# Physicochemical modelling of cell signalling pathways

**Bree B. Aldridge, John M. Burke, Douglas A. Lauffenburger and Peter K. Sorger**

**Physicochemical modelling of signal transduction links fundamental chemical and physical principles, prior knowledge about regulatory pathways, and experimental data of various types to create powerful tools for formalizing and extending traditional molecular and cellular biology.**

This review is aimed at biologists interested in mathematical modelling of biochemical pathways, but who are relatively unfamiliar with the topic. Our discussion focuses on pathways involving 'signals' rather than metabolites. In this context, physicochemical modelling is a natural extension of informal or conceptual pathway modelling. Formal modelling is much more powerful in putting molecular detail in a physiological context, uncovering principles of biological design and creating dynamic repositories of interpretable knowledge. However, to realize this power, challenges inherent in construction, verification, calibration, interpretation and publication of models must be addressed.

## MATHEMATICAL MODELS IN MOLECULAR, CELLULAR AND DEVELOPMENTAL BIOLOGY

Contemporary molecular, cellular and developmental biology seeks to describe physiological processes in terms of gene functions and specific molecular mechanism. Medicine and drug discovery add the practical goals of understanding disease and developing treatments. The 'component identification' phase of modern biology is approaching completion, and the sheer size of the cellular 'parts list' highlights the importance of understanding function, not at the level of single genes, but rather at a higher level of abstraction, involving pathways and circuits. In many cases, conceptual modelling of biology is at the breaking point[1] — it is impossible mentally to juggle large pathways involving many components. The missing ingredient is mathematics. Used appropriately, mathematical models can represent pathways in a physically and biologically realistic manner, incorporate a wide variety of empirical observations, and generate novel and useful hypotheses. Pathway modelling has existed for some time, particularly in the field of prokaryotic metabolism[2,3], but it remains at an early stage of development. It is challenging to construct accurate models and establish rigorous links to experimental data (see accompanying article by Jaqaman *et al.* in *Nature Rev. Mol. Cell Biol.*). This commentary is based on the premise that useful models of critical mammalian pathways can nonetheless be constructed using an iterative modify–measure–mine–model procedure that closely integrates experiment and mathematics (Fig. 1).

Bree B. Aldridge, John M. Burke, Douglas A. Lauffenburger and Peter K. Sorger are in the Center for Cell Decision Processes, Department Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Bree B. Aldridge and Peter K. Sorger are in the Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA.
e-mail: psorger@mit.edu

## APPROACHES TO PHYSICOCHEMICAL MODELLING

Physicochemical modelling seeks to describe biomolecular transformations (such as covalent modification, intermolecular association and intracellular localization) in terms of equations derived from established physical and chemical theory[4–8]. These 'kinetic' or 'reaction' models use prior knowledge to make specific molecular predictions and work best with pathways in which components and connectivity are relatively well established. When prior knowledge is sparse, data-driven statistical models are more appropriate (see accompanying article by Janes *et al.* in *Nature Rev. Mol. Cell Biol.*). Equations in physicochemical models refer to identifiable processes (such as catalysis and assembly) and parameters have physical interpretation (such as concentration, binding affinity, and reaction rate). The models can be viewed as translations of familiar pathway maps into mathematical form — a process that should become easier and more transparent with the adoption of common schematic standards[9].

The correct mathematical form for a physicochemical model depends on the properties of the system being studied and the goals of the modelling effort. Ordinary and partial differential equations (ODEs and PDEs) are most commonly and both can be cast in either deterministic or stochastic form. Stochastic equations include effects arising from random fluctuation around the average behaviour. Currently, the most common means of representing biochemical pathways is through a set of coupled ODEs (an ODE network). ODE networks represent the rates of production and consumption of individual biomolecular species, $d[X_i]/dt$, in terms of mass action kinetics — an empirical law stating that rates of a reaction are proportional to the concentrations of the reacting species. Each biochemical transformation is therefore represented by an elementary reaction with forward and reverse rate constants. Changes in localization, a central feature of biological pathways, are represented by compartmentalization. Each species is allowed to inhabit one or more compartments and to move among the compartments through elementary reactions. Compartments are also used to represent assembly of macromolecular complexes and other non-enzymatic changes of state. Two fundamental assumptions of the compartmentalized ODE

