5 Managing Complexity: Modeling Biological Systems Computationally

In the previous chapters we have examined modeling practices that isolate and selectively focus on specific entities and processes, separate from much of their contexts in biological systems, in order to develop understanding and control of specific behaviors. Research in the burgeoning field of computational—or integrated—systems biology (ISB) aims to get a grip on how the higher-level functionality of complex biological systems emerges from a multitude of interactions among the elements of a system. The modeling practices in this field attempt to use as much information about the biological system as the modeler can find, while keeping the computational model computationally and cognitively tractable. Although ISB is a diverse field, the modeling practices in labs we have been studying are representative of a major area that draws on the resources of engineering fields that model human-made complex dynamical systems, such as electrical engineering, control engineering, systems engineering, and telecommunications engineering, as well as on mathematical and algorithmic resources from the computational sciences to model complex biological systems.[1]

ISB researchers investigate systems that comprise a range of biological phenomena that extends from intracellular interactions to those within organs or ecosystems. There are many objectives of the field, but in general, and especially in the bioengineering stream, the overarching objectives are (1) to build large-scale models that draw out the dynamics of biological networks and enable prediction and control with respect to phenomena of interest and (2) to use models to investigate what they call "the design principles"—or organizational principles that characterize the subcomponents of the biological systems. Understanding these principles, it is hoped, will provide the basis for a general mathematical theory of biological systems, as well as aid efforts by researchers and clinicians to control and

intervene on systems. Much of the research in the field is directed toward interventions in health and the environment, such as to design new classes of antibiotics, create personalized cancer therapies, produce biofuels, or develop protective strategies for ecosystems.

ISB researchers position themselves in contrast to traditional biological fields, especially molecular biology. Although biological experimentation can reveal local causal interactions among molecular elements, biochemical functions are coordinated and controlled through large-scale networks, which are networks that have wide boundaries and that involve many interacting elements. These networks tend to function by means of nonlinear interactions (for instance, feedback loops) such that the causal properties of an element of the network depend on interactions happening upstream and downstream in the network. Further, these complex networks generate robustness and redundancy, and have nonlinear sensitivity to changes in their parameters, which give rise to variability across individual cells and organisms. All of these features make biological systems difficult to understand and control, and explain, in part, why systemic diseases such as cancer or cystic fibrosis have proven so difficult to treat (see, e.g., Hood et al. 2004). Only quantitative simulation models of such networks have the potential to capture network intricacies at the scale and size required to identify variables and predict network behavior in response to perturbations with accuracy sufficient to determine how to intervene on them effectively. As the lab G director stated, systems-level modeling "*allows us to merge diverse data and contextual pieces of information into quantitative conceptual structures; analyze these structures with the rigor of mathematics; yield novel insight into biological systems; and suggest new means of manipulation and optimization*."

Although the desideratum and philosophy of a systems-level understanding in biology has a long history (see, e.g., O'Malley and Dupré 2005; Trewavas 2006), many researchers, including the directors of the labs we investigated, look to the Human Genome Project as the origin of the contemporary field. As the lab G director stated in our initial interview, "*So if you were to put a point there, it was the Human Genome Project . . . and at the same time you had the microarray and all that stuff started to come out. They [bioscientists] said 'Wow! You can do then thousand data points in one pop. Who wants to look at all that data with the naked eye? That's not possible to do, so we need computers'—whatever that meant.*" The confluence, around the turn of the twenty-first century, of engineering developments for biological

experimentation, especially high-throughput technologies that produce reams of data from one experiment; the widespread availability of powerful computing (including, but not just high-performance computing); the development of sophisticated mathematical and algorithmic methods for solving equations computationally; and the development of Internet browsers and search engines that enable rapid searching of scientific literature and databases all have contributed to making computational modeling and simulation of complex biological systems possible (see, e.g., Kitano 2002; Krohs and Callebaut 2007; O'Malley and Dupré 2005).

The labs we studied prefer to use the descriptor "integrated," rather than "computational," to emphasize the integrative effect of putting all the pieces together in a computational structure—"*like an integrated circuit*." At the conceptual level, "integration," as one researcher noted, means "*the tasks on this new frontier require thinking beyond linear chains of causes and effects—[rather]thinking in terms of integrated functional entities, thinking in systems, networks, and models*." This kind of thinking is about the dynamic behavior of complex biological systems and requires computational modeling and simulation to carry it out. Based on our research, we would also add that ISB is integrative in another sense: it incorporates and adapts engineering concepts and methods, for instance from systems theory and control theory, computational algorithms and methods from computer science and applied mathematics, and experimental techniques, concepts, and data from experimental biology. As we have discovered, "integration" in this sense is not smooth, since the concepts and methods drawn from different fields carry with them, among other issues, conflicting epistemic values and norms, as I discuss further in chapter 7.

ISB is a heterogeneous field that brings together researchers in biosciences, computational sciences (including applied mathematics), and engineering sciences in various configurations. Although some researchers have developed into hybrids over the course of their careers, the field of ISB does not aim at the kind of hybridization through education that I discussed with respect to BME. Instead, solutions to the problems the field poses create an essential *epistemic interdependence* among the participating fields. Although there are some ongoing attempts to develop hybrid modeler-experimentalists, the nature of the problems ISB addresses, arguably, requires both specialization and collaboration. The norm in the current state of the field (and some would say, in principle) is for modelers to

be trained in engineering or applied mathematics and for experimentalists to be trained primarily in molecular biology or biochemistry. However, to function most efficiently demands a symbiotic relationship. But, with little knowledge of one another's methods, concepts, technologies, and epistemic norms and values, at the present time symbiosis is more a desideratum than a reality. Our research has focused on the modelers, who by and large are driving the field, although we did conduct interviews with their experimental bioscience collaborators, when that was feasible (most were located at distant universities or in industry), and those provided important insights into collaboration issues from their perspective.

Researchers in ISB, as well as philosophers analyzing the field, have identified two broad strands of modeling in systems biology, namely *top-down* and *bottom-up* (see, e.g., Bruggeman and Westerhoff 2007; Krohs and Callebaut 2007). The top-down strand relies on high-throughput technology that generates large quantities of time-series data (dynamic data, as opposed to steady-state) for many elements of a system, such as chemical concentrations within cells. Computational methods, especially machine learning algorithms, are then used to attempt to "reverse-engineer" the system structure through making correlations among those elements. The bottom-up strand, on the other hand, aims to "reproduce" ("simulate") the behaviors of systems with dynamic computational models built using PCs. To build a model is an intensive process that draws on what can be pieced together of the network structure of the system and such features as kinetic and physicochemical properties of its components. The initial data for building the model usually come from collaborators, especially in molecular biology. These initial data often lack much of the information modelers need, such as on the concentrations of metabolites, and are often not a time series, which requires the modeler to interpolate data she can extract from the wider literature (including in databases). Building a model in these circumstances usually also requires modelers to use, and sometimes develop, sophisticated algorithmic techniques to estimate parameters (numbers) that provide the best "fit" of the model to the real-world data.

The labs we have investigated are both situated closer to the bottom-up strand, although both build models that they consider mid-scale or "mesoscopic." Simply put, these models contain modest details of system composition and organization, in that they simplify both the target mechanisms used to build the model and the underlying system functions they seek to

represent (see, e.g., Voit et al. 2012). Such models can be informative in themselves, but they also provide the basis for incrementally and iteratively building out the system representation, in ways that can enrich both the lower (mechanistic) level and the higher (systems) level. Both labs work in the area of biochemical systems biology. Research in this area is directed toward representing, understanding, and controlling intracellular metabolic and signaling pathways. Both labs aim to build models of these kinds of pathways individually, as well as those that integrate these pathways. Lab G gets its modeling problems from collaborators, and so works on building models of pathways of a wide variety of phenomena, including, during our investigation, dopamine regulation in Parkinson's disease, biofuel production from plants, yeast response to heat shock, and arteriosclerosis. Such modeling problems also provide material for the lab's own agenda of developing novel algorithms for parameter estimation. Lab C's modeling focuses solely on pathways in complementary processes of reduction and oxidation (redox), which are thought to produce inflammation, including immunosenescence, cancers, and arteriosclerosis. These labs have quite distinct methodological practices, but they both share the feature that the researchers come predominantly from engineering backgrounds. It is an important statement about the nature of the field that, while claiming to do systems biology, the researchers did not refer to themselves as systems biologists, but rather identified themselves and their biological collaborators functionally, as "modelers" and "experimentalists" (alternatively, "experimenters"). This contrasts with BME, where we found, despite differences in subfields (tissue engineering, neural engineering), researchers identified as biomedical engineers.

We cast the differences in methodological practices between the two labs as different accommodations to numerous constraints we have identified these researchers to be operating under. These constraints are so challenging, that the reader might wonder how in the world researchers in this domain can accomplish anything. As we will see, modelers in this field develop such effective strategies to manage the complexity of building models of biological systems that they routinely produce novel and valuable insights into the behaviors of these systems and into how to manipulate, control, or modify them productively, and—as the specific case of G10 (section 5.2) demonstrates—sometimes make quite spectacular biological discoveries. "Managing complexity" is a major theme we associated with

the codes we developed with respect to the methodological practices in each lab. As we will see, epistemic aims and cognitive needs intersect to shape problem-solving practices around managing the complexity not only of the biological systems, but also of the model-building process. I begin this chapter by laying out the constraints (section 5.1), then focus on how lab G modeling practices accommodate these constraints, in general and in a specific case (section 5.2), and then examine the epistemic and cognitive affordances of the methods (section 5.3) as they enable researchers to gain epistemic access and achieve their aims of getting a grip on complex biological systems.

5.1 Adaptive Problem-Solving in ISB

A major feature of problem-solving in the labs we investigated is that the research lacks the reasonably well-structured task environments that characterize established sciences such as molecular biology and bioinformatics. Nearly every step in the processes of model-building requires the judgement of the researcher to determine how to proceed, including how to (re)structure the problem, what modeling method to use, how and what portions of the biological pathway network to construct, what literature to rely on, what programming software to use, how to determine reliable parameters, and so forth. There is little available in the way of routines or protocols. Ultimately, what is produced in the form of a computational simulation model is a *strategic adaptation* to the constraints that model-building in ISB, in general and in the specific case, operates under in its present form. We have determined many of these constraints from our lab G and lab C investigations, but our claim that these are in effect across the wider field stems from widespread discussions about similar issues in the systems biology literature, including on education; by responses to our analyses by ISB researchers in audiences we have addressed and to our publications; and from findings in other ethnographic research I have conducted in ISB beyond this study.

5.1.1 Overarching Constraints on Model-Building

Modelers in ISB rarely can simply apply a formalism or preestablished principles to build a model that accounts accurately for a biological phenomenon. In effect, they face a multidimensional problem-solving task. Any model is the result of numerous choices about what to model, and how,

with whatever resources are available. Some of the constraints on model-building we have observed operating in the labs are as follows:

1. The biological problem: Biological systems possess features that produce nonlinear behaviors, with many elements playing multiple roles. For instance, cells contain networks of genes, proteins, and metabolites that interact in feed-forward and feed-backward loops and create myriad biochemical interactions. A modeler must restrict the considerable complexity of the biological system so as to formulate a tractable problem to model, while at the same time representing it in sufficient detail for the model to simulate the target behaviors and yield predictions.
2. Knowledge constraints: Modelers usually have no familiarity with the biological system prior to starting on the problem. Today they might have to model lignin production in plants, and next, drug resistance in a cancer. They know little about biological entities and experimental methods in general, which limits their understanding of what is biologically plausible and what reliable extrapolations can be made from the available data sets. By and large, there is no reservoir of theoretical models and laws of the biological phenomena to provide the structure and dynamics from which to articulate a model, such as there is in physics-based modeling.
3. Infrastructure constraints: Comprehensive databases of experimental information for most biological systems, while growing in number, are still limited. There is little in the way of standardized modeling software, or of generally accepted routines and formalisms to apply in building a model. There are few textbooks and little in the way of educational infrastructure directed toward computational systems biology, although several initiatives are under way.
4. Data constraints: The kind of experimental data (time series) needed for building dynamic models and parameter fitting is often not available or difficult to obtain, and the available data are usually noisy. Model-building is data-intensive and routinely relies on data beyond what are collected by bioscience collaborators in small-scale experiments, leaving modelers to forage for pertinent data in the literature and databases on their own.
5. Cost constraints: New experimental data are quite costly to obtain. Experimentalists often do not see the cost-benefit of producing the specific data modelers need. On the computational side, it can be costly in time and money to update old software.

6. Computational constraints: Most biosystems modeling is carried out on PCs, not with high-performance computing resources. Although significant improvements in their speed and efficiency have facilitated the rise of simulation modeling, computational constraints still figure into the level of complexity a model can have to keep such processes as simulation and parameter fixing manageable.
7. Time-scale constraints: Processes of generating experimental data and of model construction, simulation, and testing operate on vastly different time scales. Modelers can wait for months for data to build or test a model.
8. Collaboration constraints: The significant differences in epistemic practices and educational backgrounds of experimentalists and modelers limit their ability to communicate effectively and to understand and fulfill one another's epistemic needs. Thus, it is difficult for modelers to obtain the kind of data or expert advice they need from their collaborators.
9. Cognitive constraints: Modelers need to be able to track many relations at the same time and, especially, monitor indirect influences in the system. The need to keep multiple constraints and other factors in mind as one builds a model is a multidimensional problem. In general, human cognitive constraints, such as on memory and mental modeling and simulation capacities, limit the ability of modelers to manipulate and reason about the models, and therefore limit the scale of the models they can manage.

The labs we have studied have adopted different methodological approaches to deal with these constraints. Many of the constraints on this list indicate a problem situation in which there are limited data for building a model. We began our research with lab G, and it was immediately notable how often modelers started off discussing their work with complaints about how hard it is to find sufficient data of the right kind to build their models. These complaints were frequently expressed, with considerable emotion, as the model "*needing*" data, which led us to code such expressions as a concern with "feeding the model." We frequently heard the expressions "*parameter estimation*" and "*parameter fitting*," with modelers expressing considerable worries about finding parameters (due to insufficient, inadequate, noisy data). Parameters, roughly, are the constants in the equations and are needed to control the behavior of the model, such as the rate constants of an enzyme reaction.

