Progress in Biophysics and Molecular Biology 129 (2017) 3–12

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Prospects and problems for standardizing model validation in systems biology

Fridolin Gross [a, *], Miles MacLeod [b]

[a] Institute for Philosophy, University of Kassel, Nora-Platiel-Strasse 1, 34127 Kassel, Germany
[b] Department of Philosophy, University of Twente, Drienerlolaan 5, 7522DN Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form
20 August 2016
Accepted 11 January 2017
Available online 12 January 2017

Keywords:
Systems biology
Modeling
Standardization
Validation
Model selection

ABSTRACT

There are currently no widely shared criteria by which to assess the validity of computational models in systems biology. Here we discuss the feasibility and desirability of implementing validation standards for modeling. Having such a standard would facilitate journal review, interdisciplinary collaboration, model exchange, and be especially relevant for applications close to medical practice. However, even though the production of predictively valid models is considered a central goal, in practice modeling in systems biology employs a variety of model structures and model-building practices. These serve a variety of purposes, many of which are heuristic and do not seem to require strict validation criteria and may even be restricted by them. Moreover, given the current situation in systems biology, implementing a validation standard would face serious technical obstacles mostly due to the quality of available empirical data. We advocate a cautious approach to standardization. However even though rigorous standardization seems premature at this point, raising the issue helps us develop better insights into the practices of systems biology and the technical problems modelers face validating models. Further it allows us to identify certain technical validation issues which hold regardless of modeling context and purpose. Informal guidelines could in fact play a role in the field by helping modelers handle these.

Contents

1. Introduction 3
2. Validation of models in science and engineering 4
3. Are validation standards desirable generally in systems biology? 6
4. Constraints on standardization 8
   4.1. The diversity of practice and purposes 8
   4.2. The technical feasibility of standardization 10
5. Conclusion 11
   References 12

1. Introduction

The number and diversity of computational models which are used to study biological processes at molecular, intercellular and physiological levels is steadily growing. The field of systems biology unites scholars from very different backgrounds, and consequently styles of building models vary greatly. As Jeremy Gunawardena has highlighted, systems biology will need to start “harmonizing [the] cacophony” of “concepts and techniques that are coming into the subject from the physical sciences and computer science” (Gunawardena, 2010; 42). And indeed, there have been many efforts over the past years to implement certain standards in systems biology, in the form of modeling languages, such as the systems

* Corresponding author.
E-mail addresses: fridolin.gross@uni-kassel.de (F. Gross), m.a.j.macleod@utwente.nl (M. MacLeod).

http://dx.doi.org/10.1016/j.pbiomolbio.2017.01.003
0079-6107/

biology markup language (SBML), and the collection of models in standardized formats in publicly available databases (e.g. the biomodels database). Most of these efforts, however, concern purely syntactical aspects of modeling and are not concerned with model validation. The need for standards for validation has been explicitly expressed by the systems biology community as well (Klipp et al., 2007).

Model validation refers to the process of establishing whether a "model reliably reproduces the crucial behavior and quantities of interest within the intended context of use." (Rykiel, 1996; 226) Standardization of any model-building and assessment process has much to recommend it, since it could improve clarity and communication within a field, thus promoting both productivity and efficiency. Standardized testing schemes for systems biology models could greatly facilitate the accumulation of knowledge in the field through the development of model databases which inform their users in common transparent terms on the extent of the reliability of a model without having to take this reliability on trust or investigate it themselves. Agreed upon standards could further help establishing a basis for safe reliable use of models given their increasing role in the design and testing of medical technologies and treatments, and at least in some such contexts systems biologists see a patent need for standardization (see Viceconti et al., 2016).

However, it is possible that setting standards for model validation is neither an achievable nor helpful goal for systems biology to aspire to generally, at least at this point in its development. The nature and complexity of living systems might make it intrinsically difficult to achieve the same kind of standardization achieved for instance in engineering disciplines. In this contribution we would like to discuss the prospects and problems for standardizing model validation in systems biology. By listing various challenges our goal is not to dismiss standardization out of hand but merely to point to various obstacles that any drive to standardize validation might have to contend with. We begin by exploring what is meant by validation and what previous discussions on the concept of validation contribute to discussions over standardization (Section 2). We then motivate the desirability for validation standards in systems biology (Section 3). In the fourth section of this paper we discuss both the practical and technical problems that standardization faces, given the current situation in systems biology. We close with some cautious recommendations regarding the prospects of standardization. Our reasoning for the most part is philosophical in nature, that is, we are mainly concerned with the current methodological practices in the field and the rational concepts underpinning validation. We develop the technical details only when necessary. A further part of our analysis is based on the results of an ethnographic study led by Nancy Nersessian of model-building practices in two systems biology labs.

2. Validation of models in science and engineering

Philosophy of science has investigated various aspects of scientific modeling. For example, a major discussion in philosophy of science concerns the kind of structures models are and the degree to which the primary purpose of scientific models is representational or inferential (see Suárez, 2004). Philosophy has also addressed extensively the relations between model and theories, and the role models in play in scientific discovery processes (see Frigg and Hartmann, 2012). Much less has been written about the proper or effective bases or procedures by which models can be justified or verified for their given purposes. The most important exception to this are philosophical debates about the validity of robustness analysis as a source of empirical evidence on the accuracy or reliability of a model's results (Weisberg, 2006). Otherwise, however, the literature on model validation has often focused on more abstract or fundamental questions. For example, validation has been discussed as a special case of fundamental philosophical problems such as induction or theory confirmation (for an overview see Kleindorfer et al., 1998). In a particularly influential article in this spirit Oreskes et al. (1994) argue that validation is a fundamentally misleading concept because it is impossible to establish the truth of a model. Calling a model "validated" is risky and falsely misrepresents models as "true" or "false" to policy-makers. According to this stance, models can only be falsified, and their primary role in science is heuristic, i.e., for the purpose of developing hypotheses. Of course there is a legitimate worry here. Models can be given too much credibility and authority through uncritical labeling of models as "valid". However, this analysis of validation has been criticized as of little use for practical decision making based on computational models in engineering and technology, in which engineers seek a principled epistemic basis upon which to make decisions about how to use and rely on the models they build (e.g. Oberkampf and Roy, 2010). In scientific contexts it seems unproductive to narrowly frame the issue of validation in terms of truth or falsity alone, and philosophers of science might benefit from more pragmatic approaches. On the issue of how models should be represented in the environmental sciences Peterson (2006) for instance constructs a practical system by which modelers can represent uncertainty and the sources of it to policy-makers. Küppers and Lenhard (2005) demonstrate the importance of constructing independent standards for validating social science models as opposed to natural science models, given the nature of the phenomena and practices social science deals with.

