Forthcoming in *Philosophy of Science*

COVER PAGE

Title:

Building Simulations from the Ground-Up: Modeling and Theory in Systems Biology

Authors:

Miles MacLeod
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, GA, 30332-0280
USA
mmacleod@cc.gatech.edu

Nancy J. Nersessian (corresponding)
School of Interactive Computing,
Georgia Institute of Technology

Atlanta, GA, 30332-0280
USA
nancyn@cc.gatech.edu

Abstract. In this paper we provide a case study examining how integrative systems biologists build simulation models in the absence of a theoretical base. We call this model building from the 'ground-up'. Lacking theoretical starting points ISB researchers rely cognitively on the model-building process in order to disentangle and learn the complex dynamical relationships governing the chemical elements of their systems. They build simulations in a nest-like fashion by pulling together information and techniques from a variety of possible sources and experimenting with different structures in order to discover a stable robust result. Finally, we analyze the alternative role and meaning theory has in systems biology when expressed in the form of canonical template theories like Biochemical Systems Theory.

Acknowledgements. We appreciate the support of the US National Science Foundation in conducting this research (DRL097394084). We thank the directors and members of the research labs in our investigation for welcoming us into their labs and granting us numerous interviews. We thank the members of our research group for contributing valuable insights, especially Lisa Osbeck, Sanjay Chandrasekharan, and Wendy Newstetter. We thank two anonymous reviewers and the editor of the journal for their helpful comments and advice.

MANUSCRIPT

1. Introduction. Recent years have seen a developing discussion on the role and epistemology of simulation in modern scientific practice as a subject worthy of its own attention, distinct from experimentation and modeling. With some exceptions, particularly in the case of ecology and social science, most attention in the philosophy of science literature has been given to physics-based cases, such as meteorology, climate science, and nanoscience, where theories such as fluid dynamics and quantum mechanics provide essential material from which particular simulation models take shape. This paper aims at a new contribution to this discussion by looking at cases where researchers build simulations without a theoretical starting point and without "theoretical articulation". We focus on ethnographic studies of cases from the relatively nascent field of integrative (computational) systems biology (ISB)[1], for which the principal product of investigation is a simulation. Systems biologists regard themselves as engaged in a genuinely new enterprise which cannot be subsumed within the bounds of established biological fields like molecular biology or

[1] The designations "systems biology" and "integrative systems biology" appear to be used interchangeably in the field. ISB is often used to stress the computational dimension of systems biology. We use ISB because our labs are part of an institute which uses this designation. Systems biology is in fact heterogeneous. Much attention has been given to the "top-down" or "systems theoretic stream" (see O'Malley and Dupré 2005). Our analysis focuses on the bottom up/middle-out stream.

physiology. They build computational models of the fluxes of material and information in complex genetic and metabolic networks. However instead of articulating theories into informative computational models, systems biologists compose their models by collecting essential dynamical and structural information themselves from a variety of sources including from their own simulations in highly iterative and intensely cognitive process (see also Chandrasekharan and Nersessian 2011).

This paper has two principal aims. Firstly, we illustrate with the help of a case study of one modeler's PhD research how systems biologists, at least those in the bottom-up strand of systems biology (Bruggeman and Westerhoff 2007), build their simulations from the "ground-up" by piecing together in a nest-like fashion principles from molecular biology, experimental results, literature surveys, canonical models, and computational algorithms, to create representations of systems in data poor environments. Such a study helps broaden our understanding of the range of scientific practices involved in the production of simulations. This is required if we want to attain both a general perspective or understanding of what simulation is across a broader range of disciplines and also of the important roles of simulation for novel disciplines that have come into being because of computer technology, for which simulation is the central defining methodology (see also Lenhard 2007).

Secondly in light of a developing discussion about the role of theory in modern computationally and data driven biology (see for instance Leonelli 2012a, 2012b) we use this case study and broader opinions from within the field to reflect more broadly on the

different meanings and functions of "theory" in systems biology, particularly with respect to attempts to establish mathematical templates for the field. More generally systems biology aims to invert the dictum that theory produces or facilitates simulation. It is the rhetoric of systems biology that a theory of biological systems will emerge from model-building and simulation.

2. What do we Know about Simulation in Science? We already have some good philosophical analysis of the role of simulation in various scientific contexts. For instance Humphreys, Lenhard, Winsberg, and Parker have detailed the importance of simulation to the success of various fields like nanoscience, shock wave theory, and meteorological and climate models, by applying theories of structure and dynamics from quantum mechanics and fluid dynamics to complex phenomena (see Humphreys 2002, 2004; Lenhard 2006, 2007; Winsberg 1999, 2001, 2003, 2009, 2010; Parker 2006, 2010a, 2010b). They have discussed particularly the various epistemic roles simulation plays in comparison with ordinary model building and also with experimentation. Both Lenhard and Winsberg claim strongly that simulation brings a unique methodology to scientific practice that requires its own epistemic principles. Simulations for instance necessarily require that a model be transformed into computational language which imposes its own constraints such as computational speed and memory capacity and requires its own unique decision making principles such as those for discretizing continuous differential equations. Such principles, Winsberg (1999, 2010) contends, are not given by theory. He describes simulations rather as "semiautonomous" of theories, by which he means that although one starts from a theory, "one modifies it with

extensive approximations, idealizations, falsifications, auxiliary information, and the blood, sweat, and tears of much trial and error." (2003, 109)

Winsberg further points out that simulation models are often generated because data are sparse. Simulations step in to provide a source of data, and thus a way of exploring a phenomenon for which experimental access is otherwise limited. Lenhard (2007) in turn has emphasized the role simulation experiments play exploring the consequences of a theoretical model, through a process of what he calls "explorative cooperation". This invokes many extra-theoretical techniques that allow a theory and its models to be articulated to fit the phenomena. Such explorations are largely driven "from above" by the data, and fitting the data, not by the theory. Lenhard (2006) with help from Humphrey (2004) has also argued that simulation is effecting a change in deeper aspects of the epistemology of science. In some cases a model built out of a particular theory acts in an "epistemically opaque" way, meaning the theory cannot be used to produce an understanding of a phenomenon by, for instance, analytically solving the equations. The theoretical model developed from a theory is so complex that the only way to discover its consequences is through simulation. Hence simulation provides a resource for exploring complex phenomena through the dynamics of a model. As such a kind of law-based or even mechanistic understanding cannot be had. Instead researchers produce a pragmatic level of understanding through their ability to control and manipulate a phenomenon through simulation.

Along a different but related line, Parker has been grappling with how climate science and meteorology researchers, while relying on available physical theory in building simulation models, handle the indecisiveness or indeterminancy of this theory for pinning down the best representational strategies of the many it makes available (Parker 2006, 2010a, 2010b) . Theory in these fields offers a model-building framework but no way of evaluating easily the best of the alternative models that can be built using this framework, given available data. These researchers grapple with the resulting uncertainty by collectively employing multiple models making different, sometimes conflicting, structural and parametric assumptions and approximations in order to bootstrap climate and weather predictions. These ensembles are thus particular strategic responses to the inadequacy of theory. However, Parker questions strongly whether such a strategy can justify its predictions without recourse to a theory that can determine whether an ensemble adequately represents the space of possibilities or not.