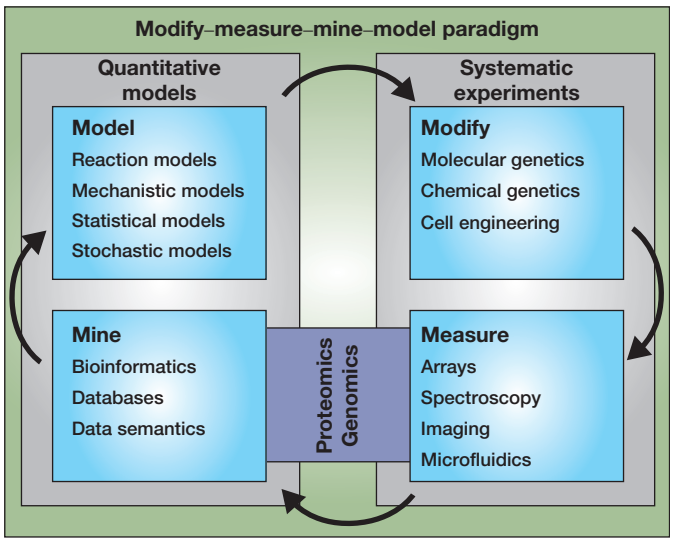


**Figure 1** The modify–measure–mine–model paradigm in systems biology. A paradigm for systems biology research involving iterative cycles of experimental modification, measurement, data mining and mathematical modelling.

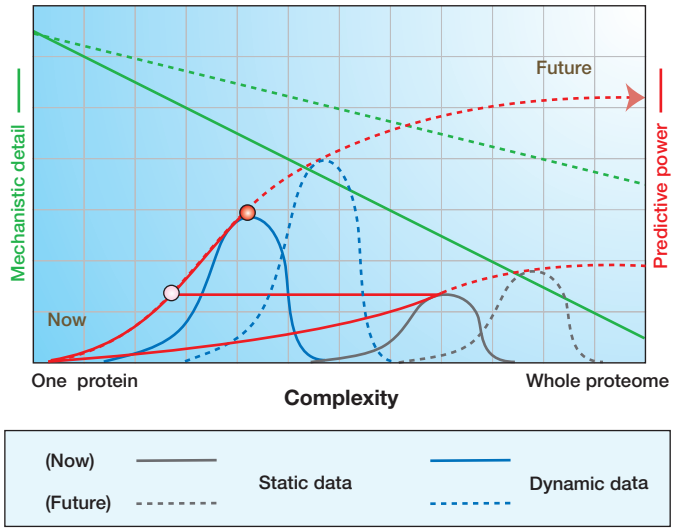


**Figure 2** Physicochemical modelling involves a trade off between increasing scope and falling detail. In this hypothetical representation, the scope of a model (the number of unique gene products) increases from left to right. The degree of mechanistic detail (green line) falls as more components are included, as many gene products are poorly studied. Physicochemical modelling involves a compromise between too narrow a scope and insufficient predictive power, and too wide a scope and overwhelming uncertainty. The result is an optimum (blue and black solid lines) at a scope below that of the complete proteome but that shifts to greater complexity over time (dotted lines) as more molecular data is collected (green dotted line). Overall, the optimal size for modelling will also increase with time (red line). It is interesting to note that the degree of predictive power that can be achieved at any scope depends on the quality of the data. Rich data resolved in time and space (blue lines) is considerably more valuable than single-time point static data (black lines).

formalism are that: first, within a compartment, the concentration of each species is high and transport essentially instantaneous, that is, the compartment is well-mixed; and second, between compartments, transport is slower and associated with an observable rate. If these assumptions are not satisfied, then it is necessary to model changes in species concentrations explicitly with respect to space (typically using PDEs).

Provided that reasonable values for rate constants and initial conditions (concentrations at $t = 0$) can be obtained, time integration of ODEs yields the concentrations of each species at subsequent times, thereby facilitating comparison of simulated and experimental time courses. ODE networks are amenable to a wide variety of analytic techniques, some of which are difficult, if not impossible, with more complex mathematical forms.

A conventional ODE represents a continuum approximation to reactions that actually involve interactions between individual molecules, which is a probabilistic process. The (bio)chemical master equation (CME)[10] shows that random fluctuations in species concentration or reaction rates scale with $(N)^{1/2}$, where $N$ is the number of molecules of a given species in the relevant compartment. Continuum approximations may be considered valid when $N$ is higher than 100–1000, yielding temporal variation of ~ $(N)^{1/2}N^{-1} = (N)^{-1/2}$ or ~3–10%; they prevail in models of cell signalling but are not suitable for describing microtubule or actin polymer dynamics[11–14] (see article by Karsenti *et al.* in this issue). However, considering the importance of macromolecular assembly and the prevalence of membranous and cytoskeletal elements in partitioning cells, even highly abundant proteins can be present in critical compartments at concentrations sufficiently low to produce stochastic behaviour. For example, stochastic fluctuations arise when abundant regulators of gene expression access a very small number of transcriptional initiation sites on DNA[15]. Stochastic effects also arise from very slow elementary reactions, because the generation of product molecules is sufficiently separated in time as to seem discontinuous. Whether stochastic events in one compartment affect overall behaviour of a network depends critically on parameter values and network structure[16]. As a general rule in cell signalling models, it is reasonable to begin with continuum approximations and proceed to more complicated stochastic representations only as required.