Model-building in ISB is not guided by theory the way the physics-based modeling that philosophers have usually studied is guided. "Theory" is, of course, a multifarious and contested notion. In positioning ISB model-building with respect to accounts of physics-based modeling that have become standard in philosophy, we take "theory" to mean a reservoir of laws, canonical theoretical models, principles of representation (such as boundary conditions), and ontological posits about the composition of the phenomena under investigation that guide, constrain, and resource building models in diverse disciplines across a wide spectrum of physical systems. Model-building in ISB starts without such a reservoir. As the lab G director noted, in the absence of the kind of theory available in physics, a "*big problem is where do we get functions from*?" Instead, modelers have to make what they call "*educated guesses*" as to the functions, guided by mathematical notions, such as growth functions, and by principles developed in molecular biology such as Michaelis-Menten enzyme kinetics (a model of the rate at which enzymes catalyze in a specific reaction), often referred to by biologists as "partial theory." There are no correlates, for example, to Navier-Stokes equations, which describe the movements of gasses and liquids, used by climate modelers. Such equations also help modelers determine significant parameters, in this case, temperature and wind speed. In physics-based modeling, theory is a resource that can inform the modeler how to go from a data set to a good representation. The models of lab G often have large numbers of unspecified or "open" parameters. Without experimental data of sufficient or good-enough quality, researchers have to rely on mathematical and computational ways to determine parameters so as to fit a model, such that it simulates the system behavior with sufficient reliability to make predictions. For this reason, a major methodological enterprise in lab G is to develop new algorithms to advance what researchers call "*the art*" of parameter estimation.

Lab C's methodological approach is to have modelers also conduct wet-lab experiments to supply data for their models—what we have called the *bimodal strategy*. This is the director's adaptation to data limitations. We rarely heard modelers in this lab talk about parameter estimation problems, since the models they built were smaller in scale, and they would conduct biological experiments to determine many parameters as they were building the model. Thus, lab C models tended to be closer to the data, and open

parameters in need of estimation were few, though the larger-scale models they built had the fitting problems encountered in lab G. Lab C modelers did experience challenges around the need to master and coordinate model-building and wet-lab experimentation on their system in the course of developing a model, as we will see in chapter 6. The lab G director's methodological choice to collaborate with experimentalists rather than produce their own data is the predominant choice in ISB at present. The differences in methodological approach with lab C mark what the lab G director calls "*a philosophical divide*" in the field, which I discuss in chapter 6.

Both labs practice forms of what we called "adaptive problem-solving." All problem-solving is adaptive to some extent, but what is remarkable about the practices we witnessed in these ISB labs is the extent to which routine problem-solving depends on the researchers' ability to think innovatively while managing a range of constraints that create a significant cognitive load. Researchers in both labs specialize in building ordinary differential equation (ODE) models of gene regulatory, cellular metabolic, and cell signaling networks. Their efforts to integrate metabolic and signaling networks are novel (at least when we began our investigations). The variables in the ODE equations represent concentrations of individual metabolites in the network in a cell. Systems of equations are used to build dynamic models that can be run to simulate the changes to the concentrations of metabolites in a cellular network over time, where each metabolite pool interacts with specific other metabolites, represented as its neighbors in the network. Running the computational model under various conditions ("simulation experiment") shows how dynamic patterns emerge through the interaction of the pathway components over time. In general, the modelers aim to produce models that, when run, make reliable predictions of the dynamic relationships among specific variables in the model and perform robustly with respect to variations in parameter and initial conditions.

All aspects of the process of building a model are open to decision or modification, including the scope of the problem, how to represent the biochemical reactions, what data sets to use, what pathway elements to include, and how to estimate and fit parameters. As I noted previously, every model is a strategic adaptation to the constraints the modeler is working under and the resources she has at hand. The main, interrelated kinds of adaptations made by the modelers in the labs we studied have to do with the scale of the models they chose to build and with how to adapt problems

to make them tractable, both of which are situated in the context of determining what kinds of conceptual and methodological adaptations to make to apply engineering and mathematical resources to biological problems.

5.1.2 Mesoscopic Modeling

As I noted at the beginning of this chapter, the overarching aspiration of the field of ISB is to build large-scale high-fidelity models of biological systems that should, in principle, facilitate understanding of the design or organizing principles of systems or predict the consequences of manipulating the systems towards desired outcomes, such as to produce biofuels efficiently or to design personalized medical treatments. The current state of the field, though, as Eberhard Voit et al. observed, is that "the vast majority [of ISB models] are neither small enough to permit elegant mathematical analyses of organizing principles not large enough to approach the reality of cells and disease processes with high fidelity. Instead, most models contain between a handful and a few dozen variables, which firmly positions them in a grey zone far outside both declared goals of systems biology" (Voit et al. 2012, 23). They call such models "mesoscopic." We agree that to attain specific goals, a mesoscopic model might be the most informative, and therefore desirable, choice in itself (see, e.g., Batterman and Green 2020; Bertolaso 2011; Bertolaso et al. 2014). However, we consider the prevalence of this kind of modeling, which falls short of the epistemic aims of the field, to be a largely pragmatic and rational response to the constraints of managing the complexity of modeling these systems.

A mesoscopic model provides a "coarse structure that allows us to investigate high-level functioning of the system at one hand—and to test to what degree we understand, at least in broad strokes, how key components of a biological system interact to generate responses" (Voit et al. 2012, 23). Such broad understanding can enable modelers to make substantive predictions—sometimes of major significance—but also, importantly, provide insight into how to expand the model in both directions ("middle-out strategy") to provide a more comprehensive representation (Noble 2006; Voit et al. 2012). The initial model creates an affordance in the problem-solving environment that modelers can use to guide and structure their investigation in stepwise fashion. Understood in this way, mesoscopic modeling is a strategy for gaining epistemic access to complex biological systems by building out the system representation so as to be able to enrich

both the lower (mechanistic) and higher (system) levels. This expansion can be carried out by the builder(s) of the initial mesoscopic model(s) or by others in the field leveraging their insights.

Interestingly, Voit et al. advance a cognitive argument for the mesoscopic strategy, which they liken to hierarchical learning in human development: "This strategy of locally increasing granularity has its (ultimately unknown) roots in semantic networks of learning and the way humans acquire complex knowledge. . . . Hierarchical learning is very effective, because we are able to start simple and add information as we are capable of grasping it" (Voit et al. 2012, 23). We have advanced an additional cognitive argument that the mesoscopic strategy is a bounded rational response to handling the complexity of the constraints under which a modeler works (Macleod and Nersessian 2020). Herbert Simon (1957) argued that when faced with complex decision-making problems, people do not seek optimally rational problem solutions, but rather settle on solutions that are good enough to make progress ("satisfice"). The mesoscopic strategy enables the modeler to make progress, while holding out the promise of producing larger-scale models as modelers gradually gain understanding and control. A further cognitive argument, developed in section 5.3.2, is that as part of a coupled inferential system, the complexity of these midsize models remains at a level at which the modeler can still develop insight and intuition about the model's behavior and therefore make inferences about how to proceed in the model-building process.

As part of the mesoscopic strategy, modelers usually find ways to adapt the problems they tackle to simplify or get better traction on the specific problem. One way to adapt the problem is to keep the network representations relatively small through careful selection of what networks to include in the model at the start. Rather than attempt to model an entire complex network, modelers tend to focus on what their experimental collaborators indicate as potentially significant subsets when selecting the experimental literature to consider. An experimental collaborator of lab G relayed an example of this kind of adaptation when he told us of the reaction of the lab director after he had come to him with a large network to model: "*I think he's been in the real world long enough doing this systems stuff—long enough that he knows to start small. . . . So, when I came to him, I had these proteomics systems. We've seen about 10% changes in all the systems of the CF [cystic fibrosis] cell vs non-CF cell. Now when you think about the number of systems that are*

in the cells, 10% changes in all of those systems . . . is a lot of information. So, he's like 'you are deluding yourself.' So, then we decided to start with glycolysis and the pentose phosphate pathway of the Krebs cycle . . . to narrow it down to energetic pathways that are very well modeled." In this example, instead of trying to build a model of a large, intractable network, the director—a highly experienced modeler—moved the model-building process toward the strategy of adapting the problem in the direction of using small models, already established, and building outward from those.

5.1.3 Engineering Transfer

One important kind of problem adaptation is to use strategies and heuristics from engineering to alter the dimensions of a problem.[2] For instance, modelers might situate a network under study within a broader network, on the basis that the broader network can reveal connections between parameters in a subnetwork in ways that have a significant effect on the behavior of the subnetwork. This strategy helps to elucidate confusing dynamics. Another quite common strategy is to use an engineering method called sensitivity analysis to isolate the elements of a network that play the most significant role in the dynamics of the network. Sensitivity analysis basically targets the uncertainty in a model by examining the change in output produced by the change of specific parameters. This method can be used to simplify the network representation or to identify parameters that do not have too great an effect on the dynamics. These parameters, then, can just be assigned arbitrary values to reduce the parameter-fixing problem. Additionally, modelers often black-box component systems or component interactions to reduce the complexity of the network and, conversely, de-black box them if it appears that a subsystem is having a nonlinear effect on the network. Further, if the modeler cannot see a means of directly building a good model for a specific network, she will work on an alternative network for related phenomena that is simpler or for which better data are available. The modelers we studied often switched systems in this way or switched cell types for the sake of better data, with the hope they would be able to modify that model in the direction of the original problem. In general, to adapt problems, modelers employ strategies that incorporate and integrate engineering methods into systems biology. These methods, themselves, have to be adapted for the new subject matter and research environment, along with the engineering epistemic values that favor precision.

Throughout the course of model-building, modelers import concepts and methods that have been used in engineering for building models of human-made systems to transform biological problems into a form appropriate for mathematical and computational analysis. Such transformation strategies range from adapting the individual problem to designing methods for classes of problems. ISB modelers, in general, draw concepts and methods primarily from engineering fields, and especially control engineering, which has developed techniques for measuring and deciphering electronic signaling networks. The lab G director claimed that modelers can tackle a range of biological problems about which they have no prior knowledge because their training in engineering methods and concepts gives them "*the right mind set*"; that is, "*the flexibility to recognize shared features of control/regulation across disparate domains*." Many of the methods used in the labs are borrowed from these fields, including, but not limited to, simulated annealing methods of parameter-fixing (approximating a global optimum of a function), and nonlinear network analysis techniques. They also used standard computational modeling tools that are used more widely than in engineering, such as Monte Carlo methods of parameter estimation (an approximation technique using random samples of numbers).

In borrowing from engineering, ISB modelers are following a practice that pre-dates modern computational systems biology. Biologists have a long history of borrowing concepts such as circuit, system control, modularity, redundancy, noise, and sensitivity to conceptualize system-level phenomena (see, e.g., Wimsatt 2007). For instance, metabolic control analysis, which began in the 1960s, is based on engineering analysis of network control, which derives from sensitivity analysis in engineering (Westerhoff et al. 2009a; Westerhoff et al. 2009b). Modelers in the labs we investigated continue to extend these practices by experimenting with their own adaptations from their different engineering backgrounds. For example, one lab G researcher we followed, who had a background in telecommunications engineering, was trying to figure out whether, and if so, how she could adapt wave-smoothing techniques from signal processing to smooth noisy biological data. We found the degree to which both lab directors allowed graduate student and postdoctoral modelers the flexibility to choose how to go about trying to solve their problems—what background methods and concepts to rely on—to be quite remarkable. A successful adaptation can require considerable ingenuity. Failed attempts along the way are par for

the course, but these are seen to provide invaluable insights into the problems, as well as information on what to try next. Importantly, as we will see, modelers rely on the model-building process, with its ongoing simulations, to develop an understanding of their systems and figure out how to adapt them to their specific epistemic goals.

It needs to be noted, though, that although engineering methods and techniques can facilitate the model-building process, some biological understanding is required to help discriminate good moves from bad ones. Modelers talked all the time about the need to get a sense of "*what is reasonable and what is not reasonable*" biologically. This is a problem for which collaboration constraints are strongly felt. As the lab G director noted, "*really good biologists have a feel for things. . . . They know what to look into, how difficult it's going to be. . . . This intuition . . . is very hard to mimic or acquire*." In our investigation, we found that graduate student and postdoctoral modelers relied heavily on the lab directors, both of whom had considerable breadth and depth of biological understanding, to help them determine whether their moves were reasonable. This was so even for modelers in lab C, who conducted wet-lab experiments on the biological systems they were modeling. Sometimes researchers could ask experimental collaborators, but lab G modelers, especially, usually found it hard to get the attention of their collaborators. They frequently expressed to us a desire, along the lines as one modeler put it, to have "*a biologist in my desk drawer*," to back up their judgment on the moves they were making. She made that comment as I sat beside her to watch how she determined the steps she took to forage for data online. On the whole, the process of model-building is mainly the responsibility of an individual modeler rather than a well-coordinated process between modeler and collaborators. This is largely because the computational model is a "black box" to most experimentalists, and modelers, especially those who have done no experimental work, do not know how to convey model details or what they require for the building process to their collaborators.[3] We made the collaboration problem the focus of the educational experiences we promoted for ISB, as discussed in chapter 7.

The lab G researchers were fortunate to have, in addition to the director, a long-time experimental collaborator who often visited the lab as he transitioned to being a modeler, with whom they frequently consulted. He found it amusing that just because he was an expert in yeasts, the modelers thought he could answer their questions about any area of biology. He also

found it problematic that often they "*don't know the right questions to ask*." However, as we will see, the model-building process, itself, although not a substitute for a biologist's intuition, also helps the modeler build some biological intuition. Modelers claimed to develop this intuition through the extensive searching in and reading of the biological literature required to build out the pathway network and from their examination of the biological system's behaviors through simulations under various conditions (including counterfactual). Such intuition is particularly important when it comes to fitting the model (a process described below). As one modeler claimed, over time "*you get a feel for what might work and what probably doesn't*" with the system under study, and eventually more broadly. The graduate student modelers usually work on specific systems for four or five years—the director much longer.

As we will see, accounting for problem-solving and discovery processes in these labs requires analysis of a D-cog system comprising modeler, experimental collaborators, lab mates, and various artifacts, including computational models, pathway representations, diagrams, graphs, pen and paper representations, and data sources (publications, databases, search engines). I turn now to the model-building practices of lab G, first providing a general overview and then examining some of the details of one of the long-term modeling projects we followed.