Our principal interest, like that of Küppers and Lenhard above, is not in higher level philosophical debates over the status of models but in the practical conditions by which models can be justified for a particular set of goals and how well practices in systems biology and the nature of biological systems afford the possibility of standardized validation procedures. We follow Carusi (2014) and Carusi et al. (2012) by treating it as important to understand the constraints on practices in order to comprehend how well these might align with robust and recognized validation procedures. Before going further then, it is wise to develop some broad understanding of the meaning of validation, and what is commonly thought relevant to it in scientific circles.

Rykiel (1996) discusses validation of simulation models with an eye to the requirements of scientific practice. A validation judgment, according to him, is an assessment of the accuracy of a model and of whether that accuracy justifies reliance on the model for at least certain goals or ends. Rykiel writes with ecology in mind but many of his findings apply generally. For instance validation assessments or procedures may take many forms in practice. Models of a system may be validated operationally, according to how well models fit the available data on that system's behavior. Such validations can be complex, involving procedures like sensitivity analysis or other form of statistical analysis, which try to discover how robustly and precisely a model mimics a system. As Rykiel puts it, these kinds of validation focus on performance (rather than representational accuracy directly) as their principal goal, and models can be tweaked or engineered to produce better performance without necessarily improving the degree to which a model soundly captures reality.

Secondly, a model can be validated to the degree to which it does capture the known properties and structure of the phenomena it models. Rykiel calls this conceptual validity. Validation in this sense uses available theories and knowledge of phenomena to assess whether a model captures accurately what is known of the phenomena and to assess how well the abstractions and idealizations the model relies on might assist or compromise its ability to

do this. These two meanings have also been captured with the distinction between 'functional validity' and 'structural validity' (Peterson, 2006).

Finally, a model may be validated according to how accurate the underlying data is from which it is built or tested. Within this framework a multiplicity of validation procedures have been developed which attempt to analyze the accuracy or robustness of models and quantify their uncertainty or risk. Sensitivity analysis, for instance, allows modelers to evaluate model uncertainty by quantifying the influence that variations in model inputs (e.g. parameters, boundary conditions) have on variation of model output (i.e. simulated results of interest).

Formal aspects of model validation have been most thoroughly developed in the realm of engineering, with the perhaps most influential definition of validation coming from the Defense Modeling and Simulation Office (DMSO), an organization within the United States Department of Defense:

> Validation: the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model (DoD, 1994).

This definition has been adopted both by the Institute of Electrical and Electronics Engineers (IEEE) and the American Society of Mechanical Engineers (ASME). Validation is distinguished from verification which refers to the evaluation of the numerical implementation of a model:

> Verification: the process of determining that a model implementation accurately represents the developer's conceptual description of the model (DoD, 1994).

Verification is concerned with the degree to which a numerical procedure or simulation software approximates the solutions of the underlying theoretical model. It can be treated as a purely mathematical problem. Validation, by contrast, directly compares the model with the target system existing in the real world. These definitions are deliberately kept general as they are intended to cover a vast diversity of applications, ranging from nuclear waste management to warfare simulation. Obviously, both the use of a model and the required degree of accuracy will strongly depend on the particular problem and on the specifics of the target system. There are, however, attempts to provide an overarching framework for verification and validation (V&V) that aims at providing guidance across different disciplines (Oberkampf and Roy, 2010).

In line with the definition given above, the V&V framework understands model validation as an assessment of accuracy which is achieved by comparing the simulated responses of a computational model to experimental data. Importantly, the accuracy of a computational model must be quantified since only in this way can the model performance be compared to a specified degree of required accuracy. Morrison (2015) argues that within this framework one has to distinguish between two distinct issues:

(1) a comparison between experimental data and computational results that focuses on accuracy, and
(2) the adequacy of the comparisons for a specific purpose. (Morrison, 2015; 270)

The first issue is properly the domain of validation and, according to Morrison, it should be objective in the sense that "standards of agreement should be specified in a formal way" (Morrison, 2015; 270). Informal validation procedures typically consist in visually assessing the agreement of graphs representing experimental measurements and simulated results. However, such procedures usually cannot properly take into account errors in the numerical solutions or uncertainties in the experimental data. The framework of V&V thus aims at providing a formal way of comparing model and reality. Two important concepts in this regard are *validation metrics* and *validation experiments*. A validation metric is a quantitative measure of agreement between measured data and model results. It has the mathematical properties of a distance measure and can be deterministic or probabilistic, depending on the type of model under consideration. Validation experiments are specifically designed experiments to assess the accuracy of simulation results by means of such a validation metric. Their aim is to yield the specific kind of information that is amenable to comparison with simulated data. It is crucial to exhaustively specify and record the characteristics of a validation experiment such that the grounds for validation are established.

Oberkampf and Roy emphasize that while close interaction of analysts and experimentalists is crucial, maintaining a certain independence is important for validation as well:

> To achieve the most value from the validation experiment, there should be in-depth, forthright, and frequent communication between analysts and experimentalists during the planning, design, and execution of the experiment. Also, after the experiment has been completed, the experimentalists should provide to the analysts all the important input quantities needed to conduct the simulation. What should *not* be provided to the analysts in a rigorous validation activity is the measured SRQ [system response quantities]. Stated differently, a blind-test prediction should be compared with experimental results so that a true measure of predictive capability can be assessed in the validation metric (Oberkampf and Roy, 2010; 477–478).

The task of model validation is often complicated by *aleatory* and *epistemic uncertainty* both in the predictions of a model and in experimental measurements. Aleatory uncertainty refers to intrinsic noise or variability in the target system, while epistemic uncertainty is due to lack of knowledge about the underlying structure. The quantification and propagation of uncertainty can to some extent be captured within a probabilistic framework, but especially for epistemic uncertainty this is often problematic.

Another important concept in the context of V&V is *calibration*. Calibration refers to the process of adjusting model parameters in order to improve agreement with experimental data, and it directly impacts on the predictive reliability of a model. Calibration of models is particularly common in situations of poor knowledge about a system or of lack of precise measurements. We will later see that the problem of calibration is especially prevalent in systems biology modeling.

Even though validation and verification are often discussed in parallel, in what follows we exclusively focus on questions of model validation. We do not think that systems biology raises problems for verification that are not already present in other fields, and to give a general discussion would go beyond the scope of this article.[1]

In contemporary systems biology, which models a range of biochemical, cellular and physiological systems, validation strategies remain relatively diverse and informal compared to engineering fields in general (see Section 3). The most ubiquitous mode of validation taught within the field as the paradigm for validation, is operational (see for instance Voit, 2012). Models are validated for predictive purposes to the extent to which they fit the available

[1] For a critical discussion of the relationship of verification and validation see Winsberg (2010).

data. Their ability to capture system behavior independently of uncertainties in input parameters requires various forms of sensitivity analysis which modelers import from their engineering backgrounds (Zi, 2011). Modelers however generally assert in practice that raw fit will not necessarily produce a reliable model (because of the problem of overfitting, see Section 4) and instead some of the data used to build a model must be put aside for testing the model after it is built for predictive testing. This testing can be both intuitive, by studying visual matches, between the phenomena and the model, or statistical, and can range in the degrees of precision it demands from qualitative correctness to strict numerical accuracy (Voit, 2012; 39). Models that pass these tests are generally thought the most well-validated for predicting at least some aspects of a system. In practice the range of system behaviors over which that reliability holds is not necessarily easy to identify, nor are the standards of accuracy, amount of data etc. necessary to ensure a given range of trustworthiness (see Section 4.2). Furthermore conceptual validation also likely plays a role with respect to what modelers rely on to argue that their models are valid, particularly if the underlying structure of a biological network is thought well-established. In general, many validation procedures in practice seem ad-hoc or informal, which results in uncertainty about the extent of validity of "validated" models in the field and inconsistency of the ways in which models are being accepted and applied. These problems have led to a dissatisfaction among at least some practitioners with the state of validation practices within the field (Viceconti et al., 2016; 71).

