Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

In silico model-based inference: an emerging approach for inverse problems in engineering better medicines

David J Klinke II[1,2] and Marc R Birtwistle[3]

Identifying the network of biochemical interactions that underpin disease pathophysiology is a key hurdle in drug discovery. While many components involved in these biological processes are identified, how components organize differently in health and disease remains unclear. In chemical engineering, mechanistic modeling provides a quantitative framework to capture our understanding of a reactive system and test this knowledge against data. Here, we describe an emerging approach to test this knowledge against data that leverages concepts from probability, Bayesian statistics, and chemical kinetics by focusing on two related inverse problems. The first problem is to identify the causal structure of the reaction network, given uncertainty as to how the reactive components interact. The second problem is to identify the values of the model parameters, when a network is known a priori.

Addresses
[1] Department of Chemical Engineering and Mary Babb Randolph Cancer Center, West Virginia University, Morgantown, WV, United States
[2] Department of Microbiology, Immunology, & Cell Biology, West Virginia University, Morgantown, WV, United States
[3] Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding author: Klinke, David J (david.klinke@mail.wvu.edu)

Current Opinion in Chemical Engineering 2015, **10**:14–24

This review comes from a themed issue on **Biotechnology and bioprocess engineering**

Edited by **Eleftherios Terry Papoutsakis** and **Nigel J Titchener-Hooker**

http://dx.doi.org/10.1016/j.coche.2015.07.006

Introduction

Chemical engineering has a rich history in using mathematical modeling to describe chemical systems [1,2]. As summarized in Figure 1, mathematical models aim to capture our understanding of the underlying mechanisms associated with a variety of reactive and physical processes that govern the behavior of chemical systems. The particular formulation of the mathematical model represents a trade-off between computational or analytical tractability and realism. The art of modeling is finding the right level of abstraction appropriate for the task at hand. The mathematical models are used in one of two scenarios. The first scenario is forward modeling, where a mathematical model is used to predict future behavior based on changing operating conditions. The second scenario is for inverse problems, where mathematical models are used to interpret observations of a chemical system as a way to identify the governing physical and reactive processes.

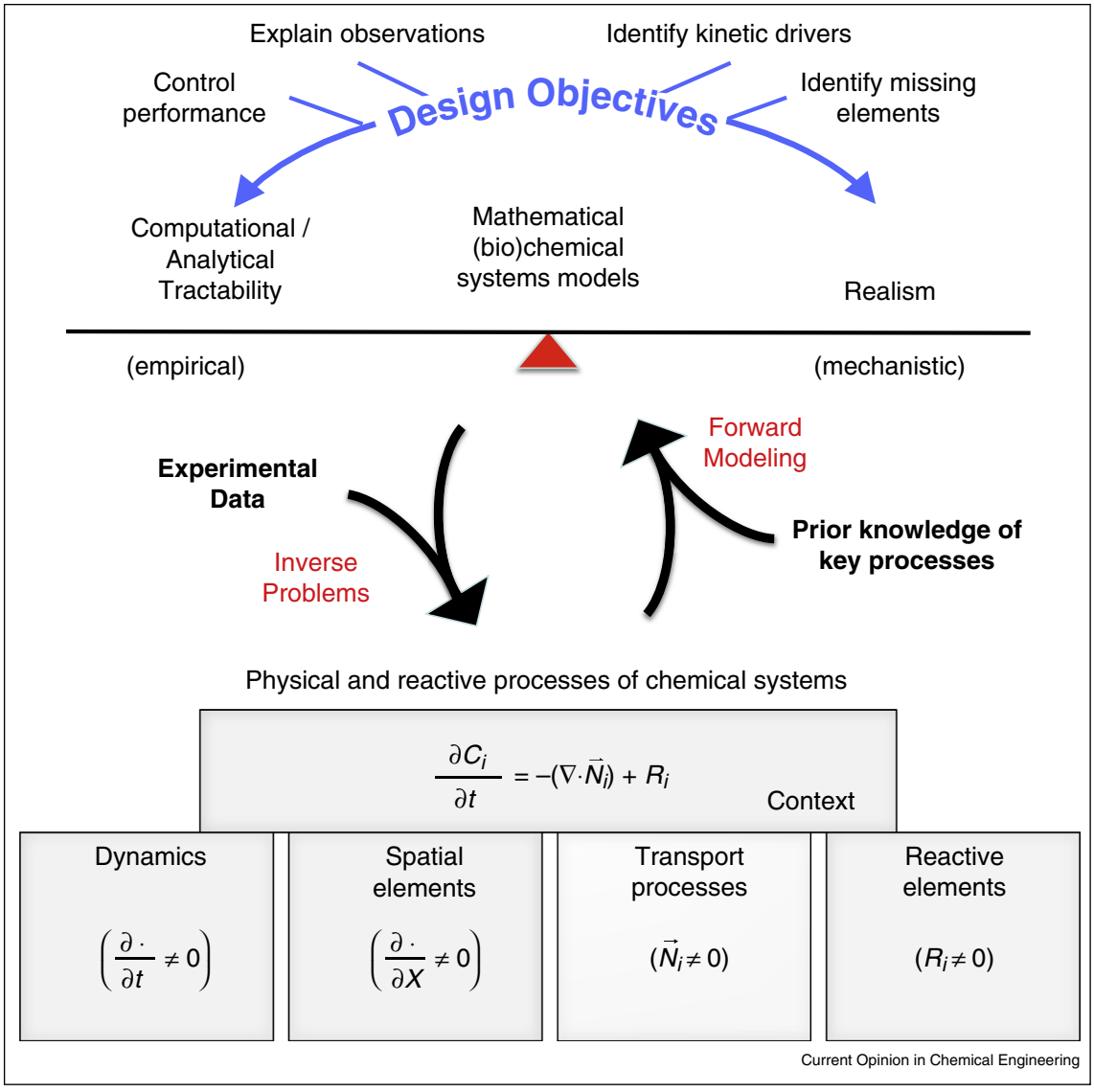
Predictions derived from a mathematical model depend on the relationships specified between the interacting species modeled, which is expressed generally as a reaction network, and the speed of information transfer by a modeled interaction, which is captured by a model parameter. Conceptually, there are then two types of inverse problems. The first type of inverse problem, which is more frequently encountered, is parameter inference. Parameter inference involves selecting a set of parameter values or initial conditions that enable the mathematical model to capture the observed data, where the reaction network is known a priori. The second type of inverse problem is network inference, which is more challenging. Network inference involves selecting an appropriate set of interactions among a set of plausible interactions based on the limited information available about the reactive and physical phenomena contained within the chemical system. In dynamic systems, the reaction network alone provides constraints as to how the model species could potentially evolve in time. The goal of network inference is then to see if the postulated reaction network can capture the observed data given any plausible combination of model parameters and initial conditions.

Historically, these two types of inverse problems have been treated similarly. However, changes in the scientific landscape and advances in technology motivate approaches specifically tailored to the specific inverse problem at hand. The focus of this review is as follows. First, we will discuss pharmaceutical drug discovery and development as a pressing area of inverse problems. Second, we will discuss some of the conceptual and technological advances that have enables a more tailored approach towards inverse problems that focus on either network inference or parameter inference, especially in cases where significant prior information exists.