5.2 "Where Numbers Come to Life": Getting a Grip on Systems Computationally

As mentioned earlier, lab G's practice is to obtain their modeling problems from experimental collaborators and, hopefully, obtain experimental data for building and testing the model from them. The director is a senior pioneer in the field of ISB and he portrayed the situation when we arrived as "*Biologists and clinicians come to my office and say, 'we have some data, so you want to work with us?'*" He also pointed out that this was a drastic change from when he started out: "*Twenty years ago that would have been utterly, totally impossible*." When he was a student in the 1970s, he found it a "*nightmare*" to figure out how to combine his interests in biology and math, because the combination "*was not only not supported, but outright considered ridiculous by biologists, and even more so by mathematicians*." He managed to get a PhD in developmental biology ("*it was really theoretical*

biology, but there was no degree in that") by developing some rudimentary computational simulations of predator-prey relations and scar patterns on budding yeast cells. The work on yeast led to a postdoctoral position with an electrical engineer working on developing methods for how to model biological systems, who, himself, had managed to get a faculty position in a microbiology department. Together, they built mathematical models of yeast and developed tools of mathematical and computational analysis for the emerging field. The future lab G director's first faculty position was in an interdisciplinary unit of epidemiology in a medical school, where "*the chair . . . was a visionary guy. . . . He hired engineers, he hired some people doing signal processing and AI.*" In his lab there, among other projects, he continued to model yeasts with local and international collaborators. He also continued to develop methods to analyze biochemical systems and for parameter-fitting for nearly twenty years, before he moved to his current position to set up lab G approximately six years before we entered.

Lab G, as we encountered it, had modelers who were working on a range of biological systems, including metabolism in yeast, atherosclerosis, neurodegenerative diseases, and sustainable biofuels production, at the request of experimentalists external to the lab (and most, external to the university). They also worked on novel algorithms for parameter estimation and structure identification in biochemical systems modeling in general. The general practice of the lab is that everyone works on an individual problem in collaboration with the lab director. The lab director has numerous one-on-one meetings with every researcher and contributes actively throughout the model-building processes. Because the researchers work on problems from quite diverse areas of systems biology, the lab director said he felt that lab meetings would not be useful. At his own initiative, though, he did arrange several group meetings to introduce us to the research of the lab. Interestingly, the researchers all expressed a desire to continue to have lab meetings, since they found it useful to see in detail what modeling issues the others were struggling with, but the director did not continue. We did witness—and were told—that it was standard for researchers to discuss specific problems with one another as they arose, which usually proved fruitful despite project differences. As one member noted, "*When I have to discuss, I grab hold of somebody and we start working on the board. It's as simple as that.*" The lab space consists of open cubicles with desktop PCs and is often empty of people, since most work from home and come in when they have

a course, a meeting, or a need to find someone to discuss modeling problems with.

The cognitive-cultural artifacts of lab G comprise computational and mathematical resources that are essential to achieving its epistemic aims. In our initial interview, the director framed their epistemic aims in terms of the overarching aims of ISB: "*We want to put pathways together and we want to predict what they do and then see if they do what is predicted. To do this rationally correct, you need to understand the types of design principles . . . regulation, adaptation, whatever. . . . So, ultimately, we need to understand these types of design principles and operating principles. We want to understand them because a) we are academicians and b) because we want to muck around with these things and change them.*" In this statement, the director expressed both objectives of getting a grip on complex biological systems: to understand how and why they exist a specific way in nature (mechanisms and design principles) and to determine what possibilities there are to manipulate them in a desired direction. Achieving these goals will both further the development of biological theory and enable bioscience and medical collaborators to manipulate the systems, for instance to produce biofuels from plants, tailor drug treatments to cancer patients, or manage bacterial populations in lakes. In the current situation, though, the kinds of in silico simulation models (mesoscopic) it is possible to build are, usually, of a scale and complexity that can provide only limited understanding of the mechanisms underlying the behaviors of complex biological systems. However, the insights they do provide are often sufficient to make novel predictions and can lead to successful experimental manipulations by collaborators. Further, as the lab G director noted, even such limited computational models can provide insights into "*why you have this one design in nature*" by comparison with a model as a "*hypothetical alternative*," which allows the modeler to examine counterfactual designs that, in principle, could exist in nature.

In practice, we found that modelers in both labs had much more limited goals. They tended to focus on modeling a system (1) to discover robust mathematical relationships among specific input and output variables in order to manipulate them in the in vivo system and (2) to infer the potential role of a specific molecular or component process in a network and its interactions, and to use this information to predict the effects of manipulations of these on system dynamics. In lab G, these goals were usually connected with requests from experimental collaborators to generate

hypotheses about what might be missing or wrong in their data, or to discover new relationships in the data. Such information could help them better direct their experiments or manipulations.

In both ISB labs, developing skills at building biosystems simulation models makes one a part of the lab cognitively and socioculturally. Many features of the basic model-building process are similar across the labs, but, as we will see, because lab C works with a smaller number of reactions, modelers are often able to use off-the-shelf modeling tools, and, significantly, the modelers conduct wet-lab experiments to collect additional data to build and test their models, which reduces the parameter-fitting problem significantly.

5.2.1 "I Always Start from Zero": Overview of the Model-Building Process

To grasp the complexity of the problem of modeling complex biological systems, a picture is indeed worth a thousand words. Figure 5.1 is the picture the lab G director gave us to illustrate the biochemical systems modeler's challenge. The left figure is the metabolic pathway (network of elements and interactions) of sphingolipid yeast, a budding yeast such as used in brewing and baking, that he has worked on for years. The pathway diagram represents, spatially, sequences of molecular interactions in the cell. It depicts a chain of reactions that result in the performance of some biological function. The right figure illustrates the limited portion one could model, and the abstraction of the elements and interactions a tractable model could handle.

The first step in model-building is to develop a representation of the biological network that shows the main reactions among the targeted elements in a system, called the "pathway diagram," which provides the basis for building the model. Most experimentalists work with only a specific subsystem within a network—often a tiny fraction of the overall pathway. It falls to the modeler to build out the pathway relevant to the biological system in the detail required to model it. Figure 5.2 provides an example of a pathway a lab G modeler was working on. In general terms, the pathway diagram is a conceptual model that represents, spatially, sequences of molecular interactions (metabolic and signaling, for our modelers) in living cells. In essence, it maps out a chain of reactions that result in some biological function being performed. The diagram also captures positive and negative regulation effects, which specify the influence of metabolites on different reactions. For the modeler, the configuration of the pathway elements specifies

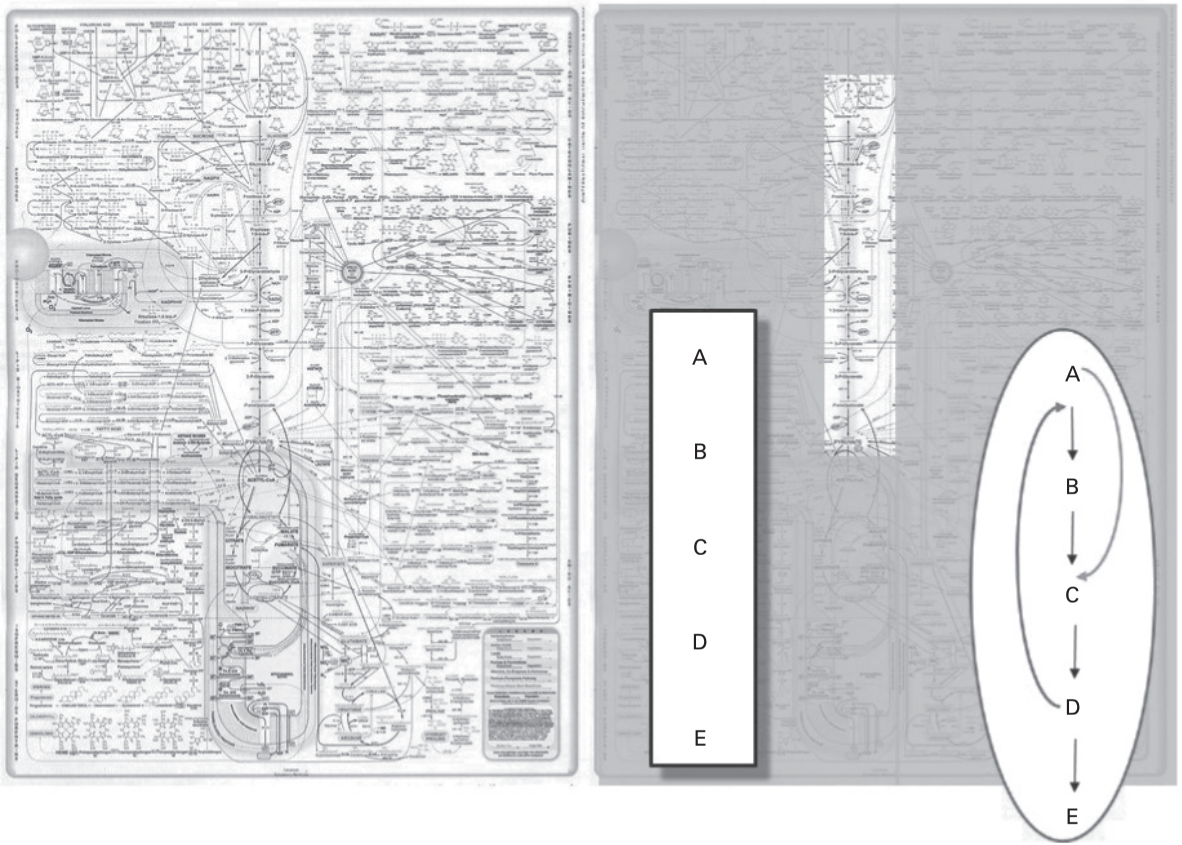


Figure 5.1
Pathway diagram of the sphingolipid yeast metabolic network. The left diagram is the pathway as currently understood and the right diagram illustrates the limited portion and relations that can be managed in a tractable model.

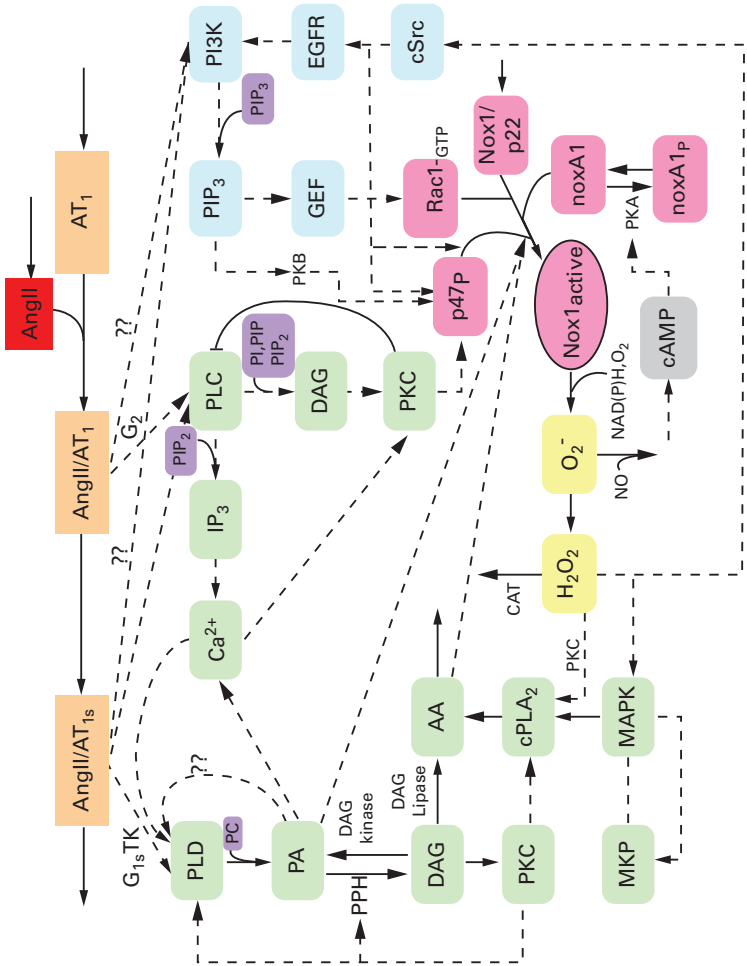


Figure 5.2
G12's preliminary pathway diagram for the angiotensin (AngII) hormone, which causes vasoconstriction, at an intermediate stage of development. Metabolite names are in the boxes. The dark lines indicate connections where material or information moves across nodes; the dotted lines indicate regulatory connections. The arrows indicate the direction of the processes. Note the question marks over some of the connections that are *"guesses"* by the modeler, which she will check with her collaborator to determine if they are *"reasonable"* biologically. The color coding indicates modules she modeled in different configurations.

the logical structure of the model. Among the affordances of the diagram format are the possibility to color code segments of the pathway, which the modeler might use to block out various combinations to examine in model-building, and the possibility to annotate it in various ways, such as to indicate uncertainties or to provide additional information on reactions.

The pathway diagram is often likened by systems biologists to "a roadmap where we wish to understand traffic patterns and their dynamics" (see, e.g., Kitano 2002 2). Lauren Ross (2018) explores this analogy in an illuminating analysis of the pathway concept from the perspective of biologists. Since developing biological pathways are a critical part of the modeling process, it is useful to elaborate on it here, before moving to lab G practices. As Ross notes, "when biologists use the pathway concept they often imply that some system can be understood in terms of causal routes or roadways. These causal routes capture interconnected paths that track the movement of some entity or informational signal through a system" (Ross 2018, 9). Her objective in that analysis is to explicate how the pathway represents the causal relational structure of biological entities and processes rather than the underlying mechanisms producing the biological phenomena. On her analysis, the pathway develops a fixed order of causal relationships that "capture the 'flow' of some entity or signal through the system. . . . Cell signaling pathways track the flow of a signal through molecular and cellular systems, metabolic pathways trace the flow of chemical substances through stepwise changes" (11). The language biologists use in discussing pathway representations—"flow," "flux," "connection," "blockage," and so forth—indicates "something that is carried over from one causal step to the next . . . something that travels along causal connections" (Ross 2018, 11) She also notes, importantly that, as with a road map, a significant amount of causal detail (for instance, temperature and Ph) is not represented in the pathway diagram, which makes it a more abstract representation than would be needed to represent causal mechanisms. This does not mean that the pathway diagram is devoid of mechanistic information, for instance, regulatory processes.

Ross does not discuss how pathway representations are used in computationally modeling biological systems, but her analysis accords with the features of those representations we have seen modelers in both labs emphasize when they build pathways to model the behavioral dynamics of a system. Modelers use the same descriptive language, but in addition, from

the modelers' perspective, they say "*the mathematics is in the arrows*" that represent the causal flow. The map itself is a static representation, but the modelers see the dynamics of the system in the arrows, which provide, as one modeler stated, the "*functional dependencies from which [they] derive . . . differential equations*." The model is built to capture the flow of interactions among the pathway components over time so that the model produced from it will act out the dynamics of the system-level behaviors of the target system that produced the data. In effect the pathway representation provides an analogue model that exemplifies the causal relational structure of the processes that produce system behavior, without specifying the underlying causal mechanisms that produce the behavior. From this representation, modelers can use the representational affordances of mathematics (such as power laws that represent relational changes) to create models that enact the causal dynamics. Understanding the dynamical behavior does not require knowing the underlying mechanism.[4]

What is remarkable about the process of developing this network map is that, usually, modelers receive only a small piece of the diagram from their collaborators, if anything at all, and it is their responsibility to build this network of reactions through foraging in the literature and available databases. The modelers we followed were engineers with little biological knowledge, and no knowledge of the specific biological systems when they started to model them. As one modeler noted, "*I always start from zero*." After sitting with this modeler to watch how she collected data and other information through Internet searches to build her model, I nicknamed this process "google biology." In publications modelers often insert numbers in parentheses at locations where specific references from the literature have been used to build out the pathway. Importantly, the primary means by which modelers begin to learn about the biological systems they are responsible for modeling are through literature searches, from which they not only gather data, but also develop conceptual understanding, and through building and simulating partial models in the course of building out the pathway network.