We think many of thephilosophical insights about simulation apply in the case of systems biology and have genuine claim to descriptive generality across scientific fields which employ simulation. Simulation must be approached on its own terms and cannot be reduced to modeling or experimental practices. We think however there needs to be more discussion of the role theory and theoretical models play in the production of simulations. For the most part the central concern of both Lenhard and Winsberg, following on from earlier on work with modeling in general (see Savageau 1969b; Morgan and Morrison 1999), has been to demonstrate how simulation is autonomous or semiautonomous from background theory by elaborating the extra-theoretical factors that contribute to simulation production. These are

usually justified by how well they produce simulations that match the phenomena not by their accordance with theory. Winsberg (1999, 277) admits himself however that his conclusions about simulation do not necessarily apply in cases where theory does not play an explicit part.

On the other hand a collection of researchers studying modeling in ecology and economics have begun to attend to how simulations in these fields come to be constructed often without any process of theoretical articulation. Instead simulations in these fields rely upon diverse methodologies and non-standardized techniques tailored to the specific problems at hand (Peck 2008). These observations are reflected in the increasing reliance on agent based modeling in ecology (Peck 2012). ABM simulations rarely start from a systematic theory, "because of the difficulty of finding general ecological theories that apply across a wide range of organisms and the deeper suspicion that we do not have any in the same sense as models found in the physical sciences." (Peck 2012, 5) ABM's provide a means for modeling without theoretical knowledge, because "we usually do know something about agential behavior through data…" (Peck 2012, 5) Ecologists themselves are grappling with how to standardize the resulting eclectic mix of practices and descriptions that characterize ABM models which lack the communicability and analyzability of traditional analytic models.(Grimm, Berger et al. 2006; Grimm, Berger et al. 2010) We think the same applies to systems biology where the models also do not admit of analytical solutions and simulation is necessary, although in this case the models for the most part are ODE rather than agent based models.

Our aim here is to advance the discussion of how simulation models are constructed without recourse to a body of theory that provides the structure and dynamics of the phenomena under investigation through drawing on our ethnographic analysis investigating the fine details of ISB model-building processes and the cognitive affordances of those processes. Nor is there generally accepted theory applying across the domain of such systems that specifies the routines and formalisms that can and should be followed and applied to model such systems. The responses the participants in our study make to the lack of theory they can articulate into models of their systems, as well as to limited data and the complexity of the subject matter, are no doubt shared with other fields like ecology in which modelers find themselves in similar situations. Thus, this case study should have broader relevance for a philosophical understanding of simulation model-building practices across a range of modern simulation-driven scientific contexts.

Our first task however is to come to some terms with what systems biology is and how it works.

3. What is Integrative Systems Biology? In principle ISB is relevant to any biological systems from ecological systems to genes, although in practice, and certainly in the labs we study, most research is directed at cellular or intracellular systems. As such systems biology presents itself as new approach to understanding and controlling systems like gene regulatory or metabolic networks through computer simulations of large scale mathematical models. It applies computer technology to what was formerly considered inaccessible complexity in

order to generate representations of the dynamics of these networks, particularly material and informational fluxes through a network of cells or biomolecules. Importantly systems biologists claim they are working to understand the *system-level properties* of a network or the dynamic patterns of a network rather than pulling apart the details of the interactions of its components. In Kitano's terms, a diagram of gene and protein interactions provides insight on how a system works, "it is like a road map where we wish to understand traffic patterns and their dynamics." (2002, 2) Systems biologists contrast this approach with traditional molecular biology that pursues these elements such as gene and protein interactions "in isolation," that is, through in vitro analyses of the properties and structure of individual bio-molecules. ISB studies these molecules in their systemic contexts as part of a functioning unit.

ISB is not however a homogenous enterprise as we have discovered in our own investigations. Researchers in the field and philosophers identify two broad strands, namely *top-down* and *bottom-up* systems biology (see Bruggeman and Westerhoff 2007; Krohs and Callebaut 2007). Top-down relies upon high-throughput technology which takes repeated simultaneous measurement of large numbers chemical concentrations within cells, enabling a great quantity of time-series data for many cellular elements to be collected. Computer methods can then be used to "reverse engineer" system structure by sorting through correlations in those elements. Bottom-up systems biology however "reproduces" (or simulates) systems with dynamical models by drawing upon knowledge of network structure

and the kinetic and physicochemical properties of its components. Its data tend to come from molecular biology sources.

Our investigations principally concern bottom-up systems biology. Since such metabolic and genetic systems are generally extremely complex cases of continuous nonlinear systems computer technology is required not only to run the dynamical models, but also to provide the power to graph their dynamics (highly important for understanding networks), estimate parameters and parameter sensitivities, and calculate steady states and analytic network measures like flux control coefficients. Since much of bottom-up systems biology works in relatively data poor environments, algorithmic techniques of evaluating networks for their solvability and estimating parameters is a computationally intensive process. One of the ultimate aims is to be able to gain sufficient fine-tuned control and understanding of the input and outputs of a biological network to be able to manipulate it to produce desired outcomes, such as increased production of a good chemical (e.g., a biofuel) or reduction of bad one (e.g., a cancer-causing agent).

Systems biologists in the labs we have been studying fit for the most part within the bottom-up tradition, although with distinct methodological differences in the ways the labs work (MacLeod and Nersessian 2014; Nersessian and Newstetter 2014). One thing they do share is that most researchers come almost uniformly from engineering rather than biological backgrounds, and bring metaphors and analogies such as electrical circuit analogies and concepts like noise, control, robustness, redundancy, modularity and information with them

to understanding biological systems. The heavy systems engineering and control theoretical perspective is very important to piecing together the thought processes driving systems biology in practice, particularly the belief that system-level understanding is not contingent on a detailed theory of the mechanical interaction of network components. It is more fairly described as "mesoscopic" or "mesoscale" or "middle-out modeling" rather than bottom-up (see also Westerhoff and Kell 2007; Voit, Qi et al. 2012). Mesoscale modeling works on the basis that nonlinear and complex system dynamics are emergent properties of network structures that do not require detailed lower level theory (whether physics, biochemistry or molecular biology) to reconstruct and understand.

4. Building from the Ground-Up. In simulations derived from theory, particularly physical theory, the modeling process starts with the theory, or at least with a theoretical model that traces its origins to a theory. Of course "theory" is a contested and multifarious category, and we do not mean to use it uncritically. What we are referring to here is a broad eclectic understanding of theory as reservoir of laws, canonical theoretical models, principles of representation (such as boundary conditions) and ontological posits about the composition of phenomena that guide, constrain, and resource the construction of models in diverse disciplines across a wide spectrum that model physical systems. From Cartwright (1983) on a number of philosophers have claimed that we need to be circumspect about the role theory plays in physics and physics-dependent disciplines, and it cannot be said in any sense that models are simply *derived* from theories (see for instance Morgan and Morrison 1999). We start from the view nonetheless that in these physics-based fields theory is playing some role

and this role is essential, even if it is not as strong as we might have once presumed. Mesoscopic modeling, however, starts without such a reservoir of fundamental models, laws and principles.