3. Are validation standards desirable generally in systems biology?

As mentioned standardization procedures for any facet of model-building or indeed any aspect of scientific practice, stand to improve many aspects of those practices, and of course provide a system others outside the field can trust without having to understand its technical details. Certification standards are thus widely used in engineering and other manufacturing contexts. Since systems biology itself has engineering aims – the production of technological and abstract structures, like models, which have practical societal value –, it is at least plausible to ask whether certification standards might be applicable for models in systems biology, too. The American Society of Mechanical Engineers has invested considerable energy itself in developing standard verification and validation schemas for a range of computational models in solid and fluid mechanics to at least determine consistently the degree of accuracy of specific variables at specific points when comparing simulations to experimental data (see Schwer et al., 2006).

Having a consensus on validation would be productive for numerous reasons in systems biology and the current lack of it is problematic. One benefit of standardized validation criteria, which would be shared by any scientific field, is the potential improvements in journal reviewing processes. Scientific publishing depends implicitly at least on the existence of shared objective criteria upon which to base the acceptance or rejection of articles. In practice it can be very hard to identify such criteria, and it can be hard to argue they actually exist given the variability by which decisions are made. In the absence of explicit rules reviewers assess the quality of proposed models based on their own tacit knowledge and more subjective preferences or weightings of different evidential factors. This requires a lot of specific expertise from the reviewers to perform. Having a well-defined standard would facilitate the reviewing process, make it much more transparent, and help prevent the publication of models of questionable quality.

A substantial benefit of a consensus on validation to systems biology, however, should be its potential to facilitate interaction and communication between modelers and experimenters working on systems biology projects. Systems biology is a highly interdisciplinary enterprise. Modelers, who come mostly from quantitative backgrounds like engineering or applied mathematics, depend on experimenters, who come mostly from molecular biology, for experimental data and biological expertise. Experimenters increasingly look to models in order to help interpret complex data sets and direct their experimentation. Systems biologists could learn from the ways in which collaboration between analysts and experimentalists has been envisioned in the engineering community, in particular in the design and execution of validation experiments. Oberkampf and Roy (2010) report an example from the Joint Computational/Experimental Aerodynamics Program (JCEAP) in the context of which guidelines directing the joint planning and execution of validation experiments have been developed. These guidelines concern, for example, the detailed quantification of experimental uncertainties and the balance between dependence and independence that should be maintained between analysts and experimentalists. Following these guidelines has been observed to lead to important synergistic effects between computation and experiment.

In general however collaborating fields may harbor disagreements over the standards for assessing the reliability of scientific claims. These standards are governed by what are often called in philosophical circles "epistemic values" (see Laudan, 2004). Disagreements over epistemic values between experimenters and modelers do exist and have been documented wherever mathematical modeling has entered biology. Rowbottom (2011) for instance explores through interviews some of the reasons experimenters give for often being skeptical of models produced in biological physics. These center on an unwillingness to accept the idealizations modelers need to rely upon to obtain tractable but informative models. As one experimental biologist he interviewed explained:

> *Biology tends to be idiosyncratic in the sense that I have some protein, I have a virus, whatever, there are some generic similarities and patterns that one can observe but if you look at the minutiae of mechanisms, there are little tweaks, they're all doing things a little bit differently and we get really het up and excited by those differences, and we emphasize these because that allows someone else to work on virus assembly somewhere else as long as it's a different virus. But I think in the physical sciences the tendency is to lump things and phenomena together and say these are all of a type and I'm understanding all of this in some way by doing the sorts of science I do.* (Rowbottom, 2011; 148)

Such a view asserts a belief by biologists that mathematical idealizations and approximations are likely incompatible with the variability and contingencies they are aware of through their experimental work. This can be used to discount the possibility that idealized mathematical models will really be useful in predicting experimental results, and can in turn be validated for such tasks.

Fagan (2016) argues further that different explanatory preferences in the context of stem cell modeling likely block the acceptance of the results of mathematical modelers by molecular biologists. Dynamical systems modeling of pluripotency in stem cells has largely been ignored by experimenters working on the same phenomenon. Fagan and her collaborators argue that part of the reason for this may well be that modelers tend to rely on deductive-nomological patterns of explanation rather than mechanistic ones to make their claim. Models in this field show how general characteristics of stem cells and cell development follow from general mathematical descriptions of cell behavior. On this

view of explanation experimental results are not accounted for unless there is some general principle that predicts them. Experimenters however tend to pursue more mechanistic accounts which aim to show how a particular sequence of cause and effect amongst the parts or components of a system give rise to the phenomenon under examination. These differences suggest implicit disagreements over what kinds of representations can serve as valid explanatory representations and different expectations over the roles models can play. They in turn help explain why communication and collaboration in these areas has been relatively sparse.

Cases from biological physics or stem cell modeling do not necessarily represent the attitudes of all modelers and experimenters nor the types of models being produced in systems biology. In both these cases models are often highly idealized. In systems biology many models and modelers are “mechanistic” in orientation, and the goal is to rely on detailed accounts of systems obtained from experimenters (their elements and interactions) in order to replicate their dynamics robustly. However, collaborative difficulties do occur in these areas of systems biology as well and some of these can be traced to implicit disagreements over how models can be validly used and under what circumstances they can be validated for these uses. Such disagreements have been recorded through empirical case studies of modelers and experimenters in systems biology (MacLeod and Nersessian, 2014; see also Calvert and Fujimura, 2011). Whereas modelers may rely on predictive success as a chief source of validation, experimenters will be more concerned with the quality of the underlying data relied on to both build and validate the model, and cite potential variability or experimental limitations and weaknesses as reason to be skeptical of what any model might contribute, even if a model seems to fit the data well. Such views have a tendency to crystallize in the form of stereotypes about models and modeling. To quote one modeler's perception from the Nersessian study of how the experimenters often saw models;

> *“Yes. They think of [systems biology] as something that's … just hooked up to … to, you know, match figures … that we're data fitting. So, for them, it's just like, you're using your data and then, you know, you're plugging in some numbers to fit … the output of your model to that. And then they would not pose a lot of faith in those models or what they predict”*

Models on this view are simply elaborate curve-fitting procedures that are likely to be very limited in what they can validly contribute. MacLeod and Nersessian (2014) have documented cases in which collaborations have largely failed when experimenters have lost faith in a modeler's ability to produce for them any new trustworthy information from their data sets, particularly in cases where modelers have requested large amounts of new data from experimenters just to make a model that can account for the original data sets. In such cases experimenters lose sight of any value to them in proceeding.