Our interviews with modelers and experimental collaborators show the pathway diagram to function as a boundary object (Star and Griesemer 1989) in that it is a representation used with sufficient flexibility in interpretation so as to provide a means of communication among the different communities. Importantly, the pathway diagram provides a tool to identify and track,

visually, causal movements within the network, for both communities, as Ross's and our analyses show. These movements provide a basis for reasoning about the network both in molecular biology (see, e.g., Sheredos et al. 2013) and in mathematical modeling. When we began our investigation, we had thought the computational model would perform the function of a boundary object, but soon realized bioscientists have little to no understanding of it. To them it is largely a black box. From what we witnessed, collaboration is rarely smooth, but one important interaction modelers have with experimentalists is to check whether a modification they have made to the pathway is "*reasonable*." This important check is possible because, as we saw above, for experimentalists the pathway diagram represents a set of qualitative causal relationships among molecular elements, and they usually can infer whether the proposed modeler modifications are plausible within the causal structure. Sometimes they are even aware of additional experimental literature that will aid the modeler in confirming their modifications.

For modelers, the pathway diagram represents a mathematical structure. As one lab G researcher expressed it, modelers "*in some sense translate it [the pathway] into a map we can deduce math from*." For the modeler, the nodes are variables and they put "*[mathematical] meaning into the arrows*," by giving them precise quantitative values for the rates of a reaction. This process necessarily involves much simplification and abstraction of what modelers often refer to as "*messy*," "*dirty*," and "*noisy*" biological systems so that they can be modeled quantitatively. The model is built on a generic pathway, such as the lignin pathway in plants, and is a generic dynamic representation in that it simulates the behavior of systems of that kind—of a generalized target.

It requires considerable effort and judgement on the part of the modeler to find the data in the literature and databases and evaluate which ones are relevant to and important for their problem. As part of their judgements, modelers need to determine what data sources are "*trustworthy*." That is, the modelers need to exercise judgment about the source, quality, and relevance of the data. The modelers we studied pointed out that they try to select data from labs that they consider to "*produce reliable data*," based on their lab's experience with them, especially those of the director. Similar judgements are made about databases, which are developed and curated in significant sociocultural negotiations (Leonelli 2016). There are many missing pieces (e.g., the pink portion of the lower right quadrant of figure 5.2 was built out entirely by the modeler) and many open questions (note the question

marks), which require the modelers to guess potential reactions. When possible, they present their guesses to experimental collaborators in the form of hypotheses to determine, as G12 explained, if the addition is "*reasonable*."

Notably, our investigations have shown that *building the pathway is an iterative and incremental process in which model simulation is itself a critical resource*. That is, the pathway structure is assembled in an exploration that involves preliminary simulations. Often small pieces of the pathway are simulated by the modeler, for instance by running through specific values for variables, using pen and paper (or marker and whiteboard) and their imaginations, before running segments in a computational simulation (see figure 5.3).

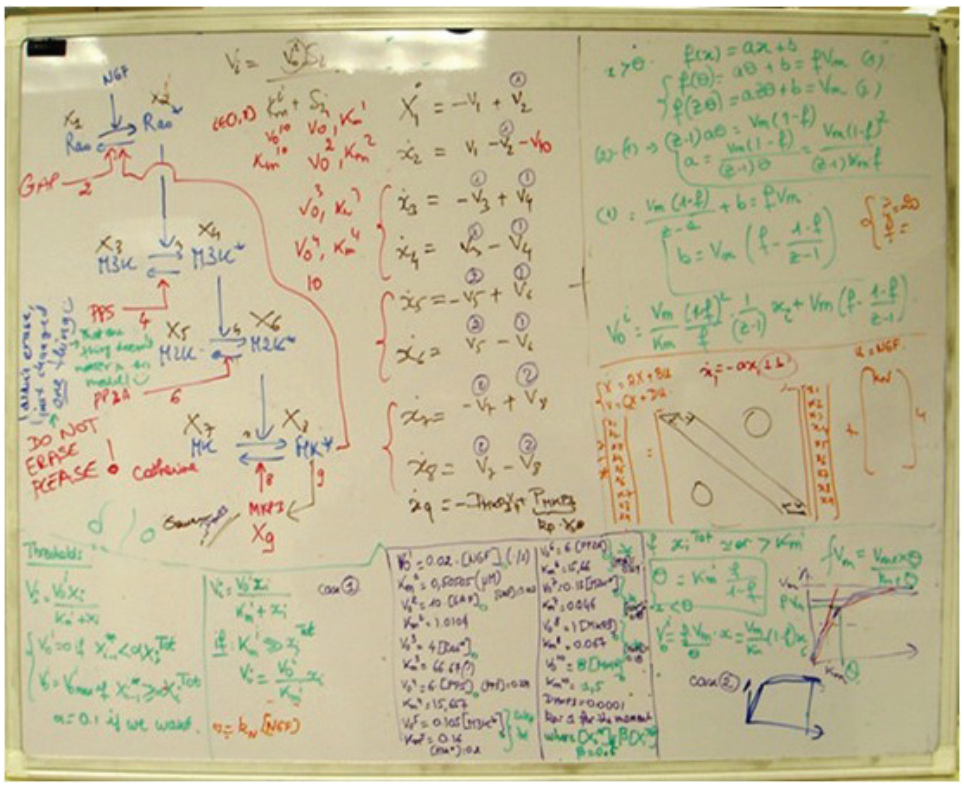


Figure 5.3
A lab G whiteboard on which a modeler is working out pieces of the pathway under investigation "by hand" and imagination.

Once the pathway is developed in sufficient detail, preliminary models usually are built in a modular fashion, with and without pieces of the pathway (such as the different color-coded sections of figure 5.2). Simulations of these help to determine, for instance, what can be trimmed or where there are possibly missing pieces. Through these simulations the modeler builds up an understanding of the dynamics and relevance of specific pathway elements. This enables her to make judgements about adding elements that might be playing a role but are not discussed in the literature on the system, or about feedback relationships that are not documented, or about what can be safely left out of the pathway. Such determinations are often based on elements in the literature that are thought to be related to the system, such as from different species and different cell lines. Modelers also use simulations to determine values of parameters often missing from the literature, such as the speed of the reaction (rate constant) and the sequence of reactions to the product (kinetic order), which experimentalists usually do not measure. Determining parameters from the literature is itself a complex process in which modelers have to reverse engineer the graphs they encounter into the numbers they need for their models. As one modeler explained, "*There might be graphs that have trends or what not, and then I have to quantify the graphs and then either figure out slopes or things of that nature to get at a particular number*."

The processes of building the pathway create a unique composite network of metabolites and parameter values. The pathway brings together pieces of information that are spread over a wide set of papers, databases, and unreported experimental data. The pathway diagram not only provides the basis from which the modeler builds the computational model but is itself a visual representation of a conceptual model of a network of causal interactions. The computational model built from it creates a synthesis that is, in effect, a *running literature review* that exists nowhere else. Thus, simulation is not used only to "sound out the consequences of a model" (Lenhard 2007, 181), but, notably, also *to learn and assemble the relevant ontological features of a system*. The process of adapting the pathway network continues throughout the model-building process until pathway, experimental data, and parameter fit coalesce into a model (or small set of models) that simulates the behavior of the target system (model output matches experimental data), at which point the model can be diagnosed and tested until it is considered validated. If the model fails testing, diagnosis is, as one

modeler explained, "*a big problem . . . because you don't know if the pathway structure may be wrong. Second, maybe the parameter is wrong. So, maybe the algorithm is wrong. So, I have to check every part of it to make sure of everything if something goes wrong. . . . It's actually a cycle, an iterative process—so we go back and forth.*"

For the modelers in our labs, model-building is a labor-intensive process that usually takes several years. Building the model requires the modeler to make numerous choices along the way. Modelers can choose a variety of formalisms to build the model. Choices include, for instance, whether to use phenomenological models, such as agent-based models, or mechanistic models, discrete or continuous models, spatial (partial differential equations, PDEs) or nonspatial (ordinary differential equations, ODEs) models, stochastic or deterministic models, and multiscale or uniscale models. Lab G most often chooses to represent the interactions in sets of coupled ODEs that capture how the concentration levels of different metabolites in the pathway change over time. The number of reactions investigated by the lab G modelers during our study ranged between fourteen and thirty-four—a number that the director characterized as "*just a handful*," when compared with those in the actual system (figure 5.1). The number of equations needed to capture these reactions varies with the specific questions the modeler is exploring, the nature and availability of data, and the computational resources. The advantage of ODE models is that they are both relatively simple conceptually and have the potential to be highly informative. In addition, there exists a wide range of computational and mathematical resources for analyzing system dynamics and for estimating parameters for ODE models.

Selecting an ODE framework opens another range of choices about whether, for instance, to model the system as steady state (static) or away from equilibrium, whether to use a mass-action stoichiometric model (based on rate of chemical reactions), or to use a canonical mathematical template such as biochemical systems theory (BST) that averages over the details of the interactions, or a mechanistic model that sticks closer to the molecular details in the form of rate laws of individual enzymatic reactions.[5] The choice depends on the nature of the problem, the goals of the modeler, and the nature of the available data. In the culture of lab G, BST plays a major role in ODE model-building. Even so, there are no set choices, and much depends also on the preferences of the modeler. As one modeler

told us in discussing the model-building process, for many choices, "*It's a pragmatic choice. That's why modeling is still an art—it's a choice people make. I make one choice and another one would make a different choice.*"

The modelers usually split the experimental data into two sets, one used to develop and fit the model (training data) and the other, to validate or test the fitted model (test data). The complexity of the tasks of fitting and testing a model is highly dependent on the nature and quality of the experimental data available. Rarely do lab G modelers have access to rich, dynamic data (time series). Most often, they have steady-state data that show how an experimental manipulation led to a change in metabolite level from a baseline. These data are reported by experimentalists usually as a single data point going up or down or holding steady ("steady-state" data), which provides the experimentalists with all the information they need, but not the modeler. This difference in needs again points to how differences in epistemic aims create problems in collaboration. A common lament we heard about experimental collaborators was expressed by one modeler as: "*They just care up/down. . . . They don't care time series . . . how this dynamically changed. They just care what is the result.*"[6] I used the word "lament" because this complaint was always expressed emotionally, with considerable frustration. In the absence of good dynamic data, the modeler faces considerable uncertainty because a range of parameter values can generate model results that fit sparse data, so the fit is not unique. The modeler can use algorithmic techniques and various computational tricks to figure out how the parameters might be changed and at least narrow down the range of acceptable fits. But it is often unclear whether the lack of a unique solution is because the parameter estimation is poor or whether some elements are missing in the pathway.

If the data generated by the model do not fit the test data, the modeler tweaks the parameters ("*tunes the model*") until the results provide a satisfactory fit. The modelers we studied do not use real-time dynamic visualizations of model behavior (as in lab D). Rather, they generate graphs that plot the concentration value of a molecule in the pathway across time for the model and for the experimental data, and compare a stack of graphs for different parameter values to judge how good the fit is. All of the modelers we interviewed pointed to parameter estimation as the most difficult part of the model-building process. In lab G, modelers often use optimization algorithms to estimate a significant number of open parameters. But just

as often they need to use novel reasoning about the problem to develop fitting options, such as to determine what reactions might be set to zero or what kinetic orders the free parameters might be. One technique we saw was to use data available on the same metabolic elements from other cell lines, such as using neural cell data to get parameters for a metabolite in smooth muscle cells. Modelers justify this move on the basis of their judgement that the systems in the diverse cells are reasonably homologous. Other common techniques modelers use include sensitivity analysis, which enables them to set parameters that do not affect network dynamics (insensitive) to a default value, or to explore the dynamics of different parameter values and ranges by running through random numbers with Monte Carlo simulations. In lab G, modelers sometimes create new algorithms for parameter estimation as part of the fitting process for the specific case, which, if useful, they will try to extend to other cases. All of these processes for parameter estimation and fit involve running numerous simulations (on the order of hundreds of thousands). Thus, simulation is not simply the end phase of problem-solving. *Simulation is a resource for iteratively building the simulation model itself*.

Once a satisfactory fit is achieved, the model is run through a series of diagnostic tests, including for stability (does not crash for a range of values), sensitivity (input is proportional to output), and consistency (reactant material is not lost or added). If these diagnostic tests fail, the modeler can tune the parameters again or modify the pathway. In general, modelers employ strategies that adapt and integrate engineering modeling methods into systems biology. These labor-intensive processes, as well as others I have not mentioned, continue until the model fits the available experimental data, as established by the data output of its simulation runs. The simulation model created by these means is generic in that it makes manifest the dynamics of all the available data on that type of system, including natural systems, in vitro systems, and engineered or modified systems. In an important sense, the "system" modeled is an abstract general system, and the dynamical behavior the model exemplifies is that of a system of that kind. As such, it enables the modeler to examine a range of behaviors, including counterfactual cases, which can provide insight into, and predictions about, how the pathway might be reengineered for specific purposes.

That the model produces a satisfactory "fit" does not mean it provides a point-by-point replication of the data, but, rather, that the behavior of the

model replicates trends (metabolite production going up or down) for most of the variables. The three main elements of the model—data fit, parameter values, and pathway structure—are mutually constraining, since they are tuned together in an incremental and iterative process until a model is considered validated. The pathway representation, for instance, is both tailored to fit the capacities of mathematical frameworks and shaped by parameter fitting in terms of available parameters and of the estimation tools used. All of these elements are kept in dialogue throughout the model-building process. Every version is "*just a version of your knowledge at the time—of what you think is going on. And it will keep changing as you learn more and more about the system.*" In the end, this modeler noted, "*the best your model can do—is a verifiable hypothesis about what you think is going on.*" That is, the objective of ISB modeling is to build a computational model of the target biological system that exemplifies its behavior under selected conditions, which in this case means that it replicates the existing experimental data and predicts new data that experimentalists can verify. At the completion of the model-building process, the goal is to have a robust model, stable for a wide range of parameter values, from which to derive novel behavioral predictions that have sufficient warrant to transfer as hypotheses to the target system, and, hopefully, will be tested experimentally by collaborators. As all of the modelers pointed out, making predictions—not just fitting the available data—is the only way to get past the underdetermination of a model.