Our investigations take the form of a 4-year ethnographic study of two biochemical systems biology laboratories. One lab ("Lab G)[2] does only modeling in collaboration with experimental researchers outside the lab. The other lab ("Lab C") conducts its own experiments in the course of building models. The labs are populated largely by graduate students. We use ethnographic data collection methods of participant observation, informant interviewing, and artifact collection. In both labs we conducted unstructured interviews with the lab members and some collaborators outside of the lab. We collected and analyzed relevant artifacts including powerpoint presentations, paper drafts, published papers, grant proposals, dissertation proposals, and completed dissertations. Thus far, we have collected 97 interviews and have audiotaped 24 research meetings. From these investigations we have built up a detailed understanding of how simulations are constructed by the researchers. (Chandrasekharan and Nersessian 2011; MacLeod and Nersessian 2014) We think their practice can suitably be described as *modeling from the ground-up.*[3] By modeling from the

---

[2] This research is supported by the US National Science Foundation and is subject to human subjects restrictions that prevent us from identifying the researchers.

[3] Note modeling from the ground up should be distinguished from Keller's (2003) notion of "modeling from above". Modeling from above is a strategy as Keller describes it, that aims

ground-up we identify a pair of entwined modeling practices. Firstly modelers assemble the structure and local dynamics of the phenomena being modeled largely from scratch (and not from a theoretical starting point) by pulling in empirical information from a variety of sources and piecing it together into an effective representation by using a variety of assumptions and simplifications and mathematical and computational techniques. Every part of this process is open to choice about the methods to use and the information to include. Secondly modelers rely cognitively on the iterative building process and their preliminary simulations to understand their systems and adapt their representations of them to their particular epistemic goals. This feature signals the more intensive role simulation and model-building have as cognitive and epistemic resources in the context of a modern computer-driven discipline.

5. Case Study: G12 and Atherosclerosis. To flesh out our ideas we have chosen as an exemplar one of the graduate researchers we followed from Lab G. We refer to her as G12. Her experiences and model-building practices are strongly representative of the practices of

---

to simulate the phenomenon itself not by trying to map its underlying causal structure or dynamics, but rather by generating the phenomenon from a simple yet artificial system of interactions and relations. Modeling from the ground-up is what she would call “modeling from below” in that is relies on information about the causal structure and dynamics of a system’s compositional elements. Both however may begin from non-theoretical starting points.

the other modelers we have tracked across both labs. As her research unfolded, G12 gave us detailed descriptions of the processes by which she put together her models, demonstrating the ways in which our ISB researchers assemble their models from the ground-up in a nest-like fashion. That is, just as a bird will gather just about anything available to make a stable nest, G12 pulled together and integrated bits of biological data and theory, mathematical and computational theory, and engineering principles, from a range of sources in order to generate stable robust simulations of biological networks. This was an intensely constructive process that lacked a coherent theoretical resource as a basic platform from which she could work. Instead she had to bring together and integrate whatever techniques and information were available and continually adapt and adjust her own representation of her systems within the affordances and constraints of these techniques. Our reconstruction of her model-building process arranges it into three segments, but this should be understood an only approximate since in fact model-building is not a linear process.

Her ostensible research topic was the relation between oxidative stress and the generation of monocytes implicated in the generation of plaques on the vascular smooth muscle cells of blood vessel walls. Such activity is thought the main cause of conditions like hypertension and atherosclerosis, which result from the hardening of arteries and loss of elasticity caused by these plaques. G12 produced three models in completing her PhD research covering different aspects of this problem in different cell lines.

Model 1	A model of the pathway in endothelial cells through which particular mechanosensitive genes produce oxidative stress in response to different vascular fluid stresses
Model 2	A model of the assembly and disassembly mechanism in smooth muscle cells of an enzyme critical in redox cycling and oxidative stress
Model 3	A model of the pathway by which a hormone A produces oxidative stress in smooth muscle cells

[Insert table 1 here]

Table 1: A description of G12's models.

Like virtually all of the student researchers in both these labs, G12's background is in engineering, not in biology. In terms of the biology she always felt she started "from zero" with every new modeling task on a different biological system. Indeed the model building

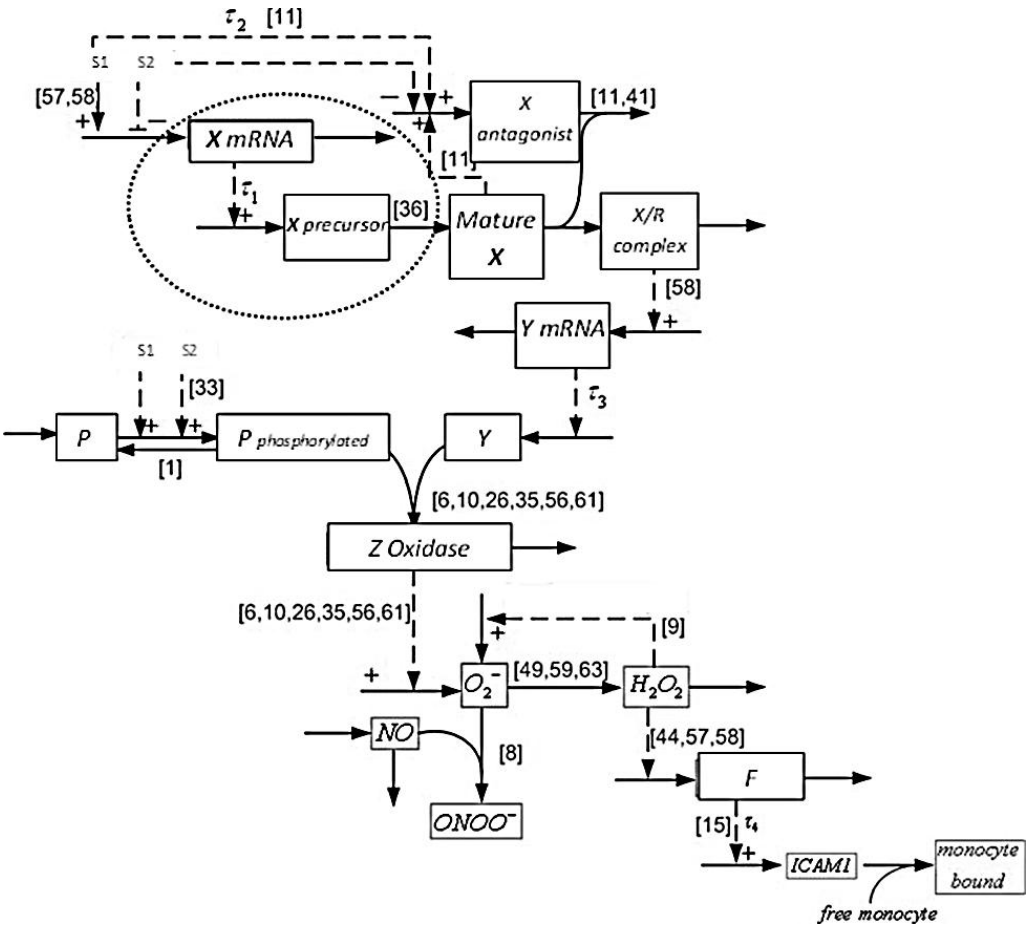
process is a way through which Lab G researchers learn the relevant biology. G12's process was built around three main interrelated tasks. Firstly the construction of a biological pathway, secondly the construction of a mathematical representation of that pathway, and thirdly an estimation of parameters for the mathematical representation that would fit it to the dynamics observed in experiment.