It is important not to generalize too much from such instances. There are many well-established productive collaborative relationships in the field of systems biology. Many collaborative problems occur when experimenters somewhat opportunistically seek out modelers, having no experience working with modelers, and a low interest in integrating modeling substantially in their practice. However, the extent to which there are real differences in epistemic values between the groups suggests the benefit of standardizing validation procedures, since standardization would at the very least bring to the surface underlying disagreements between the two groups, and promote a more robust discussion about them. Some of the difficulties that beset collaborative relationships can be traced to the lack of understanding experimenters have of modern mathematical techniques, including ensemble modeling (see Section 4.1) which offer ways of accommodating variability, or more rigorous validation schemes using sensitivity analysis which go beyond just matching model outputs to data sets. Transparent agreed-upon validation standards could be something experimenters could trust without having to have a deep epistemological understanding of how modeling techniques might handle structural and parametric uncertainty.

Further, shared validation standards which fixed acceptable criteria for validating a model could be of great help in allowing both modelers and experimenters to better align their goals with respect to modeling projects and accept the outcomes. The expectations both parties have of what modeling can achieve for them when entering projects could be much better managed, and shared progress much more easily identified. Overall this might help at least promote more effective collaboration between experimenters and modelers who lack experience with it. However it should be noted that achieving such standards may require some compromise between experimenters' and modelers' varying assessments of what is required for models to be considered trustworthy, particularly where the quality of the underlying data is concerned.

Beyond communication within collaborating groups, it is becoming increasingly common for models to be reused or modified by other research groups. The aforementioned standardization efforts and databases like the Systems Biology Mark-Up language explicitly serve the purpose of facilitating the exchange and further development of models by enabling the formulation of models in a common computational language. These models can then be easily accessed and used by modelers in their own research to a variety of different ends. As we noted already, however, standardized syntaxes for model formulation do not address issues of validity. When dealing with such complex systems as biochemical, cellular and physiological systems, knowledge requires patient careful accumulation. Successful models can serve as building blocks for others to work with and expand on or reuse for other goals (such as testing a medical device). As such, a common validation standard could be highly desirable by providing modelers clear indications of reliability of the models they are using. This would no doubt help increase the efficiency and value of model-exchange among different research groups, as well as providing a much needed index of the state of knowledge in the field.

Finally, validation standards become even more important as systems biology moves closer to the realm of clinical and medical application (a field which is commonly labeled 'systems medicine'). Quantitative models are already commonly being used to assess the absorption, distribution, and metabolism of pharmacological substances in the organism. One application of these models is the allometrical scaling of important pharmacokinetic parameters from animals to humans such that adequate dose regimens can be found. Systems medicine is expected to lead to models that can simulate the physiological mechanisms involved in disease and drug action. Initiatives such as the Avicenna Coordination Support Action are promoting the idea of *in silico* clinical trials which promise to make drug development both faster and more cost-efficient by replacing experiments *in vitro*, on animals, and on human patients. In this context standards for model validation become directly relevant for the safety and efficacy of biomedical products (Viceconti et al., 2016). From standard systems of validation government regulation can follow, which removes from the hands of individual scientists the task of having to adjudicate risk and uncertainty themselves when applying models for potential human use. Well-planned regulations can build in expert judgements from fields outside biology and medicine about what are socially, politically and ethically minimum acceptable levels of risk.

In summary, one can put together a case that validation standards in system biology, if they could be produced, would be desirable for several reasons: to accelerate peer review and ensure the quality of published models, to improve communication among scientists with different backgrounds in interdisciplinary research contexts, to facilitate the exchange and further development of models, and, eventually, to ensure the quality of *in silico* experiments in drug development in the context of systems medicine.

4. Constraints on standardization

Even though the case for objective validation criteria might appear favorable, we need to be cautious about whether validation criteria are really appropriate for the field in any broad way. There are some definite challenges standardization would have to face given the current state of systems biology. The nature of practices in the field, of biological systems themselves, and of the information modelers have to work with, raise specific challenges that any standardized approach would have to deal with.

4.1. The diversity of practice and purposes

Model validation is fundamentally purpose-dependent. Models need only be reliable enough to fulfill their particular intended purposes. One of the foremost goals of systems biology is to build models that are reliable enough for predicting responses of a system to perturbation. Achieving this is essential to achieving control of biological systems, an ambition used to sell and promote the field as superior to traditional experimental approaches alone. Predictive reliability is also essential for using models to test drugs and medical devices. For this kind of goal validation requirements need to be very strict and comprehensive, given the stakes, but also given the complex nature of the systems being dealt with, and the ever-present problem of biological variability. Models that can fulfill these goals will likely need to display a high degree of fidelity to the systems they model (MacLeod, 2016). Indeed, it should be noted that the current manual by the American Society of Mechanical Engineers (ASME) on Validation and Verification has deliberately refrained from setting any standards for interpolation or extrapolation of results outside a validation set, labeling it too much a matter of personal insight and judgment to be encapsulated within strict criteria (Schwer et al., 2006). Only the latest document from the ASME Verification and Validation 40 sub-committee is attempting to form standards for assessing the reliability of a predictive model (Popelar, 2013).

However, much modeling in the field does not follow a straightforward strategy of trying to optimize single models to capture the behavior of specific biological systems, but relies on a variety of more complex strategies and approaches for assembling models that can be used to extract useful information about systems despite their complexity and variability. Carusi (2014) suggests that in some attempts to capture and handle variability in a medically useful way the concept of an "external" validation standard is inimical to the practices modelers rely on. Modelers react to biological variability and the parameter uncertainty it generates by building populations of models which can cover a range of parameter values, and thus are better place to represent the full potential scope of system behavior across populations or sub-populations of individuals (e.g. Britton et al., 2013; Marder and Taylor, 2011). However, the process of establishing a reliable model ensemble may rely on an intensive process of parameter fitting and experimentation which, according to Carusi, serves to establish "comparability between the variability in the population of models, and that in the experimental data set" (Carusi, 2014; 33) through something like mutual exploration of the systems and the model. There is never any direct easy comparison to be made between these models and a set of experimental results since the criteria of comparison "for including and excluding models from the population, for dividing the population into sub-populations, for quantifying correlations so they can be used for comparison and so on" (Carusi, 2014 34), can only be established during the course of research and depend on factors local to the case. Hence validation grounds are not externally given, but internally established during the model-development process as researchers learn how the system and model are related. Carusi et al. (2012) make similar observations more specifically in the case of multiscale modeling which combines attempts to model phenomena that are governed across biological scales, such as intra-cellular, cellular and physiological, by integrating models of the processes functioning at each scale. When experiments at different scales have been performed under different conditions there is often no clear way of comparing integrated models with experiments. The result is variability and uncertainty which again drives both model development and the incorporation of new experimental techniques. Although Carusi does not explicitly suggest so, the strategies she describes of handling variability through ensemble and multiscale modeling, cast doubt on the idea that systems biologists can or should try to standardize validation criteria. A complex modeling situation has to feel out what an appropriate framework of comparison between sets of models and experiments might be. However, even though this modeling process seems inimical to the idea of shared validation guidelines, one has to take into account that some aspects of a multiscale setting might actually be conducive to transparent validation procedures. Submodels targeting processes at different scales that are validated independently mutually constrain each other and can thereby increase the reliability of the overarching model structure. In fact, the engineering guidelines mentioned above explicitly state that "[e]xperimental measurements should be made of a hierarchy of system response quantities, for example, from globally integrated quantities to local quantities" (Oberkampf and Roy, 2010; 418). Using a "building block, or system complexity hierarchy, approach" is recommended as the only feasible validation strategy for complex engineering systems (Oberkampf and Roy, 2010; 27).