In sum, modelers assemble the structure and local dynamics of the system being modeled largely from scratch by gathering empirical information from a variety of sources and piecing it together into an effective representation using a variety of assumptions, abstractions (modelers noted especially simplifications and approximations), and mathematical and computational techniques. Each modeler chooses the methods and strategies he or she thinks best to solve the problem without any formal procedure governing the selection process. Similar to the way a bird will gather whatever is available to build a stable nest, a modeler pulls together bits of biological data and understanding, principles developed in molecular biology, mathematical and computational theory, and engineering principles from a range of sources in order to create stable robust simulations of the behavior of a biological network ("bird-nesting process"). Modelers rely on the building process, especially their ongoing simulations, to come to understand their systems and adapt their representations of them to their

specific epistemic goals. Thus, modelers rely on the dynamical behavior of the model, itself, to make inferences about how to proceed in building both the pathway network and the model. This important role of simulation for the modeling-building process has not received sufficient attention in the literature on the epistemology of simulation.[7] A major benefit of ethnographic investigation is that it can uncover the hidden creative work modelers carry out with the choices they make in model-building, as well as the ongoing processes of developing epistemic warrant for the model and the model-building practices, which are unlikely to be included with the formal analysis presented in a publication.

Building computational simulation models in lab G requires a sophisticated grasp of mathematics, computational methods, and systems engineering analysis methods. It is notable that students begin their research with little to no prior experience in biosystems modeling. We followed three of the graduate students intensively, two (electrical engineering background) from near the start to the finish of their dissertation research (~four years) and one (telecommunications engineering) during the course of her first year, in which the lab director gave her projects to help out on, which was his usual training procedure. We also conducted numerous interviews with the other graduate students and the postdoctoral researchers, including about the algorithm development work, and with two experimental collaborators. We were able to grasp enough of their model-building practices to inform our research questions. Section 5.2.2 briefly outlines the model-building processes of one graduate student we were able to track from start to finish, to provide an exemplar of how researchers in this field achieve their epistemic aims. We did not anticipate that he would make a significant biological discovery. G10's model-building process is typical of the nature of the problems lab G modelers address and the strategies they use in handling problems. As noted earlier, there are a wide variety of modeling practices in ISB, but, in general, lab G practices are representative of practices in the field that use ODE models. It is a remarkable feature of their modeling practices that engineers with little knowledge of biology, and none of the system under study, are able to construct models that not only replicate the available data but also produce highly specific verifiable predictions about complex biological systems. The exemplar demonstrates the need to examine the *processes* of model-building, which to a large extent cannot be gleaned from published papers, in order to develop an account of

the epistemic and cognitive affordances of computational simulation (section 5.3).

5.2.2 A "Model-Based Signal Postulate": Finding a Remedy for Lignin "Recalcitrance"

G10 has an undergraduate degree in electrical engineering and a masters in bioengineering. For his MS degree he had worked on a bioinformatics modeling project for which he took a couple of biology courses (without labs), but he was not familiar with systems biology modeling when he arrived at lab G. He had read a biosystems modeling text by the lab director before deciding to apply. G10 had started on his dissertation project shortly before we entered the lab and finished in four years. In our initial interview, G10 stated that his "*engineering background contributes a lot to my way of thinking to solve a problem.*" He contrasted his engineering perspective, derived, in particular, from control theory, with that of a biochemist as evidenced in "*the biological journal literature*": "*[Engineers] look at things more at the systems level than the individual level, . . . Biochemistry look at the single protein or the single pathway—they don't really look at the whole system and how each pathway will interact with each other.*" He considered the systems perspective essential "*if you really want to understand how the human works or how the plant works.*"[8] G10's project started with a request from biofuels industry researchers for the lab to help them figure out how to tweak the lignin pathway in alfalfa to develop transgenic plants with lower lignin, so they could more easily extract sugars for the production of biofuels. Lignin is a natural polymer that hardens plant cell walls and enables the plant to grow upright. It is difficult to break down (it exhibits "*recalcitrance*") when biomass is processed into fermentable sugars using enzymes or microbes. The experimentalists had been developing genetically engineered plants with lower lignin content but were finding it difficult to determine a balance that would keep the plant structurally sound. Further, their transgenic species decreased only one of the three lignin monomer building blocks (called monolignols H, G, S). Although they had not collaborated with modelers before, they felt that modeling might be able to help them understand something about the mechanisms underlying lignin production, which would enable them to develop transgenic species with low lignin content and good growth. They also hoped, at the very least, that modeling would provide information that would enable them to develop plants with different ratios of lignin monomers, especially a lower

S/G ratio, which would improve the extraction of sugar from plant cellulose. This was a new modeling area for systems biology. Other modeling efforts in the biofuels domain were in the area of bioinformatics, or were models of organisms that are used to break up the plant mass. G10 expressed the hope that "*model-based insights will become the foundation for the rational design of metabolic engineering strategies*" for biofuel production.

G10 described his own "bird-nesting process," generally, as follows: "*We just search the literature and find the necessary data from it. . . . But most of the time you don't have much data. . . . I need to, you know, add other components from other theories, for example, the flux balance analysis . . . and I combine that with biochemical systems theory to build a model*." One unusual feature of this case is that G10 had a fairly well-established lignin pathway in the literature to start from and, in the end, lots of data for the fitting process, though he still had a considerable number of open parameters. To carry out parameter fitting for the specific lignin system he was working on, he needed to develop several novel modeling strategies for his analysis, one of which he also published separately as a potential community resource for handling systems of this kind.

In the beginning, G10's collaborators gave him few data, and what they did give him was of poor quality for modeling. As he noted, "*They don't measure the concentration, for example. And they have few kinetic data. . . . Most of the data they have is just output, the final output*." This created a significant problem because there was little literature on alfalfa, the plant they were working with. Complicating things further, they were unresponsive: "*Sometimes you want to ask question, and he would get back to you in a month—or even two months—or even don't reply. . . . That's a problem because we are not expert in the field. . . . They have more information than we know from the literature*." This is not an unusual "collaboration" situation for ISB modelers. Even when the bioscientists request the modeling, it often is low priority for them. Importantly, these bioscientists were unwilling to part with unpublished data, which constituted the bulk of their data. They seemed not to understand that the modelers would use it only for building the model and would not publish the data: "*Right now they just give us the data they have published. . . . They told me they need to publish it first—and then they can give me the data later*."

The collaborators projected it would be about six months before they would give G10 the additional data, so he decided to build a model of lignin biosynthesis in poplar—a related species for which there were ample

data in the literature, because it is the preferred biofuel species in Europe. His idea was to build the poplar model as a "*proof of concept*" for biosystems modeling in that domain, which would also help him understand the lignin pathway better. He assumed some of what he did would transfer to the alfalfa case. It turned out, unexpectedly, that to build the poplar model he needed to develop what he called "*a new two-step modeling approach*" to deal with the mathematical complexity and parameter estimation for the lignin pathway. He thought this novel method might then provide a template for modeling in the lignin domain. His approach was to integrate dynamics models with fluxes (the rate at which a metabolite is processed) derived from constraint-based models. The two steps were first to build a static, constraint-based model, which assumes the metabolic system is in steady state, and then use flux information derived from that to build the dynamic, kinetics-based model. The static model used the flux balance analysis (FBA) method, which assumes that the metabolic system is in a steady state in which, for each metabolite, the sum of fluxes coming into the pool equals the sum of fluxes coming out of the pool. The dynamic model made use of the BST modeling framework, where each differential equation in a model represents the time-dependent change in one metabolite as the sum of production fluxes minus the sum of degradation fluxes. For this model, G10 used the BST framework's generalized mass action (GMA) representations, which model the flux as a sum of the inputs minus the outputs.

The attractiveness of the BST framework in data-poor modeling is that it can account for a variety of dynamics by modeling the flux of dependent variables as a product of power law functions (relative change in one quantity gives rise to a proportional change in another), with each individual flux represented separately with one power law function. This means that even if the nature of the interactions among elements is not well known for the system, there is a good chance the model will capture the underlying causal regularities in the system, and so account for the system dynamics within the range of the realistic parameters. The parameters used are the rate constant, which determines the turnover rate of the process, and the kinetic order, which characterizes the influence of one variable on a given process.

The two-step process still left G10 with twenty-seven open parameters and required using optimization strategies to fit. To reduce the parameter space, G10 set all but the parameters considered significant (small change leads to large change in S/G) to what he considered "*biologically reasonable*"

values (determined to be so from reading the literature and discussion with the director). He then optimized the significant parameters by using various computational techniques. The results of this approach generated an ensemble of models (there was no unique model) with minimal error (SSE: sum of squared error) between model results and experimental data. Our modelers use "ensemble" to refer to a small group of models with different parameter settings that they settle on to cover uncertainties in the parameter values.[9] These models enabled G10 to identify key reactions that influence the S/G ratio in poplar, and he was able to make some predictions about how the pathway might be tweaked by knocking down specific enzymes to lower the S/G ratio.

The process of building the lignin model for poplar prepared him to deal with the more complex modeling problem presented by the alfalfa lignin system. The alfalfa model would contain twenty-four ODEs. In addition, the collaborator data included points in the growth of the plant over time (eight different internodes), where the lignin levels were different for each of these points. G10 used a slightly modified two-step procedure to analyze several internodes simultaneously, while interactively building the model and modifying the pathway in an incremental and iterative process.

Once G10's collaborators had published the relevant alfalfa research, they gave him the Excel files for all their data—which meant that, unlike the typical case, he had "*many data . . . for seven transgenic experiments and each experiment generate about seven sets of data. . . . They have more data than we need to know*." The collaborators did not give him any pathway structure, but again he was fortunate: "*The [generic lignin] pathway structure is from the literature—everybody is using it*." But, as he discovered, species-dependent data would be important in building out—and significantly altering—that lignin pathway, which had been established for twenty years. At the outset, his own literature search led him to add new elements to the pathway network, noted in red in figure 5.4.

The first model he built was for a wild-type system at steady-state, using the modified pathway (left diagram, figure 5.4). G10 discovered that this model could not produce accurate data when inputs were perturbed out of equilibrium, which suggested to him that some regulatory mechanisms controlling excess flux needed to be figured into the pathway. He tried out several pathway variations from studying the model structure with simulations, and then selected those that were the most "*biologically reasonable*"

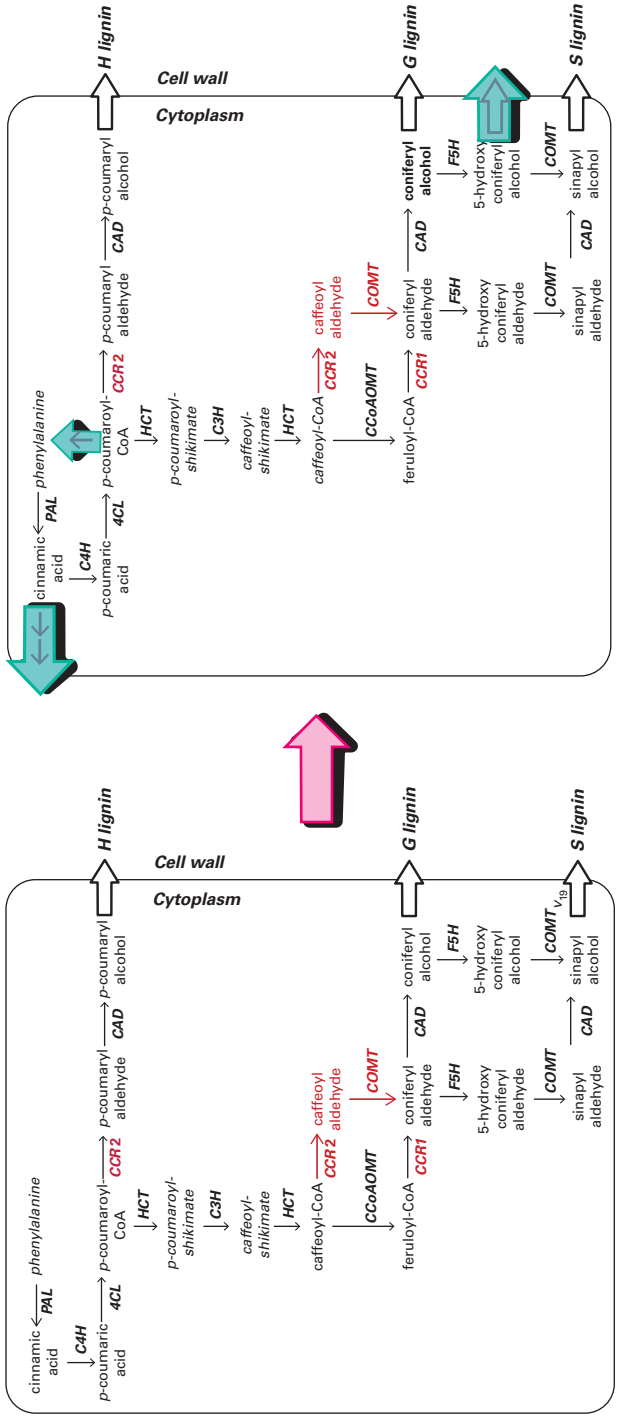


Figure 5.4
G10's initial modifications to lignin pathway. The left diagram shows additional reactions (in red) that lead to the H, G, and S monomers. In the right diagram, the highlighted blue arrows connect the extra flux to the environment at the points G10 hypothesized it leaves the system.

ways of removing flux ("*overflow fluxes*": highlighted blue arrows in the right diagram, figure 5.4). He translated these into precise mathematical modifications that would relieve the system. As he explained his process, "*We have data from our collaborators and we analyze it with very simple linear models, and based on our analysis results, we suggest there—this original pathway needs to be modified so that this data can be explained. . . . This is an important piece of knowledge that comes from the model*," that is, through the understanding of its system dynamics provided by the model. With the new pathway structure, G10 was able to build a dynamical model for each internode in each wild-type or transgenic plant and make hypotheses about the metabolic control of this pathway.

He used the data for each of the seven transgenic plants to build models on the biological assumption that the genetically modified strains would function as close to the wild-type as possible, within the limits imposed by the modification. Fitting the models was again a complex process, with numerous open parameters for each model, which he handled in a manner analogous to that of the poplar model. In the end, for the final modified alfalfa pathway, G10 arrived at a consistent convergence of five optimized models that tested well, and each gave similar predictions. He argued that the fact that this ensemble of models converged on similar mathematical relations for the target variables "*provides validation*" for the model. These models provided specific new causal information about which enzymes could potentially be targeted to decrease the S/G ratio, but did not provide an overall mechanistic explanation for the system behaviors. Altogether, G10 arrived at what he called seven "*model-based postulates*," which are mapped out on his final representation of the pathway (figure 5.5).