5.1. Constructing the Pathway. In each case G12 had to begin by putting together a pathway of the molecular interactions and signaling relations composing the biological network. In each case the pathway given to her by her collaborators was insufficient given her modeling goals, and she was forced to pull in whatever pieces of information she could find from literature searches and databases about similar systems or the molecules involved in order to generate a pathway that mapped the dominant dynamic elements. For example in the case of her first pathway diagram for Model 1 she was given a pathway by her experimental collaborator, but quickly realized this pathway would not be adequate in itself to produce a dynamic model that could explain the dynamics of the relationship between her mechanosensitive gene X and ROS (reductive oxygen species) production.

"Yeah so actually first I read the literatures and I find, I need to know even though he gave me the pathway, I need to know whether there are any other contributing factors, whether there are other important factors that are missing."

The focus of her collaborator on this project as a molecular biologist had been assembling the main direct chain of causal influence, not on what other contributing factors would be

affecting the system's dynamic behavior, which was necessary for G12's task. Reading through her collaborator's and other literature G12 began to "figure out" and pull together what other factors and interactions were involved and thus produce a more comprehensive pathway. Starting out with only five variables, the final production had expanded to fourteen.



[Insert Figure 1 here]

Figure 1. Pathway diagram used by G12 in the construction of Model 1. The chemical names of the reactants have been substituted by us. The two mechanically different types of stress

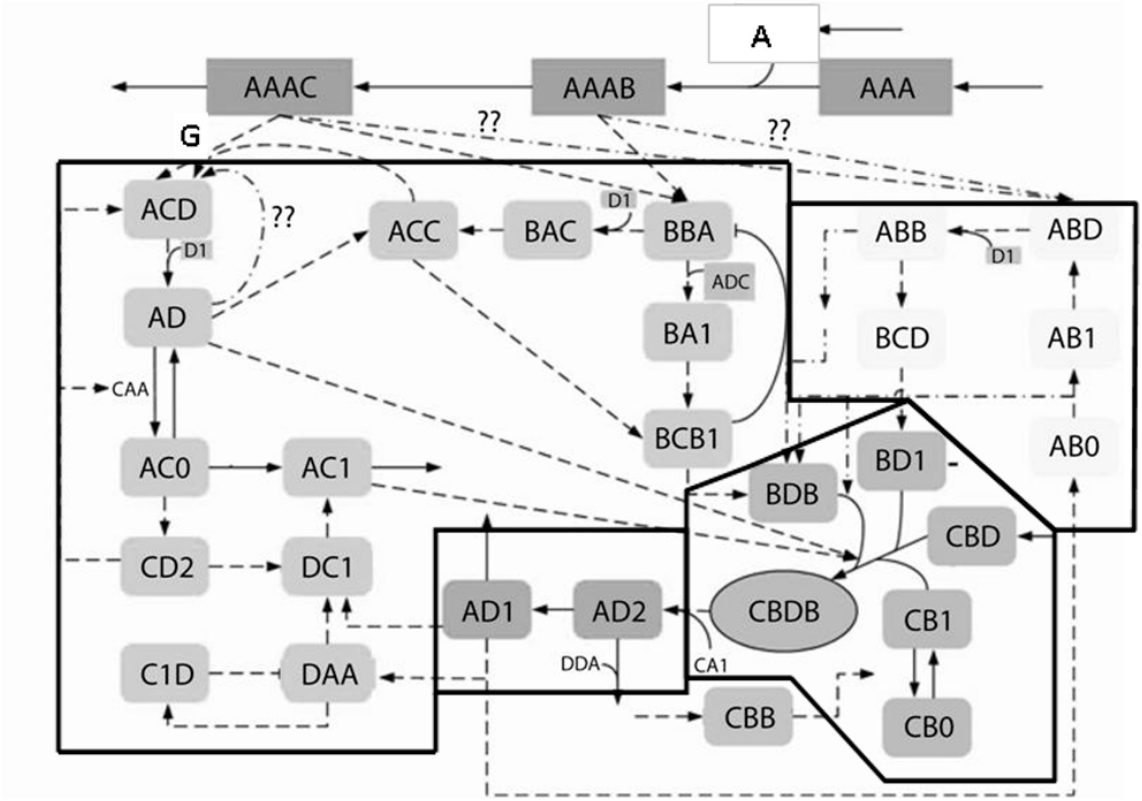
applied to the system are S1 and S2. Numbers in parentheses refer to the specific references from the literature used in building out the pathway. Solid lines indicate material flow through the pathway, while dashed lines indicate signaling without the flow of material. τ indicates time delay. The circled relation between X mRNA and X precursor she hypothesized herself.

As she put it “But um he didn’t say clearly how the things interact. Even in the, this diagram (Figure 1), there’s something missing. There are missing steps I combine them together because I don’t know how exactly this upstream one affect the downstream one. So in his papers there are several things I connect them together, and I need to make sure.... how from this, the upper one goes to the downstream. So I need to go verify.”

Verification in this context means either checking the literature or asking her collaborator. At some point running preliminary simulations she noticed that there needed to be another component that was not accounted for in the literature she had on the system in the form of signaling between X mRNA and the X precursor (circled above). She inserted this signaling relation because she had inferred the potential interaction from reading about X and its properties. This addition was made without explicit experimental evidence for its role in this system.

Building these pathways was not simply a passive process of recording information. It involved significant effort on her part to produce pathways that were sufficient to reproduce the dominant network dynamics while staying within the data constraints and without the network representation becoming too complex for parameter estimation. In this first model particular simplifications and assumptions were applied in order to be able to represent the system as simply as possible so that her mathematical frameworks and parameter estimation techniques could handle the lack of good data on the system. This is a common type of accommodation our modelers bring in. It included only considering two elements of the Z oxidation system, the protein P and enzyme Y, by assuming on basis of information in the literature that other elements did not interact with X and probably did not strongly affect the pathway. The Model 1 pathway in the final result thus tracked for G12 the dominant interactions that occur from the introduction of two mechanically different types of stress (S1 or S2) to the production of excess plaque inducing monocytes.

Model 2 was an attempt to construct a pathway lacking in the literature for the activation of an enzyme CBDB that would be critical for Model 3. This network is built into her pathway diagram for Model 3 as the module of the bottom right corner (see Figure 2 below). In the interest of space we‘ll just focus here on Model 3. In building the pathway for Model 3 she brought a range of techniques to bear. This model aimed to model the role of hormone A, rather than fluid stress (Model 1), in the generation of oxidative stress and inflammation.