Improving confidence in target selection is an important class of inverse problems in pharmaceutical R&D

One important application of mechanistic mathematical models is to help in the discovery and development of

Figure 1



Overview of the role of mathematical modeling in the context of reactive systems. The behavior of reactive systems can be described in mathematical terms that represent the underlying physical, kinetic, and reactive processes. The specific mathematical relationships incorporated into a model reflect a trade-off between computational and analytical tractability and realism that relate to the specific design objectives associated with how the mathematical model will be used. Modeling applications can be categorized into one of two applications. The first application, called forward modeling, predicts the behavior of the system based on prior knowledge of the relative importance of the physical, kinetic, and reactive processes. The second application is to identify aspects of reactive system that govern the behavior of the process using experimental observations of the system. This second application is called an inverse problem.

therapeutic drugs. Using pharmacologic agents to treat disease underpins many of the clinical successes in modern medicine. While these successes encourage an optimistic outlook, there remain unmet medical needs despite decades of intense scientific effort. The emergence of multi-drug resistant bacteria, infectious disease outbreaks, and mixed progress in the war on cancer highlight some of these unmet medical needs and the complexity of a constantly changing therapeutic landscape. Engineering better medicines to fulfill these unmet medical needs is one of the grand challenges of engineering [3]. To address these needs, new therapies are discovered and developed by a multi-stage process that relies initially on model experimental systems and then testing in humans using clinical trials. With costs associated with bringing a new therapy to market exceeding $1 billion and the high level of attrition during the research and development process, there is significant

concern over sustainability of the current model for innovation in the industry [4–6].

In a recent NIH white paper, an industrial and academic working group found that the source of attrition had shifted over the last couple of decades from Phase I clinical trials, which focus on toxicology, to Phase II clinical trials [7••]. The objective of a Phase II clinical trial is to test the therapy in patients diagnosed with the disease. A Phase II success is to improve clinical outcome relative to the current standard of care. To illustrate the challenges of demonstrating efficacy, a recent retrospective analysis of clinical trials in patients with acute myeloid leukemia found that, of the 37 therapies that received positive indications in early phase clinical trials, only one drug actually made it to clinical use [8]. This NIH working group observed that the failures in demonstrating efficacy stem from an incomplete understanding of clinical importance of a specific biological mechanism that was targeted by the therapy. Understanding how a drug interacts with a target is challenging as many diseases of therapeutic interest (e.g. cancer, heart disease, and diabetes) are multigenic, progressive, and heterogeneous in that each case may have a different mechanism of origin. While preclinical studies using model systems supported the clinical trials, the failure in translation suggests that the model systems have unclear fidelity in capturing the complexity of human disease.

One of the main recommendations to improve innovation was to focus on a network-centric view of biology to balance the 'one-gene, one-receptor, one-mechanism' (OGRM) paradigm prevalent within the industry [9]. In short, methods developed under the OGRM paradigm select drugs that modulate a specific therapeutic target in experimental systems that have been taken out of context. From the network-centric perspective, a drug target resides within a complex network of interactions that responds in dynamic and non-linear ways to therapeutic modulation. Moreover, these networks can be altered in disease such that the importance of a particular target in regulating phenotype can be quite different in health and disease. Beyond this, current multi-genic and progressive diseases are the result of a multi-variate pathology such that drug efficacy is poorly predicted by only considering a drug's primary target. For example, most so-called 'targeted' tyrosine kinase inhibitors for cancer therapy actually have broad spectrum activity against several kinases, and in fact some of the 'dirtiest' drugs that target multiple kinases are some of the most effective [10].

Upon this network-centric foundation, mechanistic modeling and simulation are integrated with quantitative wet lab studies to advance the systems-level understanding of the pathophysiology relevant for drug discovery and development. While the mechanistic modeling and simulation aspects are more aligned with the discipline of chemical engineering than pharmacology, the strong focus on translational medicine motivated the group to coin a new field called quantitative and systems pharmacology.

Quantitative and systems pharmacology versus systems biology: What's the diff?

On a simplistic level, one may view quantitative and systems pharmacology (QSP) as a simple extension of systems biology with the addition of drug dynamics. However, they are motivated by two different objectives and are orthogonal approaches to organize data and knowledge about biological systems [7••,9]. QSP is motivated by applied translational research questions that require vertical integration. To inform drug discovery, the structure of the model tends to be focused around the targeted pathways and disease mechanisms. Mathematics is used to integrate vertically data and mechanistic knowledge that span multiple levels of biological organization thereby linking molecular targets to clinical read-outs. For instance, a team of 3 PhD-level engineers and 3 PhD-level immunologists worked for two years to build a mechanistic model of the NOD mouse model of type 1 diabetes [11••]. To predict changes in blood glucose levels, cellular immune responses in the endocrine pancreas and a secondary lymphoid organ and the trafficking of cells between these two locations were modeled. In addition to these tissue-level and organism-level phenomena, cytokines and cellular decision-making processes were also represented, as these are potential points of therapeutic intervention. The model was used to evaluate alternative strategies to induce tolerance to insulin, such as the optimal dose, frequency of administration, or stage of disease progression [12,13]. Representing drug pharmacology adds an additional layer of complexity as drugs exhibit multi-organ dynamics that are important for their clinical performance and that can vary significantly from patient to patient. In contrast, systems biology studies tend to be motivated by a desire for deep understanding of a biological network. Mathematics is used to integrate horizontally data and knowledge focused at a particular scale of biological organization. For instance, Covert and coworkers developed a model of the obligate intracellular pathogen *Mycoplasma genitalium* [14••]. To capture an archtypical cell between cell division events, 28 different cellular processes were modeled including metabolism, transcriptional regulation and repair of DNA, synthesis and processing of RNA, and posttranslational modification and macromolecular assembly of proteins. The model was then used to relate genotype to cellular phenotype. While one could envision a future where these two approaches lead to similar models that link genotype to clinical read-outs, these two different approaches are tailored to achieve the research aims

given current limitations in biological knowledge and experimental methods.

While systems biology may receive more attention from academic circles, mechanistic modeling and simulation represents 'dark matter' within the pharmaceutical industry, given the financial incentives for keeping competitive advantages secret. In recent years, many of the major pharmaceutical companies have created QSP teams. Yet, there are few tangible case studies illustrating how mechanistic modeling and simulation financially impact the drug discovery process. Related examples include developing monoclonal antibodies against ErbB3 for treating cancers addicted to Epidermal Growth Factor signaling [15,16]. This target is interesting as an OGRM approach would not have selected ErbB3 as a target, since it is a catalytically inactive kinase. By focusing on the network, ErbB3 was found to be an obligate dimerization partner to other ErbB receptors targeted by trastuzumab (i.e. ErbB2) or lapatinib (i.e. ErbB1/ErbB2) [17•]. In assessing toxicity, QSP approaches have yielded better classifiers for arrhythmia risk [18•] and predictive models for hepatic injury [19].