Two important postulates are, first, the reversibility of some reactions (straight arrows pointing upward in figure 5.5A and B) in the path where he had earlier removed excess flux. Second, he hypothesized the possibility of independent pathways ("channels") for synthesis of G (blue) and S (red) monolignols. Channeling can make a metabolic pathway more rapid and efficient, and the potential role of these channels in the lignin pathway was an important new hypothesis. He offered this second postulate as a solution to what he called a "*puzzle*": given the data he had on up-regulation and down-regulation of specific variables, the S to G ratio was considerably higher in transgenic plants than in the wild-type. But now he claimed the model enabled him to "*see what happens inside the pathway*,"

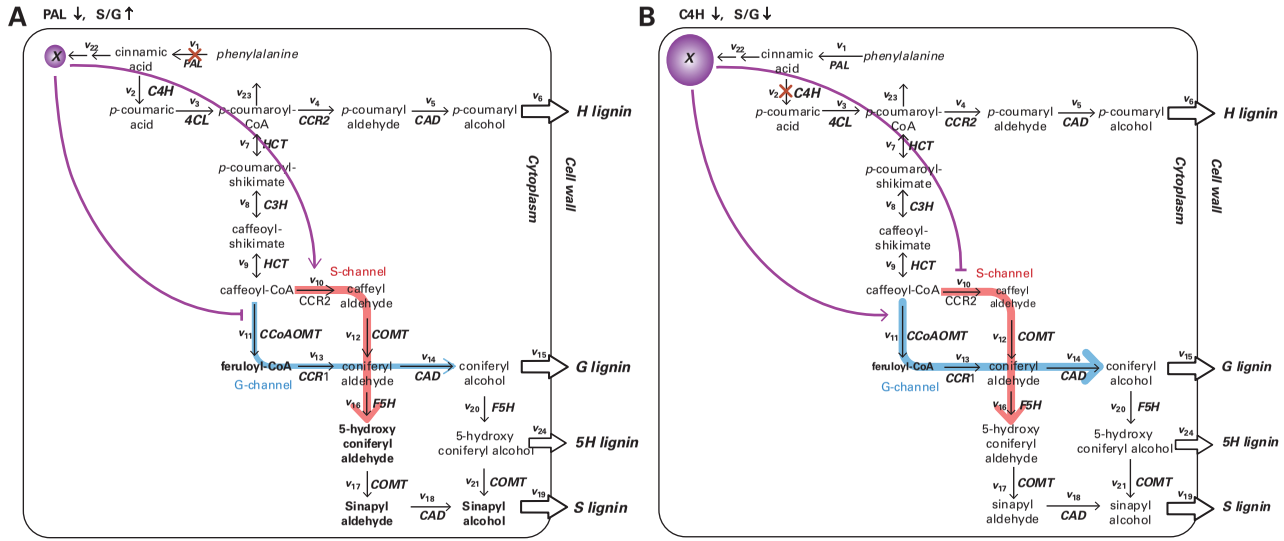


Figure 5.5

The final lignin pathways for the alfalfa model, which includes G10's most significant "model-based postulates." The upward-directed straight arrows on each indicate the reactions are reversible. The red and blue thick lines indicate channels that lead to the S and G monomers. The thickness of the lines indicates the size of the channels, which determines the rate of the reaction. The purple lines indicate the influences of the hypothesized entity X. The curved purple lines ending in arrows indicate a positive influence; those ending in a horizontal bar indicate a negative influence. The size of the circle around X indicates the concentration of X. The red cross indicates the reaction is blocked.

and this postulate made biological sense in addition to making the model work. However, it led to a significant new problem: based on hundreds of thousands of simulations of possible variations of the channelization, these channels appeared unlikely to be regulated by enzyme kinetics. G10 carried out this examination of all possible experimentally supported pathway designs with another novel method he developed of computational enumeration "*to permit an expedient and exhaustive assessment of alternative regulatory schemes*."

Based on the dynamical behaviors of his model, G10 inferred that the easiest way to resolve the problem in the model was make a spectacular biological hypothesis: the established lignin pathway of twenty years is importantly incomplete in that there appears to be an element *outside the current pathway* that has a significant regulatory effect on its behavior. This element would selectively regulate (figure 5.5 curved purple arrows) the pathways (channels) responsible for generating S (figure 5.5A) and G (figure 5.5B) lignin. Because of his limited knowledge of biology, he called the element "X" and had no way to hazard a guess as to what it might be. The postulate is warranted on the basis of the model: if excess cinnamic acid produced a substance X that both up-regulated the G channel flux and down-regulated the S channel flux, then the model produced highly accurate dynamical behavior. His postulation of a heretofore unknown metabolite in the lignin pathway derived from the understanding that the model-building process provided of the quantitative dynamics within the network and of how to control the parameters effectively. As he stated, "*So this is actually the biggest finding from our model. So, by adding this reaction you can see that we hypothesize there is another compound that can give a regulation . . . to other parts of the pathway. And this finding will not be possible if we haven't done any modeling—because, well, if you just look at the data, the data only tells you the composition of these three lignin*."

The model-building process gave G10 a comprehensive view of how the existing data on lignin fit the pathway structure and led him to question that structure as a last resort, because the model dynamics appeared to require it. This prediction finally got the attention of his collaborators, and they conducted experiments that confirmed the hypothesis, identifying "X" to be the signaling molecule salicylic acid. They determined that the molecule acts as an inhibitor of monolignol biosynthesis, which was hailed as a significant biological discovery. G10 also predicted, "*I guess this finding*

will give them more confidence in what we are doing so maybe in the future they could be more willing to give us—to share more data." This prediction was also borne out, in that they went on to collaborate further with G10 in postdoctoral research.

The G10 case leads us to a more general question about computational model-based reasoning in ISB: *How is it possible for an engineer with a few months of biosystems modeling experience and little knowledge of biology to make fundamental discoveries in biology?* To answer this question, we need to fathom how "discovery" is the outcome of processes that create cognitive-cultural systems that extend the capabilities of scientists beyond their basic human limitations to more effectively probe the natural world. Our investigations into discoveries made by in silico simulation modeling, as with in vitro simulation modeling, show these are the epistemic achievements of complex evolving distributed cognitive-cultural systems. For lab G, the distributed problem-solving systems comprise the modeler, model, lab director, other lab members "*grabbed*" for discussion, model-building resources specific to the culture of the lab (here, PCs, ODEs, BST and so forth), conceptual and methodological model-building resources from engineering and computational sciences, epistemic norms and values, experimental collaborators (even if interaction is limited), Internet resources (search engines, data bases, literature), diagrams (pathway, graph), "pen-and-paper" representations, and presentation and publication venues, which provide community feedback. As with all D-cog systems, these systems have properties that are different from those of the individual.

In what follows, I focus on epistemic affordances of specific components of the distributed model-based reasoning system, namely, the coupled system of interaction between two kinds of models, researcher mental models and computational simulation models. I consider ways in which the processes of building the artifact model enhance the inferential powers of the researcher. The back-and-forth interactions between these components of the coupled system create changes that are particularly important to account for the ability of the D-cog system to improve its investigations of a given biological system. I then address the nature of the warrant for believing the outcomes of sufficiently credible models are worthy of pursuit as hypotheses about target systems, which all the modelers noted is an important epistemic aim of these D-cog systems.

5.3 Computational Model-Based Reasoning: Building "a Feeling for the Model"

Interest in the methodology of computational modeling and simulation in science has been growing in the philosophy of science. There is now a substantial body of philosophical research that focuses primarily on physics-based modeling, such as conducted in quantum mechanics, nanoscience, and climate science (see, e.g., Galison 1997; Humphreys 2004; Lenhard 2020; Parker 2009, 2010a,b; Winsberg 2010). These analyses have produced important insights, some of which do pertain to what we have been learning about computational simulation across the board. However, as we argue, there are important differences, as I indicated earlier.

In general, the characterization of computational simulation models Eric Winsberg (Winsberg 2001) has formulated as downward, motley, and autonomous is widely accepted. "Downward" signals that established scientific theories provide the starting point from which to develop a computational simulation model, and that they contribute to the credibility of the model and to the warrant for the belief that modeling outcomes can be transferred, provisionally, to real-world phenomena. "Motley" indicates that the model-building process introduces arbitrary elements that work against any claim that the model is fully derived from theory. To build a stable, robust model requires using a range of such elements, which include abstractions, parameterizations, ad hoc assumptions, mathematical tricks, numerical methods, and much trial and error. In view of their motley nature, in particular, Paul Humphreys (2004, 148) has dubbed complex physics-based simulations as "epistemically opaque" (see also, Lenhard 2007) This means that although they begin from theory and depend on it, the ingredients needed to make a simulation work obscure the operations of the theory's laws and make analytic solutions to equations impossible. Thus, a model can be theory-driven, but in an important sense it is autonomous from theory (see also, Morgan and Morrison 1999). "Autonomous" (or, better, "semi-autonomous") in Winsberg's characterization, also underscores that simulations, customarily, are used in situations where data are sparse because real-world experiment and observation are quite difficult or not possible, and thus simulation provides a source of predictions and understanding that often cannot be checked against—or warranted fully by comparison with—data from real-world sources.

As we have seen in the previous sections, our research on computational modeling and simulation in ISB agrees with the motley characterization (which we likened to a bird building a nest) and the autonomous nature of computational simulation. However, as we also have seen, model-building is not a "downward" process; rather, lacking a theory of the system phenomena, models are built "from the ground up" (MacLeod and Nersessian 2013).[10] There are differences to be discerned from practices that lack a theoretical basis from which to draw resources to build models that are important for understanding how modelers achieve their epistemic aims.[11] One such difference derived from our analyses of ISB practices is to bring out additional, different roles for simulation than have been discussed in the physics-based literature. An important insight from our investigation is that simulation in this domain contributes to building the pathway representation and, so, to the model-building process itself. These and other findings I discussed in the section on general lab G modeling practices underscore the benefit of collecting ethnographic data on the model-building process as it is going on, rather than just relying on published scientific literature, augmented possibly with archival records, retrospective accounts, and anecdotes. There is much that is important for understanding how computational simulation affords epistemic access that is omitted from final reports or not recalled retrospectively.

Much of the recent philosophical literature on computational simulation focuses on issues about whether a new epistemology of science is needed to accommodate computational simulation as an investigative practice or on whether simulation experiments are the same as or different from wet-lab experiments (see, e.g., Beisbart 2018; Frigg and Reiss 2009; Winsberg 2009). These are interesting and important issues, but rather than address them, I consider a largely neglected issue that is important to the discovery question I raised at the end of the previous section. This important issue has only been hinted at in the philosophical literature: the need to bring considerations about human cognition into the epistemology of simulation. Humphreys, for instance, has cast the situation in which science is conducted at least partially by computers as a "hybrid scenario," by which he means "one cannot completely abstract from human cognitive abilities when dealing with representational and computational issues" (Humphreys 2009, 616).[12] Witness, also, the main title of his book, *Extending Ourselves*. Humphreys argues that computational methods and simulation belong to

a long line of "technological enhancements" that scientists have developed as tools to extend human capabilities. Some of these have targeted a specific modality, such as microscopes and telescopes, which have enhanced our native abilities to see. Computational simulation was developed to deal with the problem of processing vast amounts of data, which human cognition cannot. Computational technologies, as Humphreys claimed, provide "enhancements our native cognitive abilities required to process this information" (2009, 8). I agree, and would add, to make inferences from it. However, his claim is not backed up with any account of the nature of the native cognitive abilities that are enhanced by computational technologies—and how they are enhanced. "Extending" is left only as a metaphor.

As I indicated in chapter 3 and as we will see even more so here, there is a significant difference between Humphreys' tool view of extension by means of computational simulation and the coupled system view we have been advancing. In the tool view, over the course of science, scientists have been extending their sensory and cognitive abilities by creating instruments (e.g., telescopes) and analytical tools (e.g., models) that allow them to *use* the artifact as a tool to perform new operations, such as the fast numerical solutions to complex equations performed by computational simulation models. This view suggests the cognitive-cultural divide, from the cognitive side: the individual is able to perform different cognitive tasks—or do them better—using the new artifacts. When our perspective shifts from using a simulation model to building this artifact, we come to understand how the back-and-forth interaction with the human agent incorporates the computational model, along with other elements of culture, such as conceptual and methodological resources and epistemic norms and values, into a *hybrid, coupled human-artifact model-system that accomplishes simulative model-based reasoning*. This construal is compatible with the notion of models as "epistemic tools" (Knuuttila 2005) but focuses attention on the processes of building and incorporating the tool rather than using the final product. "Extending ourselves," in our account, is an iterative and incremental process that incorporates humans and the epistemic tool into a cognitive-cultural system with properties different from those of the individual. This process provides an example in the domain of science of what Hutchins meant more generally by "humans create their cognitive powers by creating the environments in which they exercise those powers." These system-level properties of the "hybrid scenario" facilitate epistemic

access to otherwise inaccessible processes in complex biological systems. We interpret, then, Humphreys' claim that "in extending ourselves, scientific epistemology is no longer human epistemology" (2009, 8) as meaning that *scientific epistemology is the epistemology of a D-cog system* (not only in the case of computational modeling, as we have seen in previous chapters). In the following sections I elaborate on the epistemic and cognitive affordances of building in silico simulation models that I outlined in chapter 3. Specifically, I consider how the modeler's inferential capabilities ("cognitive powers") are extended in model-building processes and the warrant for their claims that these processes can provide epistemic access to the behavior of the biological systems.

5.3.1 Extending the Capacity for Simulative Model-Based Reasoning

As I have discussed in previous chapters, analyses within the D-cog framework customarily cast the human component of a system as "off-loading" cognitive functions to specific artifacts and "coordinating" among system components to accomplish a task. These metaphors, even when explicated in terms of specific tasks, are insufficient to understand how the scientific D-cog system *improves* its ability to investigate target phenomena. Such improvement is driven by learning on the part of the human component, which in turn leads to the further development of the artifact model. We have argued, based on cognitive science research and our own data, that this kind of learning involves building more accurate mental models. We have, thus, cast model-based reasoning with in silico models as a system of interaction—a coupling—between two kinds of models (mental and artifact), which creates changes in the D-cog system that improve its ability to investigate, in the case at hand, complex biological systems.