[Insert Figure 2 here]

Figure 2. A preliminary pathway diagram for model 3. Metabolite names have been replaced. G12 color coded her modularization of the network, which we have marked off with boxes. The top left box is the BCB1 module, the top right is the AB0/AB1/ABD signaling cascade system, the bottom right is the CBDB activation system (Model 2), and the last middle-bottom box represents the reduction oxidation module. This diagram is not the completed one but represents a stage in her work. Particular unknowns are represented with '??'.

She built this pathway as her own simplified representation “having collected and interpreted all related information from literature as well as having consulted our experimental collaborators.” The pathway was broken up into manageable modules, designated in Figure 2. Constructing this complex pathway required pulling together bits and pieces of information and a full variety of techniques, ranging from offlining variable and blackboxing systems, to informed guesses. She discovered herself the role of CB1 from the literature. In a major reconstruction, after running her first simulations, her collaborator told her of the importance of BDB which G12 did not know about. This had to be added and the pathway adjusted.

In the case of the redox buffering system operating in this model (the module AD1/AD2 in Figure 2) she had black boxed many of the components (components she was aware of from another researcher in a neighboring lab) particularly the systems of degradation for AD1 both because of their complexity but also because the data she had were too poor to account for the elements of these systems. “I don’t want to separate them because it’s so complicated and then you don’t know the real concentration or the real molecule numbers.” Instead she gave the buffer the system capacities she deemed necessary by imposing for instance constant degradation rates.

Secondly she made calculated decisions about offlining variables in the model and thus treating them as controlled inputs (rather than system-embedded variables). “Offlining” involves decoupling a variable from the dynamics of the system because of a lack of

information about how it connected to the system or because of the complexity of connecting it into the system, and instead feeding in its values by hand in order to get the model to fit the data. She expressed a plan to do this for a particular enzyme G. “So I just decouple, okay, I don’t have this, if uh, I would just input this and input have this effect on this enzymes and I would just input this, the effect of the enzyme directly.” She would take the value to input take either from the literature or guess and experiment with the model.

She also had to apply her own limited biological knowledge, sometimes in contradiction to what the literature was suggesting in order to limit her model. For instance she had originally accepted the idea that protein BDB should be being recycled: that is, BDB would at some point dissociate and its component antecedents like BCB1 would reemerge (see Figure 2). Overall the number of proteins would not change at the basal level which was stable. She thought, however, that this would involve continuous translocations of the components from across the cytosol and plasma membrane and back, which made her uncomfortable because it seemed to be invoking too many assumptions she could not justify. "So before I use recycling assumption in something like this but right now I think probably it's not appropriate, I don't know why, just a feeling..." . Instead it seemed reasonable to her to just have the protein degrade at a certain rate, balanced by the rate of incoming components.

In each case these pathway representations were composed from a variety of sources and with a variety of techniques. G12 built in information that emerged in the process of building dynamical models of the pathways, but she also controlled the pathway representations

through different techniques to restrict these representations to what could be modeled manageably and reliably.

5.2. Mathematical Frameworks and Mathematical Simplifications. G12 also had to decide what kind of formalisms to use to model the data. This choice for systems biologists ranges over the very fundamental choices of whether to use phenomenal models like agent based models or mechanistic models, whether to use discrete or continuous models, or spatial (PDE) versus non-spatial (ODE) models, stochastic or deterministic models, multi-scale or uni-scale models. In our labs they almost all decide to try to put together an ODE system given its relative conceptual simplicity and potential informativeness, as well as the range of computational and mathematical infrastructure that exists for estimating parameters and analyzing model dynamics. Choosing an ODE framework opens up another range of very important choices over whether, for instance, to model the system as at steady state (static), or away from any equilibrium, whether to use a straightforward mass-action/stoichiometric model, a mathematical template like biochemical systems theory (BST: see the next section) that tends to build in or average over the details of interactions, or a mechanistic model that builds in equations closer to the molecular biological details in the form of rate laws of individual enzymatic interactions such as Michaelis-Menten or ordered uni-bi mechanisms. Such decisions are made as part of a calculated integration of a wide variety of constraints modelers have. Certainly no modeling strategy is universally the best or most accurate under all circumstances. It depends both on the objective the researcher has and certainly on the completeness and the nature of the data they have.

In G12's case she oscillated between the use of mass action models and BST in the form of generalized mass action models (which use power law representations of interactions). For her first model little dynamic data were available. Most were qualitative, in the form of observations from experiments that *x* up or down regulates *y*. At the same time the model involved multiple time-scales, since it involved genetic transcription, protein transcription, and enzyme-catalyzed biochemical reactions. This meant that she needed a framework in which time-delays could be easily incorporated.

For these tasks the GMA (generalized mass action) framework was appropriate. This framework models the system as ordinary differential equations, modeling flux as the sum of inputs minus outputs.

$$\dot{X}_i = \sum_{k=1}^{T_i} \left(\pm \gamma_{ik} \prod_{j=1}^{n+m} X_j^{f_{ikj}} \right)$$

Its prime attractive property in these circumstances is that it can account for a large variety of dynamics by modeling the flux of dependent variables as power laws. This means that even if the nature of interactions between elements are not well known for the system, there is a good chance that the systems dynamics will still be accounted for within the range of realistic parameters for the system. In her second model (Model 2) however the actual quantitative dynamical data on the system (as opposed to its composition) were extremely limited, so G12 had to fashion her aims and the model to meet this constraint. In this case because the pathway was well drawn-up in most details, a simple mass-action model sufficed and would

invoke fewer unknown parameters. Secondly because the critical enzyme CBDB (see Figure 2, the pathway for Model 3) mostly operated in a basal condition of steady state in the body, she could derive the structural features she was looking for at steady state. From steady state she was able to derive the relationships between fluxes as a set of linear relationships, and then analyze them according to various optimization conditions: namely what would produce the maximum responsiveness in least time to any stimulation of the system out of the basal condition. Hence in this case, in addition to bringing in a particular mathematical framework to help her construct a representation, she also added certain mathematical simplification techniques to transform her problem into one she could handle with such limited data.

As mentioned the information from Model 2 about CBDB activation was used to help construct that module in Model 3 on the assumption that the steady-state derived structure and parameters would hold too in the dynamic context of the system undergoing stress.

5.3. Parameter Reduction and Parameter Estimation. As all systems biologists we have interviewed contend, the hardest part of the model building process is parameter determination once a mathematical framework is in place. This process requires considerable ingenuity on their part and never follows any easily prescribed path. It was no different G12's case. For the first model G12 had neither time-series data nor rate constant data based on traditional Michaelis-Menten analysis of enzyme-catalyzed reactions. As such she was forced to develop her own parameter-fixing mechanism using novel reasoning about this problem. She reasoned that since most time-scale processes in her system happen over

hours, the fast reactions in the network should be at steady-state. By setting their fluxes to zero she was able to develop constraints among her variables that cut down the number of free parameters. Then, by making plausible inferences about the kinetic orders of her free parameters, she was able to simulate the dynamics observed in cells with a satisfactory degree of accuracy, although not without some discrepancies.