Whereas in medical, particularly clinical contexts, reliable (well-validated) models are very desirable, many of the models built in systems biology are not necessarily built to pursue these goals, at least not initially (MacLeod, 2016). Indeed standards that validate models only for predictive purposes (i.e. requiring high fidelity models of systems) will likely be too strict for the majority of uses models are actually put to in the field. There is in fact a considerable diversity of model uses. For example models may be used to,

1. Explore the consequences of an abstract theoretical model in a practical setting.
2. Evaluate and test data collection and data analysis methods.
3. Optimize experiments and experimental procedures for model-building.
4. Check the consistency of data sets – whether sets are consistent with particular mechanistic assumptions.
5. Pose and test structural hypotheses which are outside the reach of current experimental capacities to recognize or test (e.g. using information that can only be assembled and put together computationally)

In many such uses the main aim of the modeling process is not to produce a predictively-reliable concrete representation of a system. These uses treat models in a more hypothetical manner refraining from assessing their predictive reliability. In the case of (2) and (3) for instance modelers use their models to help design

and test experimental protocols that are most effective for minimizing parameter uncertainty and maximizing model performance (Kreutz and Timmer, 2009; Bandara et al., 2009). In the case of (5) for instance the goal is to explore ranges of various hypotheses about model structure, and rule sets or subsets of them in or out, by virtue of whether they can produce good data-fitting solutions. Computation plays a powerful role in this respect, allowing modelers to experiment with different possibilities.

In one of the cases (an example of (4)) which was studied in the context of the aforementioned ethnographic project, a modeler had to integrate a signaling network model as input into a larger metabolic model. However, the modeler had three potentially inconsistent data sets governing the behavior of different upstream and downstream chemical elements in this network. She made it her principal goal to use the metabolic model to help her track down which data sets were consistent with the overall system behavior and which were not. To do this she truncated the signaling network into three alternatives, starting the network from each of the three different chemical elements that were modeled separately by the inconsistent data. She added each alternative network sequence to the metabolic system model and then simulated the whole combined model alternatively. She assessed the performance of each according to how well the overall system behaved in accordance with biological expectations. One alternative stood out as the only one capable of meeting all performance conditions. She concluded that the data set representing this alternative was the most consistent with overall data on the metabolic system, since it was the only set that could give the overall model (signaling network + metabolic network) behavior that fit the data. This inference was not based on the possession of a unique model or model ensemble that could be said to form a reliable representation of the overall system. In fact the researcher never produced or tested a single parameter set that fit the best option to the data. Indeed, using Monte Carlo techniques, she found a collection of inconsistent parameter sets that worked for this particular option. Her inference was to the effect that since no parameter sets could be found that could fit the other options to the data well, they could be ruled out.

In another prominent case Duarte et al. engaged in the construction of a very large scale model of yeast metabolism involving 1149 reactions and 750 genes (Duarte et al., 2004). The model drew together a large quantity of the available experimental work and knowledge on yeast metabolism. However, the goal was not to produce a correct model of yeast metabolism, it was rather to use the framework of the model to test the consistency and accuracy of current knowledge, and then draw hypotheses about where information was missing. For example quantitative data on growth rate of individual deletion strains were compared with equivalent *in silico* deletions. False predictions given by the model were then analyzed to explain why the model failed to predict the right growth states given that all known information had been included in it. Through careful analysis using the model, the authors derived that most false predictions were primarily caused by a lack of inclusion of cellular processes from outside metabolism in the network. Current understanding failed to account for whatever external and environmental regulation was modulating the system and forcing the right growth states.

These kinds of model uses rely much more on the modeling process than the presentation of a final result. They draw for instance on the ability of a computer to explore a wide-range of options and rule them in or out depending on whether they can meet certain conditions. They also depend on using erroneous models to track error in the state of knowledge. The standards for validating models or sets of models for the purposes of drawing these inferences about network structure or the consistency of data sets should be less strict than in the case of validating a model as "good" representation.[2] Firstly, as in the case of the researcher studying the cell signaling networks, these inferences are after all often drawn from examinations of what is possible, not what is actual. Such reasoning requires much less precision, and only needs to ensure that a correct answer is at least part of the set of models examined, even if it is not found. Secondly, there is less at stake. Such uses aim principally to contribute plausible hypotheses to the field that should then be picked up by experimenters. There is not the same risk that there might be if such models were being used to make medical treatment decisions. Indeed systems biologists during such work reiterate frequently that their results are hypothetical and need experimental confirmation. The value of doing such work as they often see it in fact lies in its ability to help direct experiment towards useful avenues that would have otherwise been impossible to discover. These roles of computational modeling can be referred to as heuristic, stressing the role models play as investigative tools or as part of strategies for getting insights on real systems and generating hypotheses, rather than substituting faithfully for real systems (MacLeod, 2016). It is likely no easy task coming up with criteria for validating models for the purposes of drawing inferences like those discussed above. How and under what conditions this kind of reasoning works has not been philosophically examined. Duarte et al. for instance take the 83% success rate of their model to be a good basis for thinking that model errors should correspond to inaccuracies in the model structure, rather than deeper more systemic problems (Duarte et al., 2004; 1306). But this standard of validation for these purposes is not justified in the paper, which is unsurprising. Such uses are in fact novel to computational approaches – they were not possible before –, leaving a gap between our theories of good validation or confirmation and what is happening in practice. Certainly their common use renders the issue of standardization much more complex, since there are likely many different kinds of hypotheses that can be drawn in many different ways from such subtle uses of models. Untangling all of these and giving good foundational justification for each will be a complex problem.