The reasoning by the modelers captured in our interviews and observational studies, as well as self-reports of their reasoning processes, provide evidence that many of the inferences they make in the course of building a computational model, especially with respect to how and where to modify it, rely on simulative mental modeling. The modelers we have studied across both ISB labs articulate their reasoning in terms of causal interactions in the biological networks, which, we claim, allow them to simulate and perturb limited aspects of the network dynamics mentally. They have walked us through how these simulations—often performed in conjunction with pen and paper or whiteboard representations (see, e.g.,

figure 5.3)—enable them to perform various kinds of reasoning tasks, such as to identify possible errors, explore hypotheses about network structure or parameters in a limited fashion, and identify dominant variables. These simulations enable the modeler, in particular, to screen plausible candidates for fixing errors in the structure of the model before implementing them. This ability is important, because errors in the structure of the system model can have numerous causes and be in numerous locations, so many different manipulations of the computational model might resolve them. The modeler's ability to screen candidates by limited mental simulations cuts down on the work of parameter fitting, which, as we have seen, is a highly labor- and time-intensive process. Our findings are in line with cognitive science findings about how scientists and engineers use mental simulation as they try to solve problems in their research (Christensen and Schunn 2008; Trafton et al. 2005; Trickett and Trafton 2007). Of particular note, that research shows that the use of such mental simulations increases in cases of inferential uncertainty when scientists are trying to develop a general grasp of the phenomena under investigation.

There are three aspects of the character of the simulative mental models our study participants build that we have analyzed as especially significant. First, from the way they reason out loud with their models, we infer that their mental models are qualitative. Modelers, for instance, track qualitative effects of specific variables on other variables using terms like "*increasing*" and "*decreasing*" to describe these relations, such as "an increase in variable A produces a decrease in variable B." Modelers often do sketch out on paper or whiteboard some quantitative details of what they are thinking, but they do not compute precise numbers and values in these activities. Our characterization of their mental models as qualitative is consistent with a range of cognitive science research, especially studies of causal-mechanical reasoning by physicists and engineers (see, e.g., Roschelle and Greeno 1987; DeKleer and Brown 1983).

Second, also in accord with the cognitive science literature, modelers reason about the pathway networks and models in piecemeal fashion in interaction with pen and paper representations (Roschelle and Greeno 1987; Hegarty 1992, 2004; Schwartz and Black 1996). For instance, they track only a limited number of interactions in the network mentally or make inferences about the consequences of manipulating the values of a limited set of variables to explore what might be the effects of specific

modifications to the computational model. As one modeler recounted, the modeler "*has to visualize the pathway in his head and divide it up into parts and write codes for each part*," which is why the modeler "*draws so much and uses so much paper*."

Third, and likewise in accord with cognitive science research, modelers appear to reason by carrying out simulations with these piecemeal models. Research on nonexpert reasoning about simple mechanical pulley systems (Hegarty 2004; Schwartz and Black 1996), for instance, establishes that participants reasoned by carrying out simulations of intermediate pulleys in the system, which facilitated their ability to reason over a larger scale. This strategy is consistent with constraints on working memory that limit how much information can be processed at a time. For modelers, these constraints mean that they should be able to track and manipulate only a limited number of variables at any one time, which accords with our research. We have seen modelers use selective and piecemeal representations of the system, for example, to identify and bracket nonlinear relations into separate behaviors and simulate each separately to make inferences. In their mental simulations, modelers usually focus on elements of the pathway network that interact directly, but are not necessarily contiguous. These qualitative simulations of pieces of the network help them to understand the qualitative effects of the quantitative mathematical relations represented in computational model as they build the model. In the cognitive literature, such qualitative simulations have been called "envisioning" (DeKleer and Brown 1983). As one modeler described her envisioning process in building an intuition about her model, "*So the thing is—when you want to solve a mathematical problem . . . sometimes you use numbers and try numbers, something to give you a feel of—like intuitively how this, for example, equation works and all. So, I'm trying out numbers and then trying to make the steps kind of discrete—like sort of a state machine, kind of thinking like we're in this state. And then, now this much is going to this other metabolite pool and then, at the same time, we have less of that. So, I'm trying to see what the constraints are by actually like doing a step-by-step sort of thing*." While she was describing this to us, she was also using her finger to point to and trace out her sketches of these "*steps*" sketched in her notebook.

As I discussed in chapter 1, some cognitive scientists have proposed the way to understand how mental models and external representations work together during reasoning is as *coupled inferential processing* (see, e.g., Greeno

1989a,b; Zhang and Norman 1995; Gorman 1997; Hegarty 2004; Nersessian 2008). However, unlike the case of coupling between mental and static artifact representations considered in this literature (mainly diagrams), in the case of computational representations, both kinds of models have their own simulation capabilities. Our extension of the coupling proposal to include in silico models proposes that the incremental and iterative processes of building and simulating the computational model create a key change in the D-cog system. Namely, the process builds a close dynamic coupling between the modeler's mental model and the artifact model that incorporates modeler and model into a powerful *simulative model-based reasoning system* that significantly enhances the limited human capability to reason about the behavior of complex biological systems (Chandrasekharan and Nersessian 2015; MacLeod and Nersessian 2018). Notably, the coupling enhances the human cognitive powers used in mental modeling, such as memory, information synthesis, visualization, simulation, abstraction, imagination, and intuition. In the way we propose to understand the model-building activity, *cognitive functions are not off-loaded to the computational model, but are enriched and extended into a coupled system by virtue of it*.

As we saw, the computational model can integrate a vast amount of information from disparate sources. Further, computers have the capacity to process complex systems of quantitative representations, such as the twenty-four equations needed to build G10's model. Their speed and manipulability enable the modeler to implement changes quickly and efficiently so that he can run through pathway options or hypotheses in quick succession. The computational model can generate many kinds of visual representations, such as graphs to track only specific relations or three-dimensional visualizations to track dynamic system behaviors. The choice depends on what the modeler thinks most useful for the problem. The computational model can, also, be put through thousands of simulations of many configurations in a matter of seconds. Configurations that use, for instance, different time points or parameter values produce different network behaviors that the modeler can partition into families of mental models, which can be used to build her intuition about the behavior of the model and develop insight into how to proceed with the building process. In addition, numerous and diverse simulations enable the modeler to develop a holistic, global perspective on the system dynamics. Model simulations, in addition, enhance the modeler's ability to think about possible worlds and make counterfactual

inferences in ways that outstrip her capacity for thought experimenting alone.[13] The model's representation in variables, in particular, promotes such counterfactual explorations. Thinking in variables, too, helps to build what the lab G director called "*the flexibility to recognize shared features of control/regulation across disparate domains*." This kind of cognitive flexibility allows the modeler to move with relative ease from modeling yeast to modeling cancer, and so forth.

The overall effect of the back-and-forth exchange between these components of the D-cog system is to extend human inferential powers such that the modeler can make reasonable inferences about how to build out the pathway or improve the parameter fit of the model. The model-building process is not always successful. However, a stable and robust model (or model ensemble, such as G10 developed) can lead to hypotheses about how to understand the system-level behavior in the target or how to manipulate it, such as the significant and novel "*model-based postulates*" made by G10. G10's major biological discovery did not involve gaining expertise in biology (thus the designation "X" for the unknown biological entity) but did involve developing confidence in his judgement in the warrant for the inference that computational model did enact the behavior (exemplify) of the in vivo system (hypothesis transfer)—a confidence buoyed from numerous iterations of model-building and simulation. In the end, he could postulate with confidence that some heretofore not considered element is part of the regulation of the lignin pathway, because the addition of it produces stable dynamic behavior in the model. The warrant for making such a bold move derives from the processes in which G10 created and examined numerous model variations and found that every plausible biologically reasonable change other than this one fails to provide a good fit.

To verify, hopefully, this hypothesis and determine what "X" is required action by the collaborator component of the D-cog system, and, as we saw, they were scarcely involved beyond supplying data. Here we see that another epistemic affordance of simulation modeling in ISB is to enhance collaboration, when biologically plausible hypotheses intrigue experimentalists sufficiently to pursue them. Every experimental collaborator we interviewed expressed a degree of skepticism about computational modeling, even when they had sought out the collaboration. The collaborators often complained that modelers seemed to want to build models "*for their own sake*," and were content with just replicating data—sometimes very old

data ("*who cares about that?*")—which the experimentalist characterized as a "*tautology*." As an experimentalist stated in discussing another lab G modeler's work with us, "*I think it's absolutely essential for anybody who is going to model to build in a step of their modeling where they test its predictive power. . . . If we get some answer [experimentally], as she did, I'm going to have a lot more confidence in your model*." Having confidence in the possibilities of modeling on the part of the experimentalist is prerequisite to effective collaboration. As we saw with G10, once his collaborators had that confidence, they actively pursued further, more engaged, collaboration.

In all our investigations, we encountered the claim by computational modelers that it is of great importance to develop "*a feeling for the model*." Our analyses interpret this "feeling" as having several dimensions. One aspect refers to the intuition and confidence modelers develop about the behavior of the model through the coupling process, as well as about their ability to correct deficiencies in the desired direction of a stable and predictively robust model. The central role modelers ascribe to developing a feeling for the model in order to make progress underscores, for our analysis, that although model-based reasoning is carried out by a coupled inferential system, specific attention needs to be paid to the human component. Ultimately, it is the human agent who has to draw the inferences about how to proceed to improve the model or to flesh out the potential implications of the model's behavior, as well as possess confidence in the direction they choose. As such, the level of complexity a modeler can handle likely constrains the size of the models that are productive for the modeler to attempt to build. This consideration provides an important additional rationale for the mesoscopic modeling strategy that Voit et al. 2012 have observed to be prevalent in ISB.[14]

We think this phrase, too, is an important indicator of how the processes of building the simulation model provide insight and understanding about the target biological system; the "feeling for the model" in turn provides the modeler, by analogy, with a "feeling for the biological system." The model-building process gradually builds intuition about the dynamics of the target behavior through a large number of iterations of simulations wherein a range of factors such as sensitivity, stability, consistency, computational complexity, and so forth are explored. In the process the modeler, interactively with simulation, builds out the structure of the pathway network, which delineates a sequence of causal interactions among the

elements of the biological pathway. The simulations of the model's behavioral dynamics build an intuitive understanding of how the pathway generates the existing experimental data and what interventions might be made in the target while its stability is maintained. Simulations can also be used to explore why other sets of values are not (or have not been) seen in real-world systems, which can provide the modeler some insight into "design principles" that underlie the values seen in in vivo experimentation. These repeated interactions with the model, pathway, and biological literature develops the modeler's capacity to judge the biological "reasonableness" of the hypotheses or predictions they make about the system.

In order for the experimentalist to intervene on a target system, the model does need to provide specific causal information about the system. As we saw in the case of G10, he made causal predictions, based on the behavior of the model, about what enzymes might be knocked down to lower the S/G ratio but maintain structural integrity of the plant and predicted by what percentage these knockdowns would decrease the natural ratio. How to tweak the lignin pathway was the objective of the modeling, but the model also enabled an unanticipated, even more significant, causal prediction (figure 5.5): if cinnamic acid (postulated flux leaving the system) produced a compound ("X") that both up-regulated the G-channel (pathway) flows and down-regulated S-channel flows, then the model would produce highly accurate dynamic behavior in accord with the existing experimental data. In our discussion with the lab director about this case, he pointed out that such mesoscopic models can provide "*a certain level of explanation . . . something causal you didn't know before*," pending, of course, experimental verification.

In general, as seen in the G10 case, the director claimed, "*if you can trace out a causal pathway, then it's an explanatory model even though you may not know every single detail*." However, this kind of explanation is not mechanistic because "*with every [such] explanatory model, you have some regression in there or some association. . . . It's not pure*." That is, the top-down abstraction strategies—such as those associated with canonical mathematical frameworks, shrinking and fixing the parameter space, and global fitting algorithms—used to build the model wash out or obscure details of the mechanisms underlying the causal connections. Nevertheless, the causal information provided by the model about the pathway structure ("*trace out a causal pathway*") does provide a "*certain level of explanation*" about

the dynamical behavior of the in vivo biological system, and, in particular, how, possibly, to manipulate it.

In most instances, the kind of understanding mesoscopic modeling provides is largely pragmatic—understanding about the target system sufficient to propose ways to manipulate or control it to attain desired outcomes, but not sufficient to explain its behavior fully. Lenhard (2006) has argued, with respect to a case he investigated in nanoscience, that there are instances in this kind of understanding in physics-based computational modeling too. In such instances, models begin from theory but the equations produced for the complex phenomena are impossible to solve analytically. Instead, "simulations squeeze out the consequences in an often unintelligible and opaque way" (612), because of abstracting and averaging techniques that fit the equations to the data, as well as the numerical methods that render equations computable. He argues that, even though such simulation models do not provide the kind of explanatory understanding one derives from laws, these models do provide understanding about how, possibly, to intervene, control, or manipulate the phenomena, and thus, pragmatic understanding.

Systems biologists often write that the goal of the field is to attain "systems-level understanding" of complex biological phenomena, which they cast in terms of theories that capture general mathematical features and properties of biological systems, from which models of individual systems can be derived (to the extent possible in physics) (see, e.g., Kitano 2002; Westerhoff and Kell 2007). We interpret this aspiration to mean systems-level understanding, eventually, should be not just pragmatic, as a capacity for manipulation and control alone, but a genuine theoretical or mathematical form of understanding from which the ability to manipulate and control would follow. This is the ideal scenario. In practice in the current state of the field—at least from what we have witnessed—the complexity of the systems and the constraints on model-building are such that modelers pursue more limited goals with respect to what they can learn about their systems. They make the pragmatic decision to pursue less detailed and robust models that, in principle, can be predictively accurate for only certain elements of the systems. The understanding such models provide is pragmatic also in content, in that they provide neither a higher-level mathematical/theoretical understanding nor a mechanistic explanation (MacLeod and Nersessian 2015). Thus, the situation Lenhard describes in some physics-based modeling is the current state of computational

modeling in at least the area of ISB we have investigated, if not more widely.

Another aspect of the modeler's claim to have a "feeling for" the model is that it indicates a belief in the credibility of a validated model's predictions. In our interviews with modelers and in the presentations of their research that we witnessed, computational modelers (in lab G, lab C, and lab D) exhibit a high degree of confidence that their models, when fitted and rigorously run through diagnostic and cross-validation testing, produce simulations that do exemplify the dynamic behavior of the target systems. What warrants that confidence? In some cases, it will not be possible to conduct experiments on the system to back up this belief, and where it is possible, it often requires a considerable investment of time and resources on the part of experimental collaborators, so the modeler's confidence that predictive inferences are credible and worthy of pursuit needs to be quite high. Of course, this belief is fallible since models can be wrong even if they fit the available biological evidence, but modelers do express confidence that a validated model is correct "*in respects that matter*."