The third model is of most interest here because she had to assemble so many resources in order to handle the parameter problem. One technique she used to make up for the lack of data, which is not uncommon, was to use data pertaining to the same metabolic elements from other cell lines. In one case, for instance, she used neural cell data to get parameters for a particular metabolite in smooth muscle cells. For this she had to make the inference that the systems in these diverse cells were reasonably consistent or homologous. Another technique she used was sensitivity analysis, a technique that allows one to judge the most sensitive parameters of a network. Insensitive parameters can be set to a default value because they do not affect the network dynamics. Sensitivity analysis is part of the techniques systems biologists use to 'shrink their parameter spaces.' However the most significant work for G12 was done by Monte Carlo simulation.

The AB0/AB1/ABD signaling cascade system (the top right module in Figure 2) was considered uncertain due to a lack of information about the dynamic profile of the metabolite ABB in response to this signaling. G12 devised three mechanisms for the transmission of signals through the cascade to ABB based on the literature. It should be said though that

there was no available empirical evidence about this mechanism in the literature, only various speculations that G12 had interpret mathematically in order to construct these alternatives for her model. This meant that G12 had the task of trying to fix parameters of the whole model with each of the three versions, and testing which one met her prescribed conditions best. This is a classic seat of the pants type calculation in systems biology, one that grapples with uncertainty by testing across model candidates. While the BCB1 and AD1/AD2 module above had sufficient available experimental information, the AB0/AB1/ABD cascade system and CBDB activation module did not. Neither of them could be estimated since they possessed too many parameters and biological information and observations were limited to some qualitative evaluations and hypotheses. Thus G12 had to incorporate Monte Carlo techniques to complete her model-building process. She could not do this, however, without finding first ways of shrinking the parameter space further. She brought in numerous assumptions in order to achieve this goal. For instance, she assumed that the CBDB activation assembly and disassembly system (enclosed in a box on Figure 2) operates at steady state maintaining a low level of oxidants. Since such a system normally keeps levels of CBDB at stable levels in its basal (unstimulated) condition, this was a justifiable and important assumption. It allowed G12 to generate linear flux conditions that reduced the number of unknown parameters. Further she used her already established knowledge of the CBDB system to ignore the small contribution of one these parameters which left her with a total of 7 independent parameters. Using biological information she had available that some reaction pairs were biochemically similar, she equated them. She was,

thus, able to reduce her parameter space to just 4 free parameters. Her Monte-Carlo sampling strategy sampled these parameters over high values and low values (normally distributed), producing 16 combinations of parameters for each model and a total of 48 possibilities for the three model candidates. She evaluated these against qualitative data for the parameters (up-down regulations) that she was able to extract from experiments and the literature, and from system conditions such as that under hormone A treatment, the ROS concentration of AD1 and AD2 reaches a new higher plateau at 30 minutes. She found 3 candidate parameter sets (all from her second model choice for ABB) that gave good results. A unique solution, as she readily admitted, was never going to be possible once she resorted to Monte Carlo methods, but she managed to narrow the space of model possibilities nonetheless.

5.4. The role of simulation. Simulations were not simply the end-phase of G12's research, nor as we have intimated, were these steps simply a linear sequence of tasks. The pathway representation was shaped by issues of parameter availability and available parameter estimation tools. Likewise pathways were tailored to fit the capacities of the mathematical frameworks and whatever mathematical tools she could bring to bear. At the same time these frameworks determined the extent of the parameter fixing problem she faced. She kept these elements in dialog during the model-building process. In this regard, simulations have an important functional role for our researchers for learning how to assemble information and construct a computational model that gives the right kind of representation. Information is assembled in the course of an exploratory process that involves preliminary simulations, both computational and pen and paper, and subsequent refinements and revisions in order for the

researcher to build up his or her own understanding of the dynamics and relevancies of particular pathway elements. Sometimes this can involve making reasonable judgments about elements not discussed in the literature that a modeler hypothesizes must actually be playing a role, or about feedback relations that are not documented. In G12's case once the Monte Carlo simulations were done for her third model it became clear that there was a discrepancy in the model depending on whether upstream or downstream data were fed into the AB0/AB1/ABD cascade module. Here two data sets were available, one for AB0 and another for the AB1 activation mechanisms. AD1 seemed to be being stimulated by pathways other than just the one they had. This led her to a literature search for possible molecular candidates that be activating AD1 independently. This search revealed ACC as a candidate, which she hypothesized to be a missing signaling element. Making a pathway for it in the model resolved the inconsistencies with the data.

5.5. Conclusion. G12 was faced in each case with complex problem solving tasks defined by complex systems and sparse data, hoawever she could not start with a theory of the structure and dynamics of her system that she could apply and articulate. Rather she had to pull together information and techniques from a variety of sources and integrate the steps in her model-building process in order to find a productive manner of assembling all these elements together in a nest-like fashion that would produce a stable robust result. This information came in the form of bits of biological theory of molecular interactions, bits of data and research from the literature, and some collaborative assistance, and was assessed in the context of choices about mathematical frameworks (mass action or generalized mass action

and power laws), mathematical techniques of simplification (steady state linearizations), and algorithmic parameter estimation techniques (Monte Carlo simulations) and various other assumptions to get these to work. At the same time what we have observed with researchers like G12 is the extent to which this process of incremental model-building and its attendant processes of simulation are the means through which these researchers learn to understand their systems, which in turn allows them to make better judgments about what to include and exclude and which tools and techniques help and which will not. As per the nest analogy, they work out the best or most stable way to pack the pieces together. Thus there is a highly cognitive dimension to simulation which, in ISB practice, is integrated as an essential part of building models of complex systems that lack any core basis of theoretical knowledge that holds across the domain of such systems and could provide or prescribe a starting point for modeling them. This point goes further than Lenhard's idea of an explorative cooperation between models and simulations (Lenhard 2007). Simulation in ISB is not just for experimenting on systems in order to "sound out the consequences of a model" (181). Simulation is fundamental for assembling and learning the relevant ontological features of a system. Simulation's roles as a cognitive resource make the construction of representations of complex systems without a theoretical basis possible (see also Chandrasekharan & Nersessian 2011).

Similar observations however have been noted with respect to ecology. As mentioned this is perhaps unsurprising given the similar complexity of the problems and lack of generalizable theory that characterize both fields. As with Peck's point that "[t]here are no formal

methodological procedures for building these types of models suggesting that constructing an ecological simulation can legitimately be described as an art,"(2008, 393) our modelers, too, describe their modeling practices as an "art." Likewise the ISB modeling we have observed is always an individual project in which each modeler chooses the methods and strategies he or she thinks best resolve the problem without any formal procedure governing the process. A major benefit of an ethnographic approach is that it exposes the often hidden creative choices that are "rarely disclosed in formal descriptions of model-building" (Peck, 395).

These parallels with ecology suggest that there is a deeper fact about the methodologies employed in these kinds of simulation-building contexts. The creative model-building processes employ a mix of mathematical and computational expertise, "considered judgment" (Elgin 1996), and art.