## MODEL DESIGN

Two critical decisions in the design of ODE models are specifying the scope and level of detail. Obviously, reaction models can only encompass a small subset of all reactions taking place in cells. If the scope is too small, predictive power is lost; if the scope is too large, the uncertainty is overwhelming (Fig. 2). Thus, assumptions must be made about the extent to which species included in the model evolve independently of species excluded from the model. The issue of model scope is usually cast in terms of modules — subsets of cellular reactions assumed to work together in the execution of discrete biochemical functions[17]. Little exists in the way of rigorous theoretical or empirical evidence for modularity in cell signalling pathways, but it is an assumption implicit in all molecular approaches, not just modelling. Indeed, uncertainty as to the components, connectivity and properties of pathways is a key motivation for undertaking rigorous, quantitative analysis. For the foreseeable future, *ad hoc* assumptions are likely to determine the scope of most models, but as understanding of network architecture increases, we can expect much greater insight into modularity.

A second design decision in physicochemical modelling is the degree of detail (model granularity). As a first principle, and in the absence of countervailing evidence, the best model is the one that is most parsimonious in species and parameters, while meeting the design goals. It is not possible to derive the functional properties of proteins and other biomolecules *ab initio* from their atomic structures — pathway models are not detailed physical representations. Conversely, relatively little molecular insight is gained when molecular processes are 'lumped' together to produce few species and equations. Therefore, most pathway

models operate at a mesoscale that is intermediate between complete microscopic enumeration and broad descriptive representation. Because practical modelling aims to make predictions that can be confirmed experimentally at the levels of genes and proteins, the number of unique gene products is a starting point for specifying model granularity.

Determining the correct granularity for a pathway model is complicated by the fact that proteins assemble into large multi-component complexes, undergo extensive posttranslational modification and partition among multiple cellular compartments. If these processes are to be represented, the number of species in a model can increase dramatically relative to the number of gene products. In the case of the epidermal growth factor receptor (EGFR; a membrane-bound receptor tyrosine kinase that forms dimers with one of three other family members), 276 distinct homodimeric species can arise from three binding partners and two phospho forms[18]. In actuality, EGFR has four transmembrane, eight intracellular binding partners and ten phosphorylation sites. More generally, if interacting molecules A and B can assume $n$ and $m$ distinct states respectively, the total number of states is $2^n \times 2^m$, and of bimolecular reactions is $2^{n-1} \times 2^{m-1}$. This 'combinatorial explosion' argues in favour of models with many species. However, the greater the number of species in a model, the greater the degrees of freedom, and the larger the chance that parameters will be indeterminable. Therefore, models are frequently limited to a subset of possible species. If data support the functional equivalence of distinct biochemical forms, then 'lumping' them together is clearly warranted. Similarly, if assembly or multi-site modification is rapid and processive, then intermediate states can be ignored. However, in many situations, including EGFR signalling, the consequences of ignoring species are unknown and it is hard to establish a rigorous basis for optimizing model granularity[19,20]. Thus, being explicit with respect to design goals and discussing the implications of complete and partial enumeration of species for model performance is important.

Fortunately, even in highly granular models, it is rarely necessary to include detailed representations of every process. Core metabolic and synthetic pathways (such as energy production and gene transcription)

## BOX 1 RESOURCES FOR BUILDING CELL-CIRCUIT MODELS

Building mechanistic models involves a series of steps from model construction to validation. No existing software package can perform all of the steps, making it necessary to use multiple packages. This causes considerable difficulty because software from different sources is rarely inter-operable and methods for transporting models among software packages are incompletely developed. Here, we review the model assembly process and some available software but refer readers to an excellent recent review on the strengths and weaknesses of different packages[42].

### Model construction

Model construction typically involves the translation of prior knowledge into a list of reactants and reactions. Because large bodies of literature are rife with inconsistency and inaccuracy, expert interpretation and annotation are essential, despite the attractions of automatic text mining. One approach to codifying network topology is to manually write a list of differential equations. In many cases, this takes the form of a table of species, parameter values and comments. Manual methods become cumbersome and error prone as the number of equations grows. Graphical approaches simplify the process of transforming a network diagram into a set of linked equations. Most specialized software tools such as CellDesigner, JDesigner and SimBiology provide varying degrees of support for graphical model construction[42].

A problem with both manual and graphical modelling approaches is that every change in granularity, and thus, in the number of equations and species, requires that the reaction list or schematic be redrawn. BioNetGen addresses this problem by constructing models automatically from a set of machine-readable rules.

