of unit tests can directly serve as a programs specification, guiding its development. Following the TDD methodology allows developers to measure progress simply by looking at the proportion of tests that pass at any point

in development. When additions or modifications are made, developers can be sure that changes did not break existing functionality by ensuring that all tests that previously passed continue to do so. The success of unit testing and TDD in practice suggests that validation testing may be a practical solution to the problems discussed above.

3 Preliminary Activities

4 Research and Development Plan

5 Community and Educational Outreach

6 Personnel and Coordination

7 Results from Prior NSF Work