6. Theory in Systems Biology. Although we have examined how the systems biologists in our labs work without the reservoir of essential information about their systems that physical theory would provide on how to build models in a domain, we do not mean to be arguing that concepts of theory have no currency within ISB. As has been long understood by philosophers "theory" has diverse connotations in practice. Systems biologists use "theory" to refer to a plethora of different bits and pieces of information from, for example, molecular biology and mathematics, which they pull together to build their simulations.

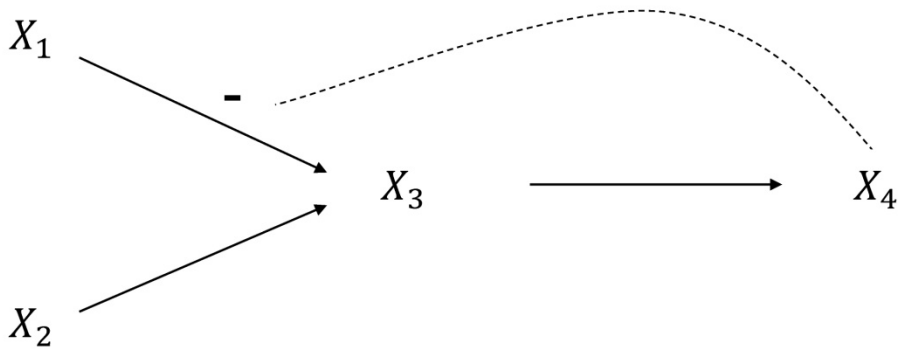
However one of the principal elements of simulation building that go by the name "theory" in the context of systems biology are "canonical" ODE models. An example is Biochemical

Systems Theory (BST), which is used at times by Lab G participants to model their systems, and comes with a collection of analytical tools. [4] BST posits that the processes of a given metabolic/signaling network can be modeled dynamically as power law functions. There are two types of canonical representation, Generalized Mass Action (GMA) and S-System formats. Firstly GMA uses ODEs of this type:

$$\dot{X}_i = \sum_{k=1}^{T_i} \left(\pm \gamma_{ik} \prod_{j=1}^{n+m} X_j^{f_{ikj}} \right)$$

Where T_i is the number of terms in the *i*th equation. The X_i's represent the concentrations of the various molecular elements in the network. Each separate flux in and out is represented by an individual power law.

[4] BST was developed originally by Michael Savageau in 1969. (see Savageau 1969a, 1969b and 1970) and has been developed further by Eberhard Voit and colleagues.



[Insert Figure 3 here]

Figure 3. A sample metabolic network with feedback signaling

For example, in the case of the system in Figure 3 we model $\dot{X}_3$ as:

$$\dot{X}_3 = \gamma_{31} X_1^{f_{311}} X_4^{f_{314}} + \gamma_{32} X_2^{f_{324}} - \gamma_{34} X_3^{f_{333}}$$

The S-system format, on the other hand, pools all the effects of ingoing and outgoing flux into one separate term:

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}$$

In this case;

$$\dot{X}_3 = \alpha_3 X_1^{g_{31}} X_2^{g_{32}} X_4^{g_{34}} - \beta_3 X_3^{h_{33}}$$

Applying these canonical models is a matter of filling in the network details (which X_i's are interacting with which others), the parameters with which such interactions take place (the kinetic orders *h*,*g*, and *f*'s) and rate constants (the α, β, γ's). Canonical templates fit with Humphrey's concept of a computational template (2004). These templates are not however constructed from theoretical templates or laws, a case Humphrey's acknowledges as perfectly possible with the example of the Poisson process (88). We think the case of BST illustrates how computational templates get used and constructed differently in different fields, particularly when compared to canonical cases in the physical sciences. The most important point is that BST and its templates are not starting points of the modeling process, in the way a theoretical template or theory often is in the physical sciences. They are chosen if suitable during the model-building process, as we have shown in the case study, and modified as required. The framework, in other words, is added to the nest in the model-building process if it serves that process given the available data and the aims of the modeler. This usage identifies a different role that these templates play in simulation building which in turn signals a difference of intent of a theory like BST.

Principally, rather than providing essential structural and dynamical information about a system out of which a simulation can be built through approximation and simplification, BST provides a schema *of* approximation and simplification for organizing effectively the

information that a systems biologist has assembled. As such BST offers a theory of how to model effectively and efficiently the dominant dynamics of complex biological systems where data are sparse and systems complex. It instructs systems biologists how to average over (or linearize) the complexities of the mechanisms that govern the interaction between elements into order to robustly match the dominant or key dynamic relationships in a network. The power laws above like $flux_{in} = \prod_{j=1}^{n+m} X_j^{g_{ij}}$ are just such averaging relations. They are motivated by sound mathematical principles of approximation theory (first order Taylor series approximations) given the biological assumption that the range of variation of concentrations of biological molecules is generally limited. BST assumes that for most biological metabolic and signaling networks suitable parameters can be found that will model to a high degree of accuracy the complex fluxes of the networks because the BST framework is flexible enough to capture them within the range of suitable parameter choices. In terms of efficiency, power law representations are telescopic. Power laws can be used generally to represent lower levels and S-systems keep their structure when variables are substituted for lower level systems. Further such systems are readily analyzable mathematically and system parameters can be clearly mapped to experimental measurements (Voit 2000).

In contrast then to the role of theory in physics-based modeling, a theory like BST does not provide essential structural or dynamical information for describing the system. Dynamical fluid laws, for instance, contain essential information about the laws that govern the

interactions of elements of mass or flux in a fluid system. In ISB the pathway and the nature of interactions have to be assembled by the researcher through the techniques and practices mentioned above, although once such templates are chosen they become routinely part of the cognitive systems researchers employ for exploring and revising their understanding of the biological system, and for deciding how to approximate its structure and dynamics so as to provide the best model for their given aims. Each canonical model has its own advantages and disadvantages and domains over which it functions well. Power-law models are more accurate than lin-log models for small substrate concentrations, but not at high substrate concentrations. However as (Voit 2013, 106) points out, in practice the choice of canonical model is less important than "the connectivity and regulatory structure of the model".

As result one cannot describe the fixing of parameters for these canonical models as a process of "theory articulation" as it is described in physics-based contexts (see Winsberg 1999, 2003) . The issue for systems biologists is not how to apply a particular background theory to a specific problem by adapting theoretical principles to the circumstances. Their project is to build a higher level or system level representation out of the lower level information they have. Canonical templates mediate this process by providing a possible structure for gluing together this lower level information.

This use of computational and mathematical templates is in fact at the heart of the philosophies and aims of ISB, which rejects an overly reductionistic approach to model-building. As mentioned above, rather than bottom-up modeling, what systems biologists do

can often be better described as mesoscale or middle-out modeling. Mesoscale models are "models of an intermediate size that are neither purely empirical nor contain complete mechanistic detail." (Noble 2008) These models try to ignore the complexities at the lower-level in order to capture the general features of a system's dynamics and thus gain generic control over the network by mapping its important input and output relations. As Voit puts it, "All interactions must follow the laws of physics and chemistry, but accounting for every physical and chemical detail in the system would quite obviously be impossible – and is also often unnecessary. Canonical modeling provides a compromise between general applicability and simplicity that circumvents some of these problems by using mathematically rigorous and convenient approximations." (2013, 99) This creates an epistemology that favors techniques that average over the complex mechanisms at work at lower levels and provides ways to build functional models not just with sparse data but also with more manageable complexity as we saw with G12. Considered in this light, systems biology inverts the familiar relationship whereby theory contributes to models which in turn produce simulations. In ISB it is simulations, according to its promoters, which are being used to build up theory, although such theory is again not necessarily the same thing as theory in physics. Major figures in systems biology like Kitano (2002) and Westerhoff and Kell (2007) have been keen to emphasize the new promise of biological theory that the computational study of biological systems brings.