5.3.2 Building Epistemic Warrant

The most detailed philosophical account of the epistemology of computational simulation modeling is that of Eric Winsberg (2010), based on an analysis of physics-based modeling largely as recorded in the published literature. As he points out, models are built in data-scarce situations to serve as alternatives to real-world experimentation, which in many instances cannot be carried out (think of colliding galaxies or climate change). He proposes that the credibility or epistemic warrant for a model rests on two pillars, which are related to his characterization of simulation models as downward, motley, and autonomous. The first pillar is the credibility of the theory of the phenomena, such as fluid mechanics, that informs the building process (downward). But the methods required to build a stable and robust model always introduce extraneous and arbitrary elements into the process (motley), which give it autonomy from theory. So, the epistemic warrant also, importantly, derives from the second pillar, the credibility of the methods used in building the model. As we have seen, the case of ISB modeling is importantly different with respect to these sources of credibility. There are no guiding theories of the biological phenomena under investigation. Instead, modelers assemble the network of reactions and

regulatory relations among elements (in our case, metabolites and signaling molecules) of the system in conjunction with literature searches and preliminary simulations of the model as they are building it. To build a computational representation, they use bits of what they sometimes refer to as "theory," such as enzyme kinetics, and canonical frameworks, such as BST, that provide a possible structure by which to glue together the lower-level information. With respect to the various methods used to build the model, these have, largely, been developed to model human-made systems. These methods have considerable credibility for modeling those kinds of systems, but are, here, being used on living systems—natural, modified (e.g., genetically), or engineered (e.g., synthetic).

In the physics-based modeling fields Winsberg considers, the methods do have considerable "antecedently established credibility," in that established "disciplinary tradition" supports their reliability in application to new cases—that is, they are "projectible" (2010, 137). Further, there are, of course, computational techniques related to fitting numerical models generally, such as Monte Carlo methods, that can be applied whatever the subject. In the ISB case, though, it is often an open question whether and what engineering modeling methods can be applied or how they might be adapted. As I noted, we often saw modelers experimenting with the application of methods from the engineering domain in which they were trained, such as wave-smoothing techniques from telecommunications to smooth noisy biological data. Still, with respect to the credibility of model-building methods, even though these are drawn from a discipline other than that in which they are used, much of what Winsberg argues about *how* they gain credibility does apply.

He calls techniques and assumptions made in applying various methods "self-vindicating," by which he means "whenever they produce results that fit well into the web of our previously accepted data, our observations, the results of our paper-and-pencil analysis, and our physical intuitions; when they make specific predictions or produce engineering accomplishments—their credibility as reliable techniques or reasonable assumptions grows" (Winsberg 2010, 122). His is a thoroughly pragmatic stance: methods are vindicated by the fruits they bear, which is the case across the history of development of scientific methods. So, too, the techniques and assumptions of the methods that are transferred from engineering systems and adapted to biological systems gain this kind of pragmatic credibility and

become projectible as they develop an interdisciplinary history in biosystems modeling. Further, productive strategies used to solve frequently encountered problems gain traction as reliable parts of the practice for both the individual and the community, as do the novel methods for building models of a specific type (e.g., G10's two-step method for modeling in the lignin domain) developed within these emerging epistemic cultures, which have not yet become established tradition.

Winsberg's other claim is that well-established theory in physics-based modeling plays an important role in helping to mitigate some of the arbitrary features of the model, and enhance its credibility. In lieu of that source of credibility, I consider what aspects of the model-building process in ISB might serve to confer credibility on the model, especially as these relate to the role of the pathway representation in the process.

First, the scope and range of the data integrated into the model cover data for all related systems. Initially, the data are split into two sets, one the modeler uses to build the model and the other the modeler uses to run cross-validation tests after the model is fitted. The integration of data develops in interaction with building out the pathway structure, which lays out the causal sequence of connections among the elements of the system. In this interactive process, the modeler can take different pieces of the network and simulate their behavior in various combinations and configurations. The modeler can run unlimited simulations (recall G10 ran ten thousand to examine just one piece of the pathway). These simulations enable the modeler to consider whether adding or deleting pieces of the pathway are biologically reasonable moves.

Second, the simulation process provides the modeler with significant ability to control and manipulate the model's behavior. The modeler can stop and start the simulation in every state, which enables her to track the system variables (nodes in the pathway) that generate specific behaviors and to determine detailed measures of significant variables. The modeler can also track every time point of the state of the simulation, which enables her to change the time at which some process kicks in, among other modifications. Such manipulations enable the modeler to interrogate the dynamics of the system and develop a sense of how the pathway as constructed could generate the experimental data, as well as a sense of what changes in the pathway might be productive, again, consistent with their biological reasonableness.

Third, the mutually constraining nature of the data, parameters, and pathway in the model-fitting process helps to mitigate some of the arbitrariness of model-building. The notion of fit is complex. It does not mean that the model provides a point-by-point replication of the data for all variables. Rather, at least for the kind of modeling we have studied, it means that the model replicates trends in the experimental data for most major variables. "Fit" is often construed as a matching process in which there is a satisfactory match between the data generated by the final model and the experimental data, usually determined by comparing graphs of each. However, from studying the practices of the modelers in our labs, we have come to understand fit as a dynamic and interactive process among the three elements that works to enable the modeler to home in on a satisfactory representation for both pathway and model.

There are three changeable components—pathway structure, experimental data, parameter values—that become increasingly constrained by their interactions in the highly recursive fitting process. To estimate unknown parameters, the modeler uses fit with experimental data as an anchor. For each change in parameter, the way the output of the model maps to the experimental results changes. Only parameter values that improve fit, or keep it at its current state, are retained. Although it cannot be done for all parameters, the modeler screens parameters for their biological plausibility to the extent possible. Each replication of experimental results in a simulation infuses more, and disparate, data into the model and changes the parameter structure. During the process, fit is used to add or delete components of the pathway network. As we saw, inferences about how to build out or modify the causal network derive largely from the behavior of the simulations. These simulations enable the modeler to infer whether her conjectures as to network structure are on the right track by running the model with and without various pieces, which requires changes in parameters. Equally important, simulation enables modelers to infer missing network structure, which they check for biological plausibility, such as G10's addition of reverse fluxes and channelization of the S and G monomers. And, although we were told such outcomes are rare, simulation has the potential to point even to the possibility that an element thought to be outside an established pathway could be affecting the behavior of the system, as in the inference about the element "X" made by G10.

Through the three-way locking-in process, the model gains complexity. The fitting process is, of course, not without risk of introducing unwarranted elements into the model, especially as there are often some parameters that can only be fit by Monte Carlo simulation. Even a well-fitted model is underdetermined. However, there are ways to further enhance its credibility. Once a satisfactory fit is obtained, for instance, the modeler performs cross-validation tests on the model (or a small ensemble if the fit is not unique) with additional data and diagnostic tests, such as how it responds to perturbations, which, if passed, add to its credibility.

At the end of these constructing, fitting, and validating processes, the modeler can build sufficient warrant to believe, provisionally, that a robust and stable model does enact the behaviors of a generic system, that is, it exemplifies the behaviors of the class of biological systems. As repeatedly expressed by the modelers in our labs, though, they aim to build models that not only replicate the available data, but also provide substantive predictions. As with the in vitro models built in BME, ISB modelers transfer predictions about behavior from the in silico models they build to the target systems using analogical inference. ISB modelers deem predictions that derive from a stable and robust model that exemplifies the known behaviors of the target system sufficiently credible to warrant investigation by experimentalists. If these predictions are verified, it further enhances the credibility not only of the model, but also of the model-building methods.

5.4 Summary: "Getting a Grip" with/on In Silico Simulation Modeling

For some time, scientists have been using computational simulation to gain epistemic access to the behaviors of complex dynamical systems, from colliding galaxies to climate systems. Only recently, though, has it been used to investigate biological systems. This has been due, in part, to the methodological problem of how to build computational models of these systems in the absence of the resources provided by a theoretical basis and, in part, to the technological and methodological problems of how to collect sufficient data of the right kind (time series) or to get around the lack of data with appropriate algorithmic strategies. Although ISB is a diverse field, the modeling practices in labs we have been studying are representative of a major area that draws conceptual and methodological resources from engineering fields, including electrical engineering, control engineering, systems

engineering, and telecommunications engineering, to build models of complex biological systems. Although the long-term objective expressed by researchers in the field is to develop a systems-level theory by which to understand and predict behaviors of complex biological systems, our labs expressed more modest aims. The lab G director is the one who expressed their more immediately obtainable aim with the particularly apt phrase, "*getting a grip*" on systems behavior—that is, an understanding sufficient for predictions that at least can enable manipulation and control. As we have seen, modelers also need to get a grip on the challenge of building models of large-scale biological systems in the face of numerous constraints. We have examined in detail some of the practices modelers in lab G have been developing to manage the complexity of this challenge.

Instead of articulating theories into informative computational models, researchers in this area of ISB need to compose their models by collecting the needed dynamical and structural information themselves from a variety of sources, including their own simulations, in an iterative and incremental fashion. Our cognitive-ethnographic investigations on how they build models provide valuable insights into the processes that are unlikely to be found in examining only the published literature, as has been the case with most of the philosophical accounts of physics-based computational modeling, or even archival material, to the extent it exists. We have been able to detail how they build models from the ground up by piecing together in nest-like fashion principles from molecular biology, experimental results, information gathered from literature surveys and databases, canonical frameworks, and computational algorithms to create representations of biological systems in data-poor environments.

Simulation, which is the central methodology for experimentation in computational modeling, is often seen as the end phase of the research. Our in situ examination of the model-building processes brings to the fore the key roles of simulation in building the model itself. Simulation is a means through which the modeler develops the biological pathway and comes to learn and assemble the relevant ontological features of a system. The modeler continues to adapt the pathway network in conjunction with simulation throughout the model-building process until pathway, experimental data, and parameter fit coalesce into a stable and robust model (or small set of models). The pathway representation is shaped by issues of available parameters and parameter estimation tools. Likewise, pathways are tailored

to fit the capacities of the mathematical frameworks and whatever mathematical tools the modeler can bring to bear. At the same time, these frameworks determine the extent of the parameter fixing problem. The modeler keeps all these elements in dialogue during the model-building process. In this regard, simulations play an important functional role in how modelers learn how to assemble information and to construct a computational model that gives the right kind of representation. Modelers assemble the needed information in the course of an exploratory process that involves preliminary simulations, both computational and pen-and-paper, and subsequent refinements and revisions. This process enables the modeler to build up her own understanding of the dynamics and relevancies of particular pathway elements. Building this understanding can require the modeler to make judgments about adding elements to the pathway not discussed in the literature, such as hypotheses about elements that must be playing a role or about feedback relations that are not documented but are required by the model, as we saw with G10.

In sum, we have analyzed ways in which the processes of incremental and iterative model-building and their attendant processes of simulation are the means through which modelers come to understand their model and their biological systems. This, in turn allows them to make better judgements about what to include or exclude and which tools and techniques will help and which, not. As per the nest analogy, simulation provides them with the means to work out the best, or most stable, way to pack the pieces together.[15] There, thus, is an important cognitive dimension to simulation in that the iterative back-and-forth interaction between the modeler's mental model and the computational model, which we characterize as "coupling" between these parts of a D-cog reasoning system, is an essential part of the ISB problem-solving practice that builds models of complex systems that lack a basis in theory. The affordances of simulation as a cognitive resource in the ways we have delineated make building representations (pathway and model) of such complex systems without a theoretical basis possible.

Modelers uniformly use the expression "*a feeling for the model*" to characterize the understanding they develop of the behavior of the model over the course of the building process. Simulation is a major source of this feeling. The modeler comes to understand the model's dynamics through numerous iterations of simulation under various conditions and uses the feeling

to guide the direction to develop the model. From our observations and interviews, other aspects of this "*getting a feeling*" include a growing insight and understanding into the target in vivo system and a growing belief in the credibility of the simulation model as a dynamic enactment of the behavior of the system. A further aspect that needs to be considered is the affective dimension. The "coupling" developed between the researcher's mental model and the in silico artifact model creates an intimate connection, the importance of which should not be discounted.

"Feeling" is intimate language. In computational modeling it is directed toward an artifact with dynamic behaviors with which the modeler is interacting intimately ("coupling") as she creates it.[16] Some readers are likely to recall Barbara McClintock's expression of a "feeling for the organism," which she deemed critical to her biological research, especially to her discovery of genetic transposition, which was dismissed at the time but, many years later, was awarded the Nobel Prize. What Evelyn Fox Keller, a scientist herself, has said in her penetrating biographical analysis of McClintock, applies equally here: "Good science cannot proceed without a deep emotional investment on the part of the scientist. It is that emotional investment that provides the motivating force for the endless hours of intense, often grueling, labor" (Keller 1983, 198). We have witnessed such emotional investment across the labs we investigated, and have analyzed how expressions of it, too, are tied, importantly and intimately, to epistemic achievement. We have examined this investment in depth in the BME labs in particular (see, especially, Osbeck et al. 2011, chapter 3). Affective engagement with the objects of one's research does not taint scientific knowledge, rather it makes it possible. The computational modelers in the ISB labs themselves recognize their "feeling" develops only through the hard, slow work of building out the model, which is necessary for them to develop insight into and understanding of the model, as well as the target system. The lab G director shows his recognition of the importance of this work when he strongly encourages his modelers to invest considerable time in exploring and playing around with their models. He does the same in the biosystems modeling classes he teaches.

Interdisciplinarity in the Making

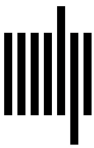
Models and Methods in Frontier Science

By: Nancy J. Nersessian

Citation:

Interdisciplinarity in the Making: Models and Methods in Frontier Science
By: Nancy J. Nersessian
DOI: 10.7551/mitpress/14667.001.0001
ISBN (electronic): 9780262372275
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Nersessian, Nancy J., author.
Title: Interdisciplinarity in the making : models and methods in frontier science / Nancy J. Nersessian.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021061880 (print) | LCCN 2021061881 (ebook) | ISBN 9780262544665 | ISBN 9780262372268 (epub) | ISBN 9780262372275 (pdf)
Subjects: LCSH: Biotechnology—Methodology. | Bioengineering—Methodology. | Biotechnology—Research—Case studies. | Bioengineering—Research—Case studies. | Biotechnology laboratories. | Scientific surveys. | Interdisciplinary research.
Classification: LCC TP248.24 .N47 2022 (print) | LCC TP248.24 (ebook) | DDC 660.6—dc23/eng/20220720
LC record available at https://lccn.loc.gov/2021061880
LC ebook record available at https://lccn.loc.gov/2021061881