### Model verification

Model verification is necessary to establish that the equations comprising a model depict a network correctly, as it is understood from the literature (or experiments). Although conceptually straightforward, model construction is error prone: it is difficult to spot errors when mechanisms are complex and many species are involved. Unfortunately, formal tools for verifying models have not yet been developed and current practice involves manual analysis and checking for conservation of species in a list of reactions (using GEPASI, for example). No mechanism exists for certifying systems biology models, or even for verifying their underlying structure, although approaches in this direction have been proposed[43].

### Model calibration

Model calibration, or regression, is the process by which parameter values are estimated so as to achieve the design goals — in our case, as close a fit to data as possible. Jacobian and JSim are examples of software that perform parameter estimation. However, it is challenging experimentally to collect enough high quality data to effectively constrain parameter values. It is also difficult to evaluate the reliability of these constraints. Thus, most biological pathway models remain poorly calibrated.

### Model validation

Model validation is the process of evaluating the ability of a calibrated model to meet post hoc constraints. In systems biology, this includes making predictions that can be subjected to experimental test. Other less easily justified constraints include robustness, bistability and simplicity.

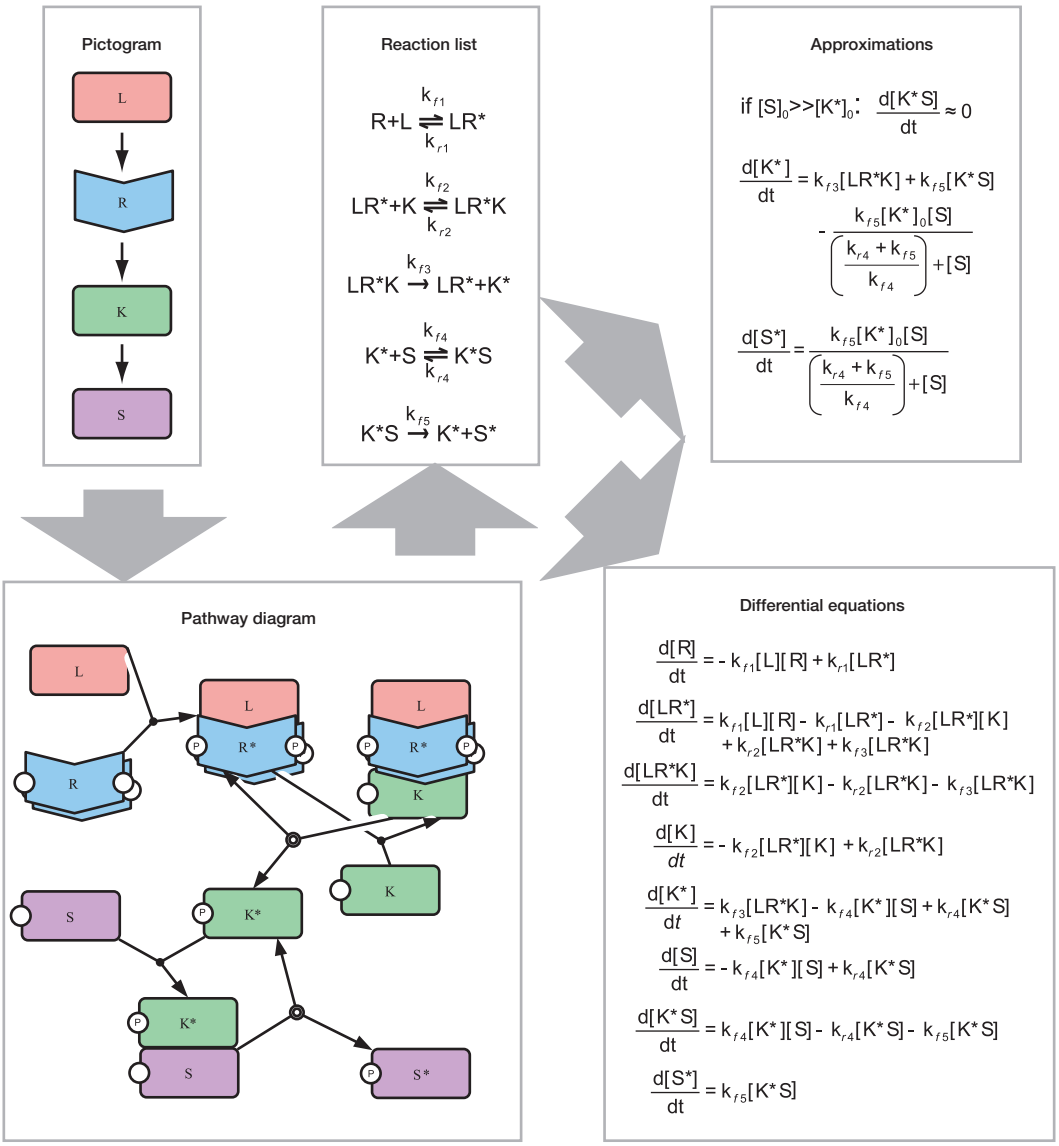


**Figure 3** Steps in physicochemical modelling. A pathway map is a highly abstracted pictogram of biomolecules and their interactions. Here, a simple linear ligand–receptor–kinase–substrate pathway is depicted. Although the pictogram conveys the general information flow in the network, mechanistic details required for mathematical modelling are absent. A formal pathway diagram drawn with CellDesigner details the reaction network[40]. Instead of representing the kinase as one object (as in the pictogram), each form of the kinase, either in complex or alone, is depicted (K, K*, LR*K, and K*S). A key challenge in developing a pathway diagram is making choices about granularity in number of species and reactions (see text). In this example, the receptor is a dimer and each subunit has two phosphorylation sites, yielding 64 possible ligand–receptor dimer complexes. However, this complexity is represented simply by two species: non-active and unphosphorylated (R) and ligand-bound, fully phosphorylated (LR*). It should be noted that approaches such as rules-based modelling may be preferred to the use of pathway diagrams (see text for details). A complete list of reactions is generated from the pathway diagram. This list can be automatically produced with several specialized software tools (Box 1). For reversible reactions, both forward and backward rate constants must be indicated. From the list of reactions, a system of differential equations is enumerated using appropriate rate laws, such as mass action kinetics, which uses the product of a rate constant and the concentrations of the reactants to calculate the reaction rates. Simplifying assumptions can be made to reduce the complexity or size of a model. The Michaelis-Menten approximation to enzyme–substrate kinetics is often applied. This particular rate form assumes rapid equilibrium of an intermediate complex (K*S), so that an equilibrium assumption is imposed ($d[K^*S]/dt = 0$), thus reducing the number of species in the model. Because this is an approximation, its use can alter model behaviour, particularly when the intermediate complex does not reach equilibrium or the reaction is tightly coupled to other processes[25–27].