"The goal of systems biology is to offer a comprehensive and consistent body of knowledge of biological systems tightly grounded on the molecular level, thus enabling us to fully integrate biological systems into more fundamental principles." (Kitano 2002, 2)

As Kitano points out however this does not mean we should expect something equivalent to fundamental laws. He prefers the example of theories like Bardeen, Cooper, and Schrieffer theory (the Cooper pairs theory) that explains super-conductivity. Such a theory imposes its own specific constraints on a background of fundamental physical principles. Kitano sees systems biology similarly as discovering the structural constraints on the interactions of molecules, governed as they are by chemistry and physics. These constraints are often referred to as design and operating principles: particular structural features of networks that exemplify general evolutionary solutions to biological principles. These cross species' lines. Design principles should help explain why nature "has settled on" some methods of building and running networks rather than others and thus, ultimately, compose a "biological theory."

7. Conclusion. The philosophy of science has long discussed and explored the relationships between theory and models, a discussion which has now been extended to the relationship of theory and models to simulation. This discussion of modeling and simulation, however, remains incomplete. New computational sciences are breaking with the traditional patterns of model-building activity philosophers have typically observed. Ideas from engineering, mathematics, and computation, and new large scale data collecting methods are being combined with biology to generate new interdisciplinary and transdisciplinary computational

biological fields. In one such, namely integrative systems biology, we have detailed a practice of *model-building from the ground-up*, which builds simulation models without a theoretical starting point. These practices assemble a wide variety of resources together from different sources, and leverage on the cognitive affordances of this incremental, nest-building process and simulation itself, in order to develop an understanding of a complex system's dynamics and ultimately provide an adequate representation. As such systems biology raises new issues about the role of theory in model-building and epistemic issues about how understanding and credibility can be generated without a theoretical base.

Bruggeman, Frank J. and Hans V. Westerhoff. 2007. "The Nature of Systems Biology." *TRENDS in Microbiology* 15(1): 45-50.

Cartwright, Nancy. 1983. *How the laws of physics lie*: Cambridge Univ Press.

Chandrasekharan, Sanjay and Nancy J. Nersessian. 2011. "Building Cognition: The Construction of External Representations for Discovery." *Proceedings of the Cognitive Science Society* 33: 264-273.

Corbin, Juliet and Anselm Strauss. 2007. *Basics of qualitative research: Techniques and procedures for developing grounded theory*: Sage Publications, Incorporated.

Grimm, Volker, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K. Heinz and Geir Huse. 2006. "A Standard Protocol for Describing Individual-Based and Agent-Based Models." *Ecological Modelling* 198(1): 115-126.

Grimm, Volker, Uta Berger, Donald L. DeAngelis, J. Gary Polhill, Jarl Giske and Steven F. Railsback. 2010. "The ODD Protocol: A Review and First Update." *Ecological modelling* 221(23): 2760-2768.

Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69(S3): S1-S11.

Kitano, Hiroaki. 2002. "Looking Beyond the Details: a Rise in System-Oriented Approaches in Genetics and Molecular Biology." *Current genetics* 41(1): 1-10.

Krohs, U. and W. Callebaut. 2007. "Data without Models merging with Models without Data." *Systems Biology: Philosophical Foundations (Boogerd FC, Bruggeman FJ, Hofmeyer J-HS, Westerhoff HV, eds)*: 181-213.

Leonelli, Sabina. 2012a. "Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies." *International Studies in the Philosophy of Science* 26(1): 47-65.

Leonelli, Sabina. 2012b. "Classificatory Theory in Biology." *Biological Theory*: 1-8.

MacLeod, Mies and Nancy J. Nersessian. 2014. "Integrating Simulation and Experiment: Hybrid Research in Systems Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* (under review).

Morgan, Mary S. and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Nersessian, Nancy J. and Wendy C. Newstetter. 2014. "Interdisciplinarity in Engineering". Cambridge Handbook of Engineering Education Research. J. Aditya and B. Olds. Cambridge: Cambridge University Press: (in press).

Noble, Denis. 2008. *The Music of Life: Biology Beyond Genes*. New York: Oxford University Press.

O'Malley, Maureen A. and John Dupré. 2005. "Fundamental Issues in Systems Biology." *BioEssays* 27(12): 1270-1276.

Parker, Wendy S. 2006. "Understanding Pluralism in Climate Modeling." *Foundations of Science* 11(4): 349-368.

Parker, Wendy S. 2010a. "Predicting Weather and Climate: Uncertainty, Ensembles and Probability." *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* 41(3): 263-272.

Parker, Wendy S. 2010b. "Whose Probabilities? Predicting Climate Change with Ensembles of Models." *Philosophy of Science* 77(5): 985-997.

Peck, Steven L. 2008. "The Hermeneutics of Ecological Simulation." *Biology & Philosophy* 23(3): 383-402.

Peck, Steven L. 2012. "Agent-based Models as Fictive Instantiations of Ecological Processes." *Philosophy & Theory in Biology* 4: 1-12.

Savageau, Michael A. 1969a. "Biochemical Systems Analysis: I. Some Mathematical Properties of the Rate Law for the Component Enzymatic Reactions." *Journal of Theoretical Biology* 25(3): 365-369.

Savageau, Michael A. 1969b. "Biochemical Systems Analysis: II. The Steady-State Solutions for an N-pool System using a Power-Law Approximation." *Journal of theoretical Biology* 25(3): 370-379.

Savageau, Michael A. 1970. "Biochemical Systems Analysis: III. Dynamic Solutions Using a Power-Law Approximation." *Journal of Theoretical Biology* 26(2): 215-226.

Schaffner, Kenneth F. 2007. "Theories, models, and equations in systems biology." *Systems Biology: Philosophical Foundations*: 145-162.

Voit, Eberhard O. 2000. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*: Cambridge Univ Pr.

Voit, Eberhard O. 2013. *A First Course in Systems Biology*. New York: Garland Science.

Voit, Eberhard O., Zhen Qi and Shinichi Kikuchi. 2012. "Mesoscopic Models of Neurotransmission as Intermediates between Disease Simulators and Tools for Discovering Design Principles." *Pharmacopsychiatry* 45(1): 22.

Westerhoff, Hans V. and Douglas B. Kell. 2007. "The Methodologies of Systems Biology". Systems Biology: Philosophical Foundations. Fred C. Boogerd, Frank J. Bruggeman, Jan-Hendrik S. Hofmeyr and Hans V. Westerhoff. Amsterdam: Elsevier: 23-70.