Finally, it is worth pointing out that in these contexts validation requirements may not in fact be beneficial and may even be detrimental to the discovery process. Whilst discussing the state of ecological modeling, which in many respects resembles the situation in systems biology, Rykiel notes that "modeling and the benefits to be gained from it can also be stifled by an overemphasis on model validation" (Rykiel, 1996). This might be because developing and investigating models is often considered a creative process, or "art", that should not be restricted by the boundaries of an imposed standard (Klipp et al., 2007). As Morrison & Morgan observe,

> There appear to be no general rules for model construction in the way that we can find detailed guidance on principles of experimental design or on methods of measurement. Some might argue that it is because modelling is a tacit skill, and has to be learnt not taught. (…) This omission in scientific texts may also point to the creative element involved in model building, it is, some argue, not only a craft but also an art, and thus not susceptible to rules. (Morrison and Morgan, 1999; 12)

In these respects, there are many reasons to think that practices in systems biology are not currently compatible with the imposition of any wide-ranging validation standards, even though such standards may be appropriate in mainstream engineering fields.

[2] See also Oreskes et al. (1994) who argue that such uses of models are the only empirically sound uses.

This is not to say that all aspects of current practices are inimical to validation standards, noting the potential value of them in multi-scale modeling contexts. Further, given this is a relatively new field (unlike most engineering fields), we do need to be wary of imposing standards that cut down the potential for creative exploration of different modeling approaches. Indeed much of the diversity and complexity of practices in systems biology may be a consequence of just this freedom and flexibility (MacLeod and Nersessian, 2016).

4.2. The technical feasibility of standardization

One aspect that makes modeling and model validation particularly challenging in systems biology is the extreme complexity and variability of living systems. Modelers are typically faced with systems involving a complex interplay of many components whose behavior and interactions are only incompletely described and understood. At the same time it is extremely difficult to get precise and quantitative measurements of properties that would be relevant for model building. Moreover, in contrast to engineering or more physics-based disciplines, such as the earth sciences, there are no 'first principles' that serve as a unique starting point and recipe for model construction. As a result, models in systems biology often involve hypotheses about certain components and their specific behavior. A mathematical consequence of this is that models in systems biology usually contain a large number of free parameters that need to be calibrated before they can be used to make quantitative predictions. Thus, before the accuracy of a model can be tested in terms of the standards proposed for engineering, systems biologists have to face the problem of the underdetermination of model structure and parameters.

Sometimes systems biologists are already satisfied if a model reproduces empirical data and take this as evidence that the proposed structure corresponds to the real mechanism. Goodness of fit alone is not sufficient, however, if the model contains unknown parameters due to the risk of overfitting the data. Overfitting occurs when the number of free parameters is large compared to the size of the dataset. In that case it is possible to achieve a very good fit simply by calibrating free parameters. The problem is that one captures features of the dataset that are not representative of the underlying mechanism (i.e. noise), and the model will perform badly when used to predict different datasets. Moreover, if a model is overfitted, the goodness of fit does not necessarily indicate that the structure of the model is related to the target system. So apart from not being predictive, an overfitted model is not useful for gaining mechanistic insight. One general lesson for modelers is that the complexity of a proposed model should always be in proportion to the amount and the quality of available empirical information.

The following quote shows that systems biologists are aware of the problem of overfitting, but often deal with it in informal ways:

> With >100 parameters at our disposal, is it any surprise that we can fit the model to the phenotypes of lots of mutants? After all, with four parameters, one is supposed to be able to fit an elephant. That is true, if the model is elephant shaped to begin with. But if the model is yeast shaped, it will not fit any particular elephant and vice versa. Hence, it is essential to prove that the model is yeast shaped by displaying a particular parameter set that brings the model into agreement with the observed properties of yeast cell growth and division. In our experience, many reasonable assumptions about the wiring diagram must be rejected because no amount of parameter "twiddling" can bring the model into agreement with the phenotypes of all the mutants (Chen et al., 2004; 3859).

As noted a more systematic and very common way for systems biologists to control for overfitting consists in partitioning empirical data into two sets. One part of the data (the "training set") is used to determine the parameters of the model, while the rest of the data (the "test set") is compared to the model's simulated behavior in order to assess the predictive performance of the model. This method can be criticized for two reasons. First, it still leaves to subjective judgement the question how good the fit should be. And second, it uses only part of the available data to determine the parameters of the model, which seems wasteful (see Hitchcock and Sober, 2004).

We will focus on one way of implementing a standard to deal with this problem, while being aware that there may be other and possibly better ways. We consider it plausible, however, that the problems raised in our discussion will be similar to those that any other reasonable validation standard would face. The standard that we propose instead is well-known and comes from a part of statistics called "model selection theory". This is an information theoretical approach that allows one to find models that are maximally justified by the data. As the name suggests, model selection is not only concerned with fitting free parameters to data, but also with choosing the best model among a set of candidates.[3] It controls for overfitting by penalizing models with many free parameters in a statistically rigorous way. One important result that lies at the heart of model selection theory is expressed by Akaike's Information Criterion (AIC):

$$AIC = -2\log\left(\mathrm{L}\left(\widehat{\theta}|\mathrm{y}\right)\right) + 2\mathrm{K}$$

The first term is the likelihood of the model (i.e. the goodness of fit) with a vector of estimated parameters $\widehat{\theta}$ given data y, while K is the number of estimable parameters of the model. Ideally, one should find the model that minimizes *AIC* because it is the one that strikes the best balance between goodness of fit and model complexity. Model selection thus implements a version of Ockham's razor by ensuring that a more complicated model is chosen only if it provides significant improvement over a simpler model. Note that the *2K* term penalizing model complexity is not arbitrarily introduced, but derived from rigorous statistical principles. Parsimony is thus not imposed as an independent epistemic virtue, but emerges as a direct consequence of making the best use of available information. *AIC* can be interpreted as an estimate of the expected, relative distance of a candidate model to the unknown "true" mechanism that generated the data. "Relative distance" in this context means that it is not possible to determine how close a candidate model is to the truth, but given two models one can decide which one is closer. To put it differently, model selection does not tell us what the correct model looks like, but gives us the model whose inferences are maximally supported by the available data. As Burnham & Anderson stress, "whether any of the models is actually 'good' depends primarily on the quality of the data and the science and a priori thinking that went into the modeling" (Burnham and Anderson, 1998; 20). We have chosen the approach of model selection to illustrate the problem of a validation standard because it is a very general framework that avoids many of the criticisms which might be put forward against other possible candidates.

For obvious reasons model selection is not useful if the structure of the correct model is already known. In that case one might be more interested in determining how well the model performs in

[3] Model selection does not necessarily mean that one model eventually has to be selected as 'the best,' it also allows for drawing inferences from multiple selected models (for details see Burnham and Anderson, 1998).

absolute terms. Then one could go ahead and attempt to apply systematic validation procedures known from engineering, such as setting up a validation metric and performing specific validation experiments. This is, however, rarely the case in the practice of systems biology. It is indeed a very common situation to have different alternative hypotheses about a mechanism of interest.