can be introduced as simplified 'lumped' rates. At the same time, metabolic and synthetic processes are themselves being subjected to quantitative modelling. Thus, hybrid models can be constructed in which specific biological processes are alternately modelled in detail or in aggregate. For example, a highly simplified 'lumped rate' representation of a detailed metabolic model could be embedded in a physicochemical model of signal transduction to yield a hybrid. Realistic regulation could be reproduced by adding an adjustable parameter to the grouped metabolic model that makes metabolism dependent on signalling.

The issue of model granularity also arises with equations representing elementary reactions. For example, when a reaction is a hundred times or more faster than other reactions, it can be assumed that the fast process

operates in equilibrium. Differential equations can then be substituted by time-independent algebraic relations[21,22], resulting in a smaller model with more complex rate terms. A common example of algebraic substitution is replacing mass action kinetics with the Michaelis-Menten approximation (Fig. 3). Another example is representing a series of reactions as a 'transfer function' in which inputs (such as the concentrations of substrates) are translated into outputs (such as rate of formation of products) through an assumed or fitted algebraic function. A common biological transfer function is a Hill function, which captures cooperative behaviour (for example, of an enzyme cascade) in a single exponential function[6,23]. Substitution of ODEs by transfer functions and other algebraic terms constitutes a very simple type of model order reduction — an important topic that is unfortunately beyond the scope of this commentary[24]. As with all model order reductions, it is important to ascertain whether the implied assumptions are valid. It is not always appreciated, for example, that Michaelis-Menten kinetics represents a simplification of mass action kinetics that are valid in equilibrium, but not necessarily in the case of rapidly evolving reactions (Fig. 3)[25–27]. Caution is also warranted when transfer functions are introduced for the purpose of model simplification, and the discovery of switch-like and other non-linear behaviour is then treated as an insight into mechanism. The non-linear behaviour can be implicit in the mathematical forms themselves (ultrasensitivity in models with Hill functions, for example). In our opinion, circuit-level properties (such as switching and robustness) are more interesting when discovered in the course of analysing models based only on elementary reactions and mass action kinetics, including their stochastic representations.

## MODEL VERIFICATION, CALIBRATION AND VALIDATION

Once a preliminary model has been constructed it must be subjected to verification, calibration and validation. Verification is the process of determining how accurately prior knowledge and underlying assumptions have been translated into mathematical form (that is, whether the structure of the model is correct[28,29]). Calibration (also known as model regression or training) is the process by which parameters in a model are adjusted so as to match model performance to experimental data. Finally, model validation is the process of evaluating model performance against the primary design goal. In the case of biological models, this is usually a close match between model and experiment. Robustness and bistability are sometimes used as additional *post hoc* validation criteria; but only at the risk of introducing potentially incorrect bias.