Biological systems present unique challenges in contrast to more traditional chemical processes

The conceptual toolkit developed to analyze and design chemical systems provides a rich framework to aid in understanding how biological networks function in health and disease. The focus on how cells process information is a natural point of focus as the transfer of information with a cell involves the reactive conversion and intracellular transport of a number of biochemical species, as described in Box 1. Moreover, decades of detailed biochemical and molecular biology studies have identified the major components of these intracellular signaling networks. This prior information can significantly reduce the set of possible networks that can explain observed behavior. For instance, candidate networks can be obtained by mining the published literature [20,21]. However, this prior information can also significantly bias how we interpret experimental observations, given the limited observability of biological systems.

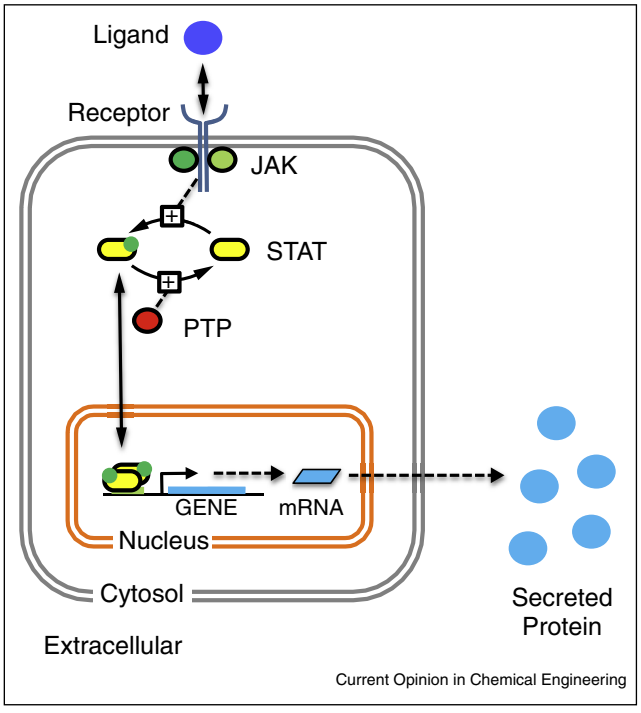
Experimental observability is probably one of the biggest differences between traditional chemical processes and biological systems. Cellular decision making is regulated by the spatial localization and posttranslational modification of proteins within a cell. Given technical limits in detecting small numbers of proteins, many assays measure cellular decision making indirectly or using significant assumptions. Given that nucleotides can be easily amplified using PCR, changes in gene expression are used as a surrogate measure of nuclear localization of a signaling protein. Assuming that a large population of cells (e.g. 1×10^6 cells) behave identically, protein abundance and posttranslational modifications can be quantified by Western blot [22]. In practice, these experimental limitations imply that multiple approaches must be used and the resulting data should provide a self-consistent picture of cellular decision making. As described in the next section, math models aid in testing whether these data are self-consistent and consistent with what we currently know about the biology. For example, single-cell biochemical and imaging measurements acquired over time and in different cell types in

Box 1 Translating cartoons to mechanistic models: A case study of the JAK–STAT pathway

In contrast to more traditional chemical systems, the analysis of reaction networks associated with biological systems is characterized by lumped reactions and poor observability. To illustrate this point, consider the canonical Janus kinase (JAK)-signal transducer and activator of transcription (STAT) signaling pathway, which is highly conserved across eukaryotic organisms. As summarized in Figure 2, this compact signaling pathway transmits extracellular polypeptide signals, through transmembrane receptors, to control gene expression in a variety of fundamental cellular processes, including innate and adaptive immunity [24], regulation of cell growth and apoptosis [25], and control of embryonic stem cell self-renewal [26].

While cartoons like Figure 3 summarize the flow of molecular information down particular pathways within cells, identifying the network associated with a specific JAK–STAT signaling pathway within a specific cell type is more challenging. To illustrate this point, we will consider how Interleukin-12 (IL12) activates a JAK–STAT signaling pathway in type 1 CD4+ T helper cells [27,28,29•]. In the case of IL12, the integrated use of modeling, simulation, and experimentation identified that multiple STAT proteins (STAT4 and STAT1) can become activated in response to receptor ligation and the activity of these STAT proteins are differentially regulated through uncharacterized negative feedback mechanisms, as summarized in Figure 3. The relationships between the abundance of activated STAT4 in the nucleus and de novo protein production and release are unique for each secreted protein, such as Interferon-γ (IFNG) and IL10 [29•]. In addition, IL12 stimulation also enhances cell survival, which suggests that receptor ligation activates additional signaling pathways like the phosphoinositide-3-kinase-protein kinase B/Akt (PI3K-PKB/Akt) pathway [30]. Given the emerging complexity of the IL12 signaling pathway, experimental observability of this pathway is a challenge. For instance, flow cytometry can be used to obtain multiplex single-cell measures of protein phosphorylation, protein copy numbers, and mRNA abundance. Assuming that bench skills and reagents are up to the task, an experiment to monitor intracellular signaling activation and a functional response as a function of time in a cell line can cost $10,000 (we are assuming an experimental design involving a negative control and cells stimulated with a single concentration of ligand that is observed at the following time points: 0, 15 min, 30 min, 1 hours, 2 hours, 4 hours, 8 hours, and 16 hours. Flow cytometric measurements include copy numbers of a heterogeneous receptor (IL12RB1 and IL12RB2), viability, phosphorylation of signaling proteins (STAT1, STAT4, and AKT), abundance of inducible negative regulators of cytokine signaling (SOCS1 and SOCS3), and abundance of three mRNAs using single-cell RNA Fluorescence-in situ hybridization (FISH). Experiments would be performed twice with three replicates per independent trial.) Inevitably, limitations in bench skills, reagents, or financial resources impose a suboptimal experimental design. The mathematical models then aid in interpreting the acquired data in light of our current understanding of how a cell interprets IL12 to orchestrate a response.

Figure 2



A cartoon illustrating the canonical JAK–STAT signaling pathway that initiates a cellular response, as depicted by secreted protein production, in response to stimulation with an extracellular ligand. An extracellular soluble cue binds to a multi-protein complex that is comprised of transmembrane receptor proteins and associated Janus kinases. Upon ligand binding, the receptor changes conformation enabling the JAKs to phosphorylate STAT binding sites within the cytoplasmic tails of the receptors. The STAT proteins then associate with the activated receptor complex and subsequently become phosphorylated by the JAKs, as indicated by the green dot. The phosphorylated STAT proteins dimerize and migrate to the nucleus to initiate the transcription and translation of the corresponding STAT-responsive genes, which include cytokines that are released by the cell to help coordinate cellular response. STAT proteins become deactivated following dephosphorylation by a number of different phosphatases, including protein tyrosine phosphatases (PTP) that are present within the cell.

response to different perturbations were used to identify the dynamic regulation of adhesive contacts between adjacent cells [23•].