One problem that may be raised is the fact that many different kinds of models are used in systems biology, in addition to the diversity of purposes and modeling strategies we saw in section 4.1, and not all of them might be amenable to the same kind of validation method. In particular, many models are deterministic models (e.g. ODE models) and different from the statistical models that model selection has been developed for. For complex stochastic models, however, it might not be possible to even specify the likelihood function required for applying a criterion like *AIC*. At least for deterministic models there is a straightforward technical solution: one can modify them by introducing stochastic terms that account for measurement error and thereby make them amenable to validation. In general, the framework of model selection provides methods "for nearly all classes of models we might expect to see in the theoretical or applied biological sciences" (Burnham and Anderson, 1998; 29). It seems, however, that the problem of standardization reappears in a different guise for the statistician who has to ensure that the ways of implementing methods for different types of models yield equivalent results. It should also be mentioned that the numerical implementations of validation algorithms, where analytical solutions do not exist, can be computationally expensive for large models.

Another objection might be that validation in terms of a statistical comparison of finished models against data is not adequate for systems biology. One might argue that validation is part of the modeling process, as we saw, and that it is a misconception that models are first built and then validated. Instead, one might consider modeling as an iterative process without definite end point (Carusi, 2014). Here one possible response is that model selection is not only concerned with assessing finished models in isolation, but can also serve as a guide for the process of model construction. The following quote proposes one possible mode of action:

> Development of the a priori set of candidate models often should include a global model: a model that has many parameters, includes all potentially relevant effects, and reflects causal mechanisms thought likely, based on the science of the situation. (…) At some early point, one should investigate the fit of the global model to the data (…) and proceed with analysis only if it is judged that the global model provides an acceptable fit to the data. Models with fewer parameters can then be derived as special cases of the global model. This set of reduced models represents plausible alternatives based on what is known or hypothesized about the process under study (Burnham and Anderson, 1998; 17).

So the strategy of comparing alternative candidate models actually fits quite naturally with the informal way in which many systems biologists develop their models in practice.

The diversity of model uses in systems biology that was already discussed in Section 4.1 also represents a technical obstacle to standardization efforts. Does the very idea of a standard not contradict the idea that validation requirements are strongly dependent on model use? And aren't statistical validation methods exclusively concerned with the predictive use of models? Our example of a validation standard can serve to illustrate that this is not necessarily the case. As mentioned before, model selection is used to find the model that serves best the generation of valid inferences. This notion of validity goes beyond mere prediction of data and covers also inferences regarding the structural parameters of the model. Moreover, by comparing and ranking alternative structures, model selection also contributes to the investigation of the underlying causal structure by suggesting which candidate mechanism best explains the observed phenomena. It seems plausible that the validity of inferences to an extent also translates to the hypotheses that are being generated by investigating a model. So even for heuristic uses it might be useful to stick to the general ideas behind model selection and make sure to balance model complexity against the amount and quality of available data.

The issue of data quality, however, raises the perhaps most severe problem for validation standards. The typical situation in bottom-up systems biology is that one has relatively detailed background knowledge about the mechanism of interest, while the data to specify model parameters and to test the model are comparably scarce. This situation is thus exactly the reverse of the ideal proposed by model selection theory. Moreover, the data are often of low quality and stem from heterogeneous sources (e.g. from different model organisms or from a mixture of *in vitro* and *in vivo* experiments). Sometimes data are qualitative, for instance when a biologists can determine only that a gene is "on" or "off" without specifying expression in quantitative terms. Although some of these problems can probably be solved technically by appropriately taking into account measurement uncertainty and biological variability, these issues represent serious problems for a quantitative statistical framework like model selection theory. This suggests that for a significant part of systems biology standardization regimes may not set realistic goals in current circumstances. In particular, it means that it would be difficult to apply the quantitative validation concepts from engineering even in the favorable case where a single model structure can be identified.

Our discussion of this one candidate validation standard highlights several important issues. Within a sound statistical framework validation goes beyond checking how well a model fits the data. In order to be able to draw valid inferences from a model, one has to make sure that the complexity of the model is adequate given the quality of data and available background information. As such a validation standard is not necessarily restricted to the predictive use of models, but can also capture other modeling purposes. However, even though it would be technically possible to implement a standard like model selection for most types of models in systems biology, it appears that there is generally a mismatch between the complexity of the models that are typically investigated and the data currently available for their validation. Insisting on a validation standard would be beneficial only for those areas of systems biology where sufficient amounts of reliable and quantitative data are available. This might for instance be the case for the recent attempt of building a whole-cell model of the bacterium *Mycoplasma genitalium* that relies on an impressive amount of quantitative information assembled from over 900 primary sources, reviews, and databases (Karr et al., 2012).

5. Conclusion

In this paper we have explored some of the principal challenges that systems biology would face attempting to standardize validation in the field. The nature of practices in systems biology, of the subject matter being dealt, and the data available and obtainable in the field, make it hard to imagine that we can pin down criteria for validating systems biological models in a meaningful way at this point in time. The purposes to which models are applied in the field are diverse and have different epistemic requirements. There are many problems that would pervade a standardized platform for measuring validity based solely on how well the models fit or

accord with the data, given the quality and quantity of data usually available to systems biologists. Measuring the validity of a model in just these terms is likely to be uninformative or inaccurate independent of the assessment of the quality of the data and of the theory upon which a model is based. However the inclusion of these in any assessment criteria likely creates a much more complex task for standardization, and one can argue that such assessments might well be better left to the intuitive abilities of experts to pull together and integrate diverse evidential factors.

Having said this there is no reason to rule out that validation might be plausible in certain restricted domains, such as physiological models for testing medical devices for which the quality and availability of data might be very good. But for the most significant part of systems biology standardization has to be looked at cautiously. Getting to a position where standardization might be possible will require not only better development of modeling techniques, but also better philosophical understanding of modeling and confirmation practices in the field, and ultimately the development of better interdisciplinary relationships so that richer and better quality data that can support modeling efforts becomes available. In this regard, the organization of interdisciplinary projects in systems biology might profit from existing guidelines for collaborative efforts in engineering, such as the ones mentioned in Section 3. Standardization needs to be balanced against the complex practices and complex modeling structures (like ensemble and multiscale models) systems biologists and experimenters rely upon particularly when trying to produce medically applicable models. It also needs to be balanced against the legitimate interest the field might have in flexibility insofar as best practices for handling different types of biological systems given different types and quality of data have not yet been fully realized.

Nevertheless, despite our cautious message, thinking about criteria for model validation is useful even at this stage and even without the aim of implementing rigorous validation standards. It promotes the development of better insight into practices in the field and the challenges systems biologists face. Further our discussion indicates that there are certain aspects about modeling that are largely independent from modeling context and modeling purpose, such as the issue of balancing model complexity against the amount of available empirical information. We would suggest as a result that developing shared informal guidelines based on rules like these might provide many of the benefits of a true validation standard, while respecting the diversity of modeling in systems biology.