The difficulty in assembling physicochemical models from pathway maps is frequently underestimated. The problem is not the mathematics itself, which in the case of ODEs is quite straightforward, but the sheer size, complexity and topological uncertainty of network diagrams. A wide variety of open source and commercial software has been developed to assist in translating diagrams into equations (Box 1). However, pathway maps are not necessarily the ideal starting point for modelling, as they ignore details of protein–protein association, particularly those that give rise to combinatorial complexity (Fig. 3). One attractive possibility, recently reviewed in detail[19], is to construct models automatically from sets of rules that encapsulate prior knowledge in human and machine-readable from. Rule-based models are potentially more amenable to automatic verification and model composition (joining small models together) than models with hand-written equations. Rules can also be used to generate families of models compatible with prior knowledge, but that are divergent in structure. However, formal methods for optimizing model structure have not yet been widely applied to biological networks. Thus, the issue arises whether a particular model structure and set of parameter values is the only way to model the data. Unfortunately, for large models the issue of 'model uniqueness' is difficult to address conclusively.

Assuming that a model with correct components and connectivity has been constructed, the next critical step is determining the values of parameters — the rate constants and initial conditions. Parameters can be measured directly, particularly in the case of protein and mRNA concentrations, or obtained from the literature. Rate constants can also be measured *in vitro*, although it is debatable whether rates in dilute solution are similar to those in the crowded intracellular environment. Even when considerable experimental data are available, it is common for many parameters in a pathway model to remain unmeasured and require estimation. Estimation involves computing the range of parameter values over which the model most closely matches experimental observation, given uncertainty in the data. The difficulty and reliability of model regression is closely tied to the number of free parameters and to the amount and quality of training data. In the noise-free case, $2n + 1$ observations are needed to estimate $n$ parameters in an ODE model, but the presence of noise makes the relationship between observation and parameter determinability more complex[30]. Regression with noisy data proceeds through a series of statistical tests whose power is proportional to the amount of data, and inversely proportional to the noise. Model calibration has the potential to yield either correct values or weakly determined values (the variable can take on a wide range of values without altering the goodness-of-fit to the observation). Unconstrained parameters are correctly regarded as ones about which modelling (and experimentation) give little insight. Moreover, parameters in complex models can be tightly coupled so that uncertainty in some parameters can affect many others (see accompanying article by Jaqaman *et al.* in *Nature Rev. Mol. Cell Biol.*).

Most physicochemical models, particularly those of eukaryotic networks, await rigorous regression, therefore parameter values are largely unconstrained. A useful starting point is to set values within physically plausible ranges and conservative catalytic rates, which we take to be: $k_f \sim 10^{-6}$ (number per cell) $^{-1}$ sec$^{-1}$; $k_r \sim 10^{-2}$–$10^{-3}$ sec$^{-1}$; $k_{cat} \sim 1$–$10$ sec$^{-1}$, $K_d$ (for complexes) $\sim 10^{-8}$ M and concentrations in the range of $10^3$–$10^6$ molecules per cell with a volume of 1 pl. Uncalibrated or partly calibrated models are useful for simulating results, but definitive conclusions cannot be drawn about specific rates and species conentrations. In addition, different models, or models with the same structures but regressed against different sets of data, cannot be rigorously compared.

## MODEL ANALYSIS

When model structure and parameter values have been determined or estimated, mathematical exploration and analysis can begin. Simulation is a simple but powerful tool for studying behaviour and guiding experiment. Time-dependent concentrations of key species can be compared over a range of concentrations, network topologies and rate parameters (Box 2; parameter determinability affects the accuracy of these computed time courses). To simulate the effects of RNA interference (RNAi), for example, a model is run with experimentally derived concentrations for the depleted protein. To simulate the effects of a kinase inhibitor, the catalytic rate constant, or levels of bound ATP, are reduced. In each case, simulated data can be compared to data such as flow-cytometry, quantitative western blotting and time-lapse microscopy[31].