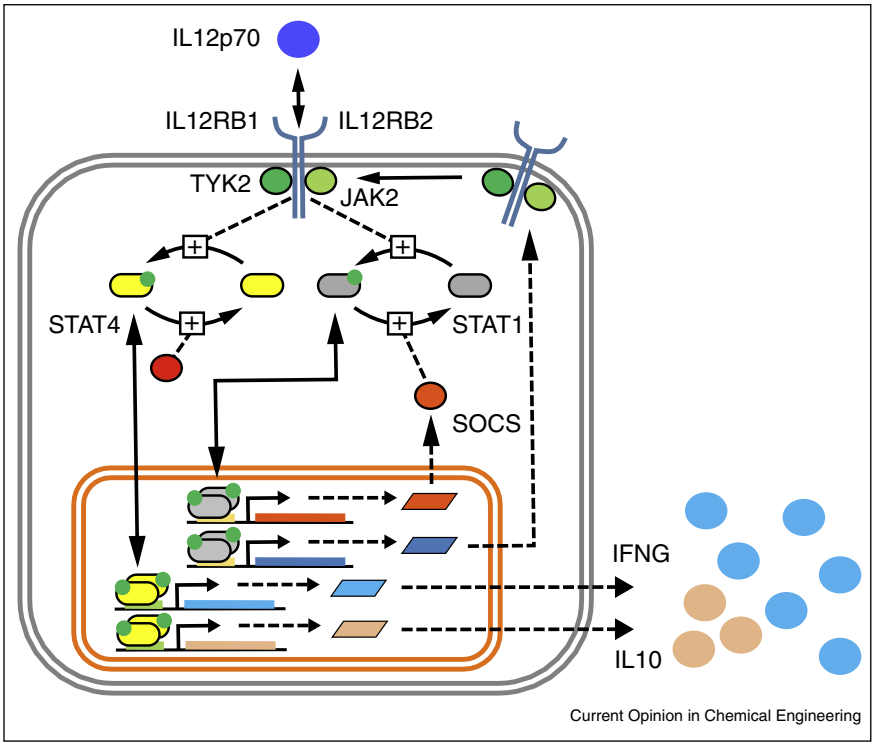
In addition, biological systems contain a number of complex biological processes that confound identifying the underlying physical and reaction processes from experimental observations of cellular behavior. Cellular responses to extracellular cues are governed by a variety of biological processes that can influence protein structure, protein abundance, the functional response to signaling protein activation, and other contextual cues present within the local microenvironment of the cell (Figure 4). Each of these biological processes also has associated time scales in which a change can be observed in response to a perturbation. Introducing a time delay between an experimental perturbation and assaying a cellular response implies that many of these different biological processes can become involved in influencing the cellular response. The challenge in improving our understanding of biological systems comes from deconvoluting the contributions of these different biological processes, which remains a pressing problem in identifying the networks that regulate cellular responses [31].

Bayesian statistics and Markov chain Monte Carlo methods are reshaping how we approach inverse problems in reactive networks where prior knowledge exists

In part, these two types of inverse problems reflect a distinction between concepts associated with statistics and causality [32]. Statistical concepts are applied to quantify uncertainty, which is captured through the use of distributions [33]. In contrast, causality concepts are used to identify how observable events are structured into independent and dependent variables, which implies that the value of one variable is conditioned on the value of another variable. Causality can be depicted as a directed graph, where the variables comprise the vertices, such as the ligand and receptor in Box 1, and causal relationships are depicted as directed edges, such as an arrow that indicates that a ligand binds to the receptor to form a multi-protein complex. A directed edge is a generalization of a reactive event that can range from elementary steps to lumped reactions, which is more common in biology. The key idea drawn from chemical kinetics is that causality among reactive species is determined based on how the variables dynamically respond to perturbations [34]. For instance, observed species can be rank ordered into primary, that is those species impacted directly by the perturbation, and secondary, that is those species impacted indirectly by the perturbation through intermediates, based on their kinetic responses [35–37]. Given the uncertainty associated with experimental observations, the statistical and causal concepts are integrated through the use of conditional probability: $P(Y|X, M)$, which is the probability of observing the value of a dependent variable (Y), given the value of the independent variable (X) and the causal model (M) that captures our understanding of the relationship between X and Y.

In silico model-based inference is an emerging approach that can be applied to inverse problems especially where prior knowledge exists [38], as summarized in Figure 5. To test the implications of the causal structure of a network model, the posterior distributions in the model predictions ($P(\hat{Y}|M)$) need to be established, which also depend on the available data (Y) and the uncertainty in the model parameters ($P(\Theta|M)$). To account for these

Figure 3



The emerging JAK–STAT pathways associated with IL12 signaling in type 1 T helper cells. IL12 binds to a multi-protein receptor complex comprised of two transmembrane receptor proteins, IL12RB1 and IL12RB2 that are bound to the Janus kinases TYK2 and JAK2, respectively. Binding of the receptor complex by IL12 results in the phosphorylation of both STAT1 and STAT4. STAT1 plays a role in the expression of IL12RB2 while STAT4 promotes the transcription and translation of Interferon-γ (IFNG) and Interleukin-10 (IL10). While initially activated by IL12, STAT1 becomes dephosphorylated through an unclear mechanism that may involve the inducible expression of a phosphatase, like a Suppressor of Cytokine Signaling (SOCS). The intracellular transport of the receptor complex (i.e. receptor trafficking), dilution of proteins due to cell proliferation, and signaling pathways that connect IL12 stimulation to enhanced cell viability are some of the biological processes not depicted in this diagram.

confounding influences, we can formulate the problem as an integral:

$$P(\hat{Y}|M) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P(\hat{Y}|\Theta, M) \cdot P(\Theta|M, Y) \cdot P \times (Y) d\Theta dY. \quad (1)$$

As $P(\Theta|M, Y)$ is difficult to obtain directly, Bayes theorem is used to re-express this in terms of quantities that we can calculate,

$$P(\Theta|M, Y) \cdot P(Y) = P(Y|\Theta, M) \cdot P(\Theta|M), \quad (2)$$

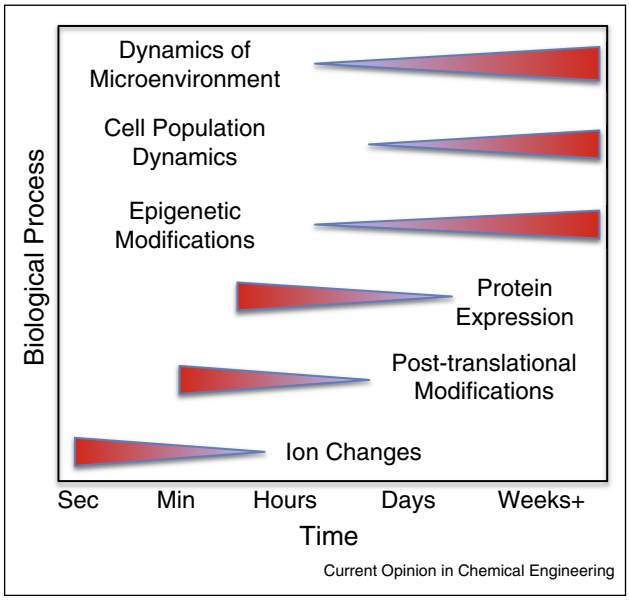
to give:

$$P(\hat{Y}|M) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P(\hat{Y}|\Theta, M) \cdot P(Y|\Theta, M) \cdot P \times (\Theta|M) d\Theta dY. \quad (3)$$

In Eqn 3, $P(\Theta|M)$ is the probability of sampling a point (Θ_i) in parameter space Θ before any knowledge about data Y (i.e. the prior for Θ) and $P(Y|\Theta_i, M)$ is the conditional probability of observing data similar to the simulated response Y when Θ_i and M are given (i.e. the likelihood of Y, given Θ_i and M). Generally, $P(\hat{Y}|\Theta, M)$ represents how the modeled variables will evolve in time, based on a set of parameter values and a mathematical model. In the case of a deterministic model, this conditional probability collapses down to a single path that describe how the variables evolve in time for a single set of parameter values.