References

Bandara, S., Schlöder, J.P., Eils, R., Bock, H.G., Meyer, T., 2009. Optimal experimental design for parameter estimation of a cell signaling model. PLoS Comput. Biol. 5 (11), e1000558.

Britton, O.J., Bueno-Orovio, A., Van Ammel, K., Lu, H.R., Towart, R., Gallacher, D.J., Rodriguez, B., 2013. Experimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. Proc. Natl. Acad. Sci. U. S. A. 110, E2098–E2105. http://dx.doi.org/10.1073/pnas.1304382110.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Multimodel Inference. Springer, New York. http://dx.doi.org/10.1007/b97636.

Calvert, J., Fujimura, J.H., 2011. Calculating life? Duelling discourses in interdisciplinary systems biology. Stud. Hist. Philos. Biol. Biomed. Sci. 42 (2), 155–163.

Carusi, A., 2014. Validation and variability: dual challenges on the path from systems biology to systems medicine. Stud. Hist. Philos. Biol. Biomed. Sci. 48, 28–37.

Carusi, A., Burrage, K., Rodríguez, B., 2012. Bridging experiments, models and simulations: an integrative approach to validation in computational cardiac electrophysiology. Am. J. Physiol. Heart Circ. Physiol. 303 (2), H144–H155.

Chen, K.C., Calzone, L., Csikasz-Nagy, A., Cross, F.R., Novak, B., Tyson, J.J., 2004. Integrative analysis of cell cycle control in budding yeast. Mol. Biol. Cell 15, 3841–3862. http://dx.doi.org/10.1091/mbc.E03.

DoD, 1994. DoD Directive No. 5000.59: Modeling and Simulation (M&S) Management. http://www.dtic.mil/whs/directives/corres/pdf/500059p.pdf (Accessed 19 August 2016).

Duarte, N.C., Herrgård, M.J., Palsson, B.Ø., 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res. 14, 1298–1309. http://dx.doi.org/10.1101/gr.2250904.

Fagan, M.B., 2016. Stem cells and systems models: clashing views of explanation. Synthese 193 (3), 873–907.

Frigg, R., Hartmann, S., 2012. Models in science. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2012 Edition). http://plato.stanford.edu/archives/fall2012/entries/models-science/ (Accessed 19 August 2016).

Gunawardena, J., 2010. Models in systems biology: the parameter problem and the meanings of robustness. In: Lodhi, H.M., Muggleton, S.H. (Eds.), Elements of Computational Systems Biology. John Wiley & Sons, Hoboken, pp. 21–47.

Hitchcock, C., Sober, E., 2004. Prediction versus accommodation and the risk of overfitting. Br. J. Philos. Sci. 55, 1–34.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., Covert, M.W., 2012. A whole-cell computational model predicts phenotype from genotype. Cell 150 (2), 389–401.

Kleindorfer, G.B., O'Neill, L., Ganeshan, R., 1998. Validation in simulation: various positions in the philosophy of science. Manag. Sci. 44 (8), 1087–1099.

Klipp, E., Liebermeister, W., Helbig, A., Kowald, A., Schaber, J., 2007. Systems Biology standards – the community speaks. Nat. Biotechnol. 25, 390–391.

Kreutz, C., Timmer, J., 2009. Systems biology: experimental design. FEBS J. 276 (4), 923–942.

Küppers, G., Lenhard, J., 2005. Validation of simulation: patterns in the social and natural sciences. J. Artif. Soc. Soc. Simul. 8 (4), 3.

Laudan, L., 2004. The epistemic, the cognitive, and the social. In: Machamer, P., Wolters, G. (Eds.), Science, Values, and objectivity. University of Pittsburgh Press, Pittsburgh, pp. 14–23.

MacLeod, M., 2016. Heuristic approaches to models and modeling in systems biology. Biol. Philos. 31 (3), 353–372.

MacLeod, M., Nersessian, N.J., 2014. Strategies for coordinating experimentation and modeling in integrative systems biology. J. Exp. Zool. Part B Mol. Dev. Evol. 322, 230–239. http://dx.doi.org/10.1002/jez.b.22568.

MacLeod, M., Nersessian, N.J., 2016. Interdisciplinary problem-solving: emerging modes in integrative systems biology. Eur. J. Philos. Sci. 6 (3), 401–418. http://dx.doi.org/10.1007/s13194-016-0157-x.

Marder, E., Taylor, A.L., 2011. Multiple models to capture the variability in biological neurons and networks. Nat. Neurosci. 14, 133–138. http://dx.doi.org/10.1038/nn.2735.

Morrison, M., 2015. Reconstructing reality. Models, Mathematics, and Simulation. Oxford University Press, Oxford.

Morrison, M., Morgan, M.S., 1999. Models as Mediating Instruments. In: Morgan, M.S., Morrison, M. (Eds.), Models as Mediators. Cambridge University Press, Cambridge, pp. 10–37.

Oberkampf, W.L., Roy, C.J., 2010. Verification and Validation in Scientific Computing. Cambridge University Press, Cambridge.

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

Popelar, C.F., 2013. Verification & validation in computational modeling of medical devices (V&V-40). In: FDA/NIH/NSF Workshop on Computer Models and Validation for Medical Devices. ASME subcommitee and guide, Silver Spring, MD.

Peterson, A.C., 2006. Simulation uncertainty and the challenge of postnormal science. In: Lenhard, J., Küppers, G., Shinn, T. (Eds.), Simulation: Pragmatic Constructions of Reality – Sociology of the Sciences, vol. 25. Springer, Dordrecht, pp. 173–185.

Rowbottom, D.P., 2011. Approximations, idealizations and 'experiments' at the physics–biology interface. Stud. Hist. Philos. Biol. Biomed. Sci. 42 (2), 145–154.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Modell. 90, 229–244. http://dx.doi.org/10.1016/0304-3800(95)00152-2.

Schwer, L.E., Mair, H.U., Crane, R.L., 2006. Guide for verification and validation in computational solid mechanics. Am. Soc. Mech. Eng. ASME V&V 10, 2006.

Suárez, M., 2004. An inferential conception of scientific representation. Philos. Sci. 71 (5), 767–779.

Viceconti, M., Henney, A., Morley-Fletcher, E., 2016. In Silico Clinical Trials: How Computer Simulation Will Transform the Biomedical Industry. Research and Technological Development Roadmap, Avicenna Consortium. http://dx.doi.org/10.13140/RG.2.1.2756.6164.

Voit, E.O., 2012. A First Course in Systems Biology. Garland Science, New York.

Weisberg, M., 2006. Robustness analysis. Philos. Sci. 73 (5), 730–742.

Winsberg, E.B., 2010. Science in the Age of Computer Simulation. The Chicago University Press, Chicago.

Zi, Z., 2011. Sensitivity analysis approaches applied to systems biology models. IET Syst. Biol. 5 (6), 336–346.