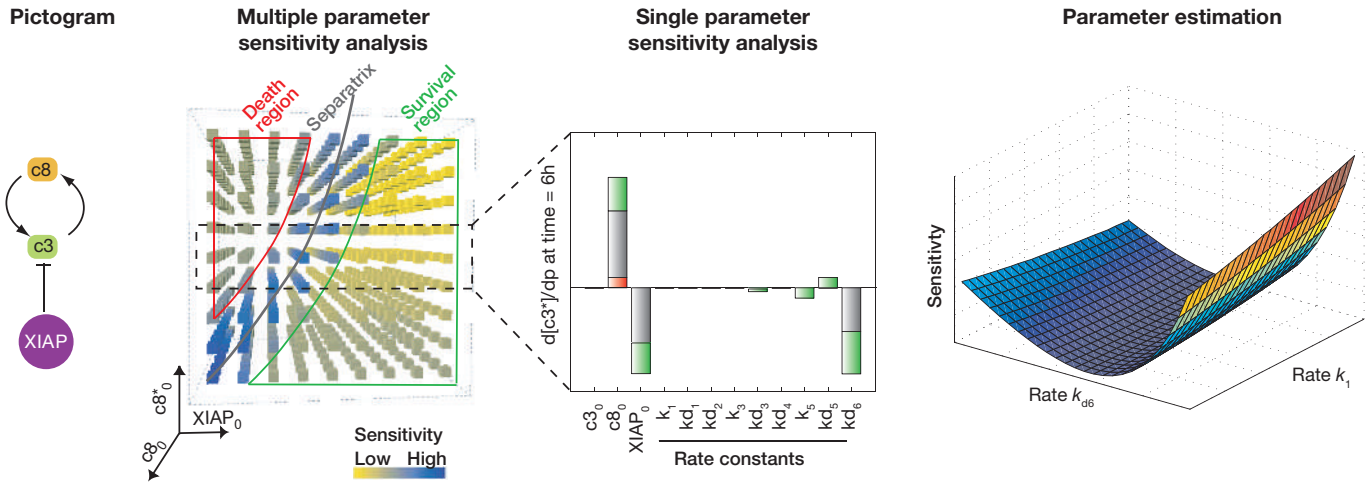


**Figure 4** Sensitivity analysis and parameter estimation are context specific. Sensitivity analysis evaluates the relative importance of specific initial conditions and rate constants for model output. Typically, a sensitivity objective function, d[output]/d[parameter], is evaluated at a finite or integrated time and the parameter with the largest effects on output identified. Sensitivity is context specific, that is, it is performed around a particular operating point or set of parameters. However, calculating sensitivity by simultaneously altering multiple parameters provides a more global view of network behaviour. In either case, a challenge lies in choosing an appropriate objective function and operating point for evaluation. As an example, multiple parameter sensitivity analysis performed on a model of caspase-3 activation (by caspase-8 and inhibited by XIAP), shows a region of high sensitivity (blue, the separatrix) separating the two different outcomes of death and survival (reproduced with permission from[32]). The separatrix is invariant to the initial amount of inactive caspase-8, but not the active form or XIAP. The critical amount of caspase-8 or XIAP needed to alter behaviour is dependent on the initial conditions of the other network species. Individual parameter sensitivity analysis performed from three different locales of phase-space (low, medium, and high initial concentrations of XIAP shown in red, grey and green, respectively) shows high sensitivity exists near the separatrix (gray) to caspase-8, XIAP, and $k_{d6}$ (the rate constant associated with ubiquitination of caspase-3 by XIAP). Parameter estimation uses an objective function to optimize parameter sets, with the goal of fitting models to data. The results of parameter estimation are context specific (dependent on time, the parameters of the model and the objective function) and many methods exist (such as Monte-Carlo simulations) to ensure that the estimation does not end in a local minimum of the objective function[41]. The shape of the minimum is a reflection of parameter sensitivity — long, thin valleys are sensitive to some, but not all parameters. As with sensitivity analysis, parameter estimation can elucidate which parameters should be measured experimentally (in our case, $k_{d6}$ is more important than $k_1$).

Sensitivity analysis is a powerful method for systematically determining which concentrations and rate constants in a model have the biggest influence on overall behaviour. Both the objective function (such as maximum integrated output or rate of change of output) and the set of variables over which the sensitivity analysis is performed can be defined, as can the range of values to be evaluated. Because the objective function is typically sensitive to changes in multiple parameters, multidimensional sensitivity analysis is the preferred approach[32,33]. Sensitivity is often visualized as a landscape of 'hills' and 'valleys' representing areas of parameter space in which small changes have significant effects on behaviour. In addition to revealing key parameters in a network, sensitivity analysis is valuable in ascertaining which parameters should be the focus of direct measurement or experimental perturbation. Insensitivity of a model to parameter variation has been equated with robustness, but robustness is better defined with respect to the sensitivity of a model to noise, either in the experimental data or in stochastic reactions (in product design, robustness involves searching the space of adjustable parameters for values that minimize the influence of noise on system behaviour[34]; Fig. 4).

A second approach to analysing models is qualitative and based on determining the classes of behaviour it can produce (such as bistability or oscillation). Sets of differential equations can be solved for stable and unstable steady states (that is, sets of parameter values for which the model does or does not return to a steady state when perturbed) by setting the rate of change to zero. Bifurcation analysis makes it possible to determine whether different trajectories through phase space lead to different qualitative behaviours around the steady state[35]. Stability and bifurcation analysis are of interest because they help explain how a network can switch between different states, (for example, 'on' and 'off' states). However, in the case of transient processes, other techniques from dynamical systems theory are required to determine how the output of a model will change over time (or time and space) when initial conditions or parameter values change. These methods include singular-perturbation theory and finite-time Lyapunov exponents (Fig. 4)[32,36].