To solve for $P(\hat{Y}|M)$, integration over the finite discrete set of experimental observations (Y) is equivalent to a sum over the comparisons between each model prediction, $\hat{Y}$, and the corresponding observation, Y, as represented by the likelihood term: $P(Y|\Theta, M)$. Integrating with respect to the parameters, Θ, is more difficult. Exponential leaps in computational power have enabled Markov Chain Monte Carlo (MCMC) methods to integrate Eqn 3. Given the explosion of MCMC methods in general, resources for MCMC integration are abundant, including reference texts [39,40], stand-alone software

Figure 4



Besides the physical and reactive processes typically associated with chemical systems, a variety of additional biological processes influence how cells respond to extracellular cues. A sampling of biological processes that orchestrate the cellular responses to an extracellular cue (ligand) are shown as a function of their associated kinetic time scales. With the fastest time scales, ion changes in the cytosol and post-translational modifications influence protein structure. Changes in protein structure alter the affinity of protein interactions. Protein abundance can be altered due to signaling-induced changes in abundance via de novo synthesis, degradation, or dilution within an expanding cell population. Epigenetic changes influence the relationship between transcription factor activation and the resulting protein synthesis by altering promoter binding sites, mRNA stability, or translation. Finally, cellular response can be modified by additional secreted or metabolic cues present within the local microenvironment that can, in turn, be shaped by the cell populations themselves.

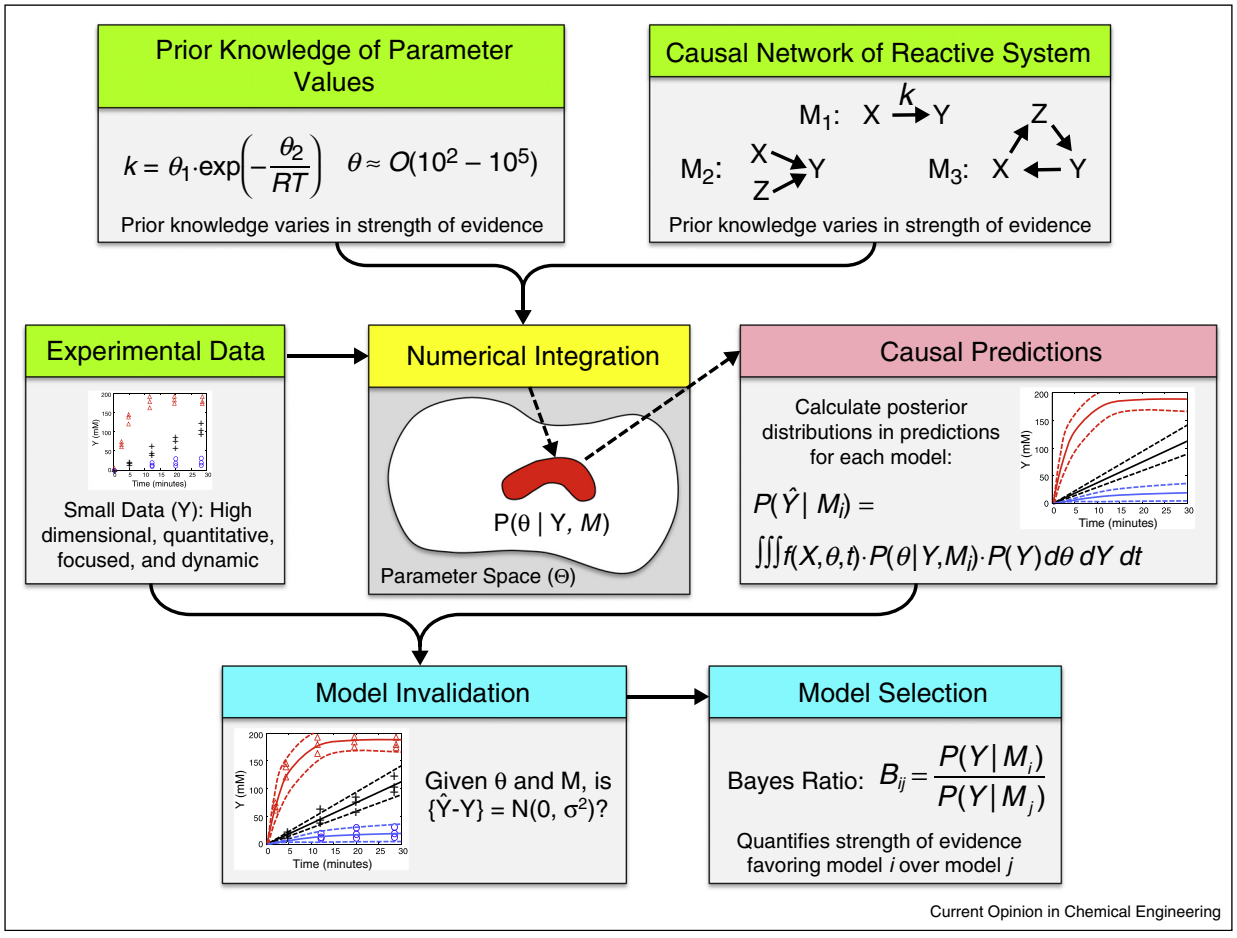
[41–45], and R packages [46]. While MCMC algorithms are relatively simple to program, they can be a challenge to implement correctly such that the results provide an estimate of $P(\hat{Y}|M)$. For instance, the criteria to assess convergence of the Markov Chains should be applied to the model predictions and not the model parameters due to the presence of kinetic slaving in dynamic systems [47,28]. Kinetic slaving means that the overall speed of information transfer within a network is slaved to the slowest step within the network. Other steps in the network are either near equilibrium or are kinetically unimportant.

Applications to network inference

To achieve meaningful learning, one must first identify and address misconceptions that are specific to a scientific domain (i.e. prior knowledge) [48]. In the context of network inference in biology, a discrepancy between our model and an observation identifies flaws in our current understanding of the modeled system. Historically, the models most frequently used are mental models, which are communicated as cartoons (see Box 1). However, biological systems, like cell signaling networks, exhibit characteristics that make it extremely difficult to test our mental models, namely embedded dynamics, feedback regulation, or competing pathways. Dynamic mathematical models, like systems of coupled ordinary differential equations, provide a quantitative framework for encoding our causal knowledge about systems [32]. Predictions derived from the models can be used to test our models against data. Historically, these predictions are a single dynamic trajectory that is dependent on the values of the underlying model parameters. However, uncertainty in the model predictions is convoluted with the uncertainty of the model parameters. Thus, it is impossible to make confident statements about model inadequacy, given our ignorance of the underlying parameter values and the biology (e.g. protein–protein interaction energies or protein abundance) that they represent.

In a network inference context, in silico model-based inference methods can be used to encode competing hypotheses regarding the causal network and then generate posterior distributions in the model predictions by statistically sampling over all parameter values that give predictions that are consistent with the observed data. As the error between a model prediction and observed data should be normally distributed with a zero mean value, competing hypotheses can be rejected if they systematically deviate from the observed experimental data. Competing hypotheses that pass model invalidation criteria can then be evaluated using a Bayes Ratio, which quantifies the strength of evidence that favors one model over another, as illustrated in [49,45]. Alternatively, a number of other model selection criteria, like the Akaike Information Criterion, have been proposed, as reviewed in [50]. These criteria quantify the perceived trade-off between predictive power, as commonly quantified by the summed squared error between the observations and model predictions, and a penalty term associated with model complexity, which can be related to the number of parameters. While these model selection criteria provide a simple metric, there are a number of underlying assumptions in developing these relationships that are not commonly encountered in biology. First, these criteria are applied in the asymptotic limit of empirical clarity, which means that the states of the system are all observed and all of the model parameters can be identified. Second, the penalty terms are ad hoc, which makes the criteria qualitative as a different weighting scheme would select a different winner among similarly scoring models [51]. In biology, the complexity of the model is influenced heavily by prior knowledge of the reactive network and observability is limited. A biologically realistic model inevitably includes many parameters that cannot be identified in

Figure 5



A schematic of an emerging approach for inverse problems that use mechanistic mathematical models of reactive systems. This approach leverages concepts from Bayesian statistics, probability, and advances in computational power to test competing hypotheses regarding causal structure of the reaction network. A mechanistic model of the reactive system, prior knowledge of parameter values, and experimental data are inputs to a computational filter. If prior knowledge of the key processes that govern the behavior of the reactive systems is weak, multiple competing hypotheses could be proposed as a reaction network. If prior knowledge of the parameter values is also weak, the computational filter uses this information as it searches parameter space to select a statistically based ensemble of parameter values (red region of parameter space), given the uncertainty in the experimental data and model. This ensemble of parameter values are then used to generate a corresponding ensemble of predictions that describe probabilistically how the system evolves in time from the initial values, given the specific data and network model. A model invalidation step involves testing whether the difference between the model predictions ($\hat{Y}$) and experimental data (Y) do not have systematic differences (i.e. $\{\hat{Y} - Y\} \neq N(0, \sigma^2)$). Finally, a Bayes Ratio can be used to select among competing hypothesis as to the governing processes associated with the reactive system.

practice. To improve parameter identifiability, timescale analysis of reactive networks provides a data-based approach to select the appropriate complexity [28].

As all models are abstractions of reality, the value of the model ultimately depends on the fitness-of-purpose of the model for aiding inductive/deductive reasoning. The goal, then, is not necessarily to confirm our existing knowledge, but to use mathematical models to capture our cognitive understanding of the system and challenge these models with experimental data to identify flaws in our current understanding, as discussed in [38]. One example of the approach is mentioned in Box 1, where the differential regulation of STAT1 versus STAT4 phosphorylation, an indirect measure of activity, in response to IL12 stimulation was identified after an initial data set was unable to distinguish between competing hypotheses [29•]. Another example focuses on the dynamic regulation of adherens junctions [23•], which maintain the integrity of epithelial tissues through extracellular homotypic bonds. Experimentally, quantitative single-cell and population-level *in vitro* assays were used to quantify the endogenous pathway dynamics following the proteolytic disruption of the adherens junctions.

Using prior knowledge of isolated elements of the overall network, these data were interpreted using *in silico* model-based inference to identify the topology of the regulatory network. While not previously recognized as playing a role in this network, an endocytic recycling pathway was essential to capture the observed data. Collectively, the data suggest that the regulatory network contains interlocked network motifs consisting of a positive feedback loop, which is used to restore the integrity of adherens junctions, and a negative feedback loop, which is used to limit beta-catenin-induced gene expression.

Applications to parameter inference

The parameters of a model can include rate parameters that quantify the propensity of a reaction to proceed, initial values of the model, and nuisance parameters that are required to map a mathematical model onto experimentally observed values. By integrating Eqn 3 using a MCMC approach, the predictions contained within the converged segments of the Markov Chains represent samples from posterior distribution in the model predictions, $P(\hat{Y}|M)$. In addition, the corresponding parameter values from the converged segments of the Markov Chains represent samples from the posterior distributions in the parameters ($P(\Theta|Y, M)$):

$$P(\Theta|M, Y) = \frac{P(Y|\Theta, M) \cdot P(\Theta|M)}{P(Y)}, \quad (4)$$

where $P(\Theta|M)$ is the prior for the model parameters and $P(Y)$ serves effectively as a normalization constant.

In implementing a MCMC approach for parameter inference using mechanistic models of dynamic systems, there are a couple of points to keep in mind. First, the parameter priors should have heavy tails, meaning that prior probability for a particular set of parameter values ($P(\Theta|M)$) should be greater than the posterior probability. In traditional chemical kinetic systems, the priors for parameters may be well defined in thermodynamic terms and calculated using *ab initio* methods (e.g. [52]). In models where multiple kinetic processes are grouped together as a lumped reaction, the prior distribution may be broad as possible values of the parameters are known only within a couple of orders of magnitude. Second, the data should 'swamp' the prior, which means that the posterior distribution should reflect the combination of the data and the model used to interpret the data rather than an arbitrary choice of prior distributions for the parameters. Although in practice, this is difficult to diagnose, which leads to the last point. Finally, the structure of mechanistic models of biological systems is created based on the prior knowledge of the key variables involved in a system. In creating a more realistic model of the biology, additional parameters become incorporated into the model. However, the available data may only constrain a subset of the parameters through a process described as kinetic slaving.

In kinetic slaving, parameters associated with reactions that are fast, such as pre-formed multi-protein complexes, or that are kinetically unimportant, such as stationary reactions, exhibit one-sided distributions [28]. This means that parameters associated with fast (or stationary) reactions are only constrained such that the value has to provide a time scale that is faster (or slower) than the observable time scales associated with the rate-limiting steps. This is a subtle but important point as many studies that apply statistical inference methods to inverse problems related to dynamic biological networks provide distributions in the model parameters that have bounds all supposedly informed by data (i.e. a posterior distribution) but that are in the form of a multivariate Gaussian distribution (i.e. all the parameters are bounded). Distributions in parameters of dynamic system models that are inconsistent with kinetic slaving is an indication that distributions do not reflect the experimental data but are constrained by the prior or that the model has been overly simplified to achieve an empirical fit.

Conclusion

In understanding traditional chemical processes and in engineering better medicines, making mechanistic models of these reactive systems involve similar challenges, which involve establishing causal relationships among reactive components of a system based on their observed dynamics. Moreover our existing knowledge of these systems is incomplete, as we have some prior knowledge of the likely reactive components and how they influence system response, and our ability to observe the system is limited. Nonetheless, these mechanistic models aim to assist our natural intuition and facilitate communication by providing a concrete realization of how we think the system works. While mathematics and simulation play central roles in addressing these problems, advances in computational power have enabled more sophisticated methods that provide a solid statistical and probabilistic foundation to test our mechanistic understanding against data.

Here, we provide an overview of some of the emerging model-based inference tools that leverage Markov Chain Monte Carlo methods enabled by increases in computational power and concepts drawn from Bayesian statistics, probability, and chemical kinetics. We have focused on applications related to cellular signal transduction networks, which transmit molecular information within a cell to regulate a functional response. Given the importance of understanding cellular behavior to engineer better medicines, cellular systems are particularly challenging as limited observability of these intracellular networks is more acute and potential reactive components of these systems are being discovered at an increasing pace. Computational approaches built upon solid statistical and causal foundations, as described here, will be increasingly used to help think more clearly about the dynamic

relationships among the components of reactive networks.

Conflict of interest

The authors declare that they have no financial/commercial conflicts of interest.

Acknowledgements

This work was supported by grants from the National Science Foundation (NSF) CAREER 1053490 (DJK) and the National Institutes of Health R15CA123123 (DJK), P50GM071558 (MRB), R01GM104184 (MRB), and U54HG008098 (MRB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or the National Institutes of Health.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ramkrishna D, Amundson N: **Mathematics in chemical engineering**. *AIChE J* 2004, **50**:7-23.

2. Kevrekidis IG: **Matrices are forever: on applied mathematics and computing in chemical engineering**. *Chem Eng Sci* 1995, **50**:4005-4025.

3. National Academy of Engineering: *Grand Challenges for Engineering*. Washington, DC: National Academies Press; 2008 http://www.engineeringchallenges.org/.

4. Gilbert J, Henske P, Ashish S: **Rebuilding big pharm.'s business model**. *In Vivo Bus Med Rep* 2003, **21**:1-10.

5. Ocana A, Pandiella A, Siu LL, Tannock IF: **Preclinical development of molecular-targeted agents for cancer**. *Nat Rev Clin Oncol* 2011, **8**:200-209.

6. Birtwistle MR, Mager DE, Gallo JM: **Mechanistic vs. empirical network models of drug action**. *CPT Pharmacomet Syst Pharmacol* 2013, **2**:e72.

7. •• Sorger PK, Allerheiligen SRB, Abernethy DR, Altman RB, Brouwer KLR, Califano A *et al.*: *Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. An NIH white paper by the QSP Workshop Group*. 2011, October http://www.nigms.nih.gov/training/documents/systemspharmawpsorger2011.pdf. (accessed 15.12.14).
This white paper provides one of the first comprehensive definition of quantitative and systems pharmacology as a discipline and summarizes what grand challenges the field is poised to tackle.

8. Walter RB, Appelbaum FR, Tallman MS, Weiss NS, Larson RA, Estey EH: **Shortcomings in the clinical evaluation of new drugs: acute myeloid leukemia as paradigm**. *Blood* 2010, **116**:2420-2428.

9. Vicini P, van der Graaf PH: **Systems pharmacology for drug discovery and development: paradigm shift or flash in the pan?** *Clin Pharmacol Ther* 2013, **93**:379-381.

10. Gossage L, Eisen T: **Targeting multiple kinase pathways: a change in paradigm**. *Clin Cancer Res* 2010, **16**:1973-1978.

11. •• Shoda L, Kreuwel H, Gadkar K, Zheng Y, Whiting C, Atkinson M, Bluestone J, Mathis D, Young D, Ramanujan S: **The Type 1 Diabetes PhysioLab Platform: a validated physiologically based mathematical model of pathogenesis in the non-obese diabetic mouse**. *Clin Exp Immunol* 2010, **161**:250-267.
As one of the first published examples, this paper describes a large multi-scale mechanistic model that illustrates a quantitative and systems pharmacology approach. This mechanistic model of the NOD mouse, a mouse model for type 1 diabetes, was used to evaluate different targets and clinical trial protocols *in silico*. The model can be explored using proprietary software that runs on Windows XP.

12. Fousteri G, Chan JR, Zheng Y, Whiting C, Dave A, Bresson D, Croft M, von Herrath M: **Virtual optimization of nasal insulin therapy predicts immunization frequency to be crucial for diabetes protection**. *Diabetes* 2010, **59**:3148-3158.

13. Mamchak AA, Manenkova Y, Leconet W, Zheng Y, Chan JR, Stokes CL, Shoda LK, von Herrath M, Bresson D: **Preexisting autoantibodies predict efficacy of oral insulin to cure autoimmune diabetes in combination with anti-CD3**. *Diabetes* 2012, **61**:1490-1499.

14. •• Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**:389-401.
This was the first paper that incorporated arguably every known biochemical process in a single bacterium into a single integrated computational model. This systems biology model captures a range of experimentally observed phenotypes.

15. Schoeberl B, Faber AC, Li D, Liang M, Crosby K, Onsum M, Burenkova O, Pace E, Walton Z, Nie L, Fulgham A, Song Y, Nielsen UB, Engelman JA, Wong K: **An ErbB3 antibody MM-121 is active in cancers with ligand-dependent activation**. *Cancer Res* 2010, **70**:2485-2494.

16. McDonagh CF, Huhalov A, Harms BD, Adams S, Paragas V, Oyama S, Zhang B, Luus L, Overland R, Nguyen S, Gu J, Kohli N, Wallace M, Feldhaus MJ, Kudla AJ, Schoeberl B, Nielsen UB: **Antitumor activity of a novel bispecific antibody that targets the ErbB2/ErbB3 oncogenic unit and inhibits heregulin-induced activation of ErbB3**. *Mol Cancer Ther* 2012, **11**:582-593.

17. • Schoeberl B, Pace EA, Fitzgerald JB, Harms BD, Xu L, Nie L, Linggi B, Kalra A, Paragas V, Bukhalid R, Grantcharova V, Kohli N, West KA, Leszczyniecka M, Feldhaus MJ, Kudla AJ, Nielsen UB: **Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis**. *Sci Signal* 2009, **2**:ra31.
One of the first published applications of systems pharmacology modeling that led to the development of a clinically tested therapeutic. As a drug target, ErbB3 is arguably a poor choice under a one drug, one target paradigm, since it is an obligate dimerization partner and has no kinase activity on its own.

18. • Cummins MA, Dalal PJ, Bugana M, Severi S, Sobie EA: **Comprehensive analyses of ventricular myocyte models identify targets exhibiting favorable rate dependence**. *PLoS Comput Biol* 2014, **10**:e100354.
A pioneering study using systems pharmacology methods to predict toxicity and focuses on arrhythmia risk. By integrating a systems view of cardiac action potentials with poly-pharmacology drug action, they were able to improve prediction performance above the state of the art for arrhythmia risk (e.g. univariate hERG channel measurements).

19. Howell BA, Yang Y, Kumar R, Woodhead JL, Harrill AH, Clewell HJ, Andersen ME, Siler SQ, Watkins PB: **In vitro to in vivo extrapolation and species response comparisons for drug-induced liver injury (DILI) using DILIsym (TM): a mechanistic mathematical model of DILI**. *J Pharmacokinet Pharmacodyn* 2012, **39**:527-541.

20. D'Alessandro LA, Samaga R, Maiwald T, Rho SH, Bonefas S, Raue A, Iwamoto N, Kienast A, Waldow K, Meyer R, Schilling M, Timmer J, Klamt S, Klingmuller U: **Disentangling the complexity of HGF signaling by combining qualitative and quantitative modeling**. *PLoS Comput Biol* 2015, **11**:e1004192.

21. Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK: **Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks**. *BMC Syst Biol* 2012, **6**:29.

22. Janes KA: **An analysis of critical factors for quantitative immunoblotting**. *Sci Signal* 2015, **8**:rs2.

23. • Klinke D, Horvath N, Cuppett V, Wu Y, Deng W, Kanj R: **Interlocked positive and negative feedback network motifs regulate beta-catenin activity in the adherens junction pathway**. *Mol Biol Cell* 2015. (in press).
See annotation to Ref [29•].

24. O'Shea JJ, Plenge R: **JAK and ST, signaling molecules in immunoregulation and immune-mediated disease**. *Immunity* 2012, **36**:542-550.

25. Fagard R, Metelev V, Souissi I, Baran-Marszak F: **STAT3 inhibitors for cancer therapy: have all roads been explored?** *JAKSTAT* 2013, **2**:e22882.

26. Tang Y, Tian XC: **JAK–STAT3 and somatic cell reprogramming**. *JAKSTAT* 2013, **2**:e24935.

27. Finley SD, Gupta D, Cheng N, Klinke DJ: **Inferring relevant control mechanisms for Interleukin-12 signaling within naive CD4+ T cells**. *Immunol Cell Biol* 2011, **89**:100-110.

28. Klinke DJ, Finley SD: **Timescale analysis of rule-based biochemical reaction networks**. *Biotechnol Prog* 2012, **28**:33-44.

29. • Klinke DJ, Cheng N, Chambers E: **Quantifying crosstalk among interferon-gamma, interleukin-12 and tumor necrosis factor signaling pathways within a Th1 cell model**. *Sci Signal* 2012, **5**:ra32.

This reference along with [23•] illustrates an approach to network inference that combines quantitative and dynamic experimental measurements with mechanistic mathematical modeling to test whether our current understanding of the corresponding signaling networks is consistent with the observed data. In each of these two papers, this integrated approach uncovered previously unappreciated aspects of the corresponding signaling networks that regulate cellular responses and serves to illustrate how mathematical modeling can refine our understanding of signaling pathways.

30. Hemmings BA, Restuccia DF: **The PI3K-PKB/Akt pathway**. *Cold Spring Harb Perspect Biol* 2012, **4**:a011189.

31. Purvis JE, Lahav G: **Encoding and decoding cellular information through signaling dynamics**. *Cell* 2013, **152**:945-956.

32. Pearl J: *Causality: models, reasoning, and inference*. edn 2. Cambridge, UK: Cambridge University Press; 2009.

33. Jaynes E: *Probability Theory: The Logic of science*. edn 1. Cambridge, UK: Cambridge University Press; 2003.

34. Haken H: *Synergetics*. Berlin: Springer-Verlag; 2004.

35. Bhore N, Klein M, Bischoff K: **The Delplot technique – a new method for reaction pathway analysis**. *Ind Eng Chem Res* 1990, **29**:313-316.

36. Sontag E, Kiyatkin A, Kholodenko BN: **Inferring dynamic architecture of cellular networks using time series of gene expression protein and metabolite data**. *Bioinformatics* 2004, **20**:1877-1886.

37. Klein M, Hou Z, Bennett C: **Reaction network elucidation: interpreting Delplots for mixed generation products**. *Energy Fuels* 2012, **26**:52-54.

38. Klinke DJ: **In silico model-based inference: a contemporary approach for hypothesis testing in network biology**. *Biotechnol Prog* 2014, **30**:1247-1261.

39. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis, Texts in Statistical Science*. Boca Raton, FL: Chapman and Hall; 2004.

40. Brooks S, Gelman A, Jones G, Meng X (Eds): *Handbook of Markov Chain Monte Carlo, Handbooks of Modern Statistical Methods*. Boca Raton, FL: Chapman and Hall/CRC; 2011.

41. Lunn D, Spiegelhalter D, Thomas A, Best N: **The BUGS project: evolution critique and future directions (with discussion)**. *Stat Med* 2009, **28**:3049-3082.

42. Vyshemirsky V, Girolami M: **BioBayes: a software package for Bayesian inference in systems biology**. *Bioinformatics* 2008, **24**:1933-1934.

43. Hug S, Raue A, Hasenauer J, Bachmann J, Klingmuller U, Timmer J, Theis FJ: **High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling**. *Math Biosci* 2013, **246**:293-304.

44. Chen Y, Lawless C, Gillespie CS, Wu J, Boys RJ, Wilkinson DJ: **CaliBayes and BASIS: integrated tools for the calibration simulation and storage of biological simulation models**. *Brief Bioinform* 2010, **11**:278-289.

45. Eydgahi H, Chen WW, Muhlich JL, Vitkup D, Tsitsiklis JN, Sorger PK: **Properties of cell death models calibrated and compared using Bayesian approaches**. *Mol Syst Biol* 2013, **9**:644.

46. Geyer C, Johnson L: *R Package MCMC (Markov Chain Monte Carlo), Tech. rep.*. 2012 http://cran.r-project.org/package=mcmc.

47. Klinke DJ: **An empirical Bayesian approach for model-based inference of cellular signaling networks**. *BMC Bioinform* 2009, **10**:371.

48. National Research Council (U.S.): *Committee on Learning, Research, Practice and Education, How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academies Press; 2000.

49. Xu TR, Vyshemirsky V, Gormand A, von Kriegsheim A, Girolami M, Baillie GS, Ketley D, Dunlop AJ, Milligan G, Houslay MD, Kolch W: **Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species**. *Sci Signal* 2010, **3**:ra20.

50. Kirk P, Thorne T, Stumpf MP: **Model selection in systems and synthetic biology**. *Curr Opin Biotechnol* 2013, **24**:767-774.

51. Kass RE, Raftery AE: **Bayes factors**. *J Am Stat Assoc* 1995, **90**:773-795.

52. Klinke DJ, Broadbelt LJ: **Construction of a mechanistic model of Fischer–Tropsch synthesis on Ni(1 1 1) and Co(0 0 0 1) surfaces**. *Chem Eng Sci* 1999, **54**:3379-3389.