## CHALLENGES

Current challenges in building physicochemical pathway models include developing more efficient ways to summarize prior knowledge and specify model structure, as well as better methods for combining and sharing models. Rules-based model assembly is likely to be important in this area, as is the still evolving systems biology markup language (SBML; see Swedlow *et al.* in this issue)[37]. Implementing models as Web services should also facilitate model exploration by non-experts (Box 2). Additional important areas for the future are acquiring sufficiently rich data (see accompanying article by Albeck *et al.* in *Nature Rev. Mol. Cell Biol.*), careful assessment of data reliability and rigorous model calibration.

Models of the type described in this review are largely phenomenological in that they aim to reproduce empirical data, often with considerable descriptive complexity. These models are largely explored through

simulation and are experimentally evaluated in terms of their predictive power. Given the difficulty and cost associated with experimentation, good phenomenological models are valuable not only for basic research but also industry and medicine. However, it is important to ask whether phenomenological models are truly explanatory. Explanatory models yield insight into circuit design, network dynamics and biophysical mechanism, and advance general understanding of biology. Models can be phenomenologically rich and predictive without being immediately explanatory. Careful analysis, including analytical approximation and algebraic manipulation, is usually valuable in deriving general understanding.

A wide range of opinion exists as to what constitutes scientific explanation[38]. Physicists typically seek explanations in terms of relatively few, but very powerful, and widely applicable theories. In contrast, biologists are often satisfied when the components of a process are enumerated, their connectivity determined and plausible (if not necessarily correct) biophysical principles invoked to describe mechanism. The large number of free variables in biochemical pathway models often provokes skepticism: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (a quip subsequently shown to be untrue — actually 30 variables are necessary!)[39]. Clearly models should be formulated with as few species and free parameters as possible, but engineering models are often similar to biological models in having many variables. Pathway models in biology represent a compromise between excessive complexity and too many degrees of freedom, and oversimplification and loss of mechanistic insight.

The issue of complexity is exacerbated by the difficulty in setting appropriate levels of biological abstraction. Ideally, all properties of a natural system would be deducible *ab initio* from fundamental physics. However, as physical systems become more complex they exhibit behaviours qualitatively different from those of simple systems. Fortunately, new axioms and theories can be formulated (for example, of solid matter) that are abstracted from, but compatible with, fundamental physical laws. In the physical sciences, great effort has gone into identifying appropriate levels of abstraction, and abstraction in engineering is an essential feature of good design. In biology, genes, cells and organisms represent useful layers of abstraction, but the wide disparity of complexity between genes and cells — the range at which pathway models operate – is not easily abstracted. Determining the most appropriate way to subdivide intracellular circuitry into discrete modules therefore remains a significant challenge.

## SUMMARY

Physicochemical models of biological pathways are attractive because their mathematics provides a means to merge prior knowledge with experimental data and underlying physical principles. Pathway models make it possible to examine in detail the effects of protein dysregulation and pharmacological intervention. Formal analysis should also help to uncover design features and common motifs, as well as reveal the extent to which pathways are truly modular. Models also have the potential to serve as transmittable repositories of knowledge. In our opinion, mathematical models, rather than databases, will dominate the dissemination of biological knowledge in the future. Our belief in the primacy of models is bolstered by historical experience with celestial mechanics, combustion chemistry, semiconductor fabrication and metabolic engineering. In each case, large sets of empirical data describing complex time-dependent processes were organized into models that evolved over time and gained considerable predictive power. As the models matured, they were adopted commercially as central components of industrial design. By analogy, accurate biological models of cell signalling circuits are likely to have a key role in the future of pharmaceutical discovery and medical treatment. □

## BOX 2 RESOURCES FOR ANALYSING AND SHARING CELL CIRCUIT MODELS

Once a model has been constructed (Box 1), it must be subjected to detailed analysis and shared.

### Model analysis

Model analysis refers to a wide range of techniques for probing model dynamics, estimating parameter sensitivity and identifying bifurcations and hysteresis. Simulating the temporal dynamics of species X is a particularly simple form of analysis in which d[X]/dt is integrated with respect to time. However, it is common for a single gene product to be represented in a model by multiple species, reflecting changes in, for example, compartmentalization. If these species cannot be distinguished experimentally (for example, on a whole cell western blot), then it is necessary to re-aggregate the model into observables. This process suffers from the same complexities and potential sources of error as species enumeration during model construction. Many programs have built in functionality for regular analysis tasks, including manual re-aggregation (COPASI, JDesigner/Jarnac, JSim, Jacobian and SimBiology). Often, more specialized or customized analyses require the use of a flexible scientific computing language, such as Matlab.

### Model publication and maintenance

Model publication and maintenance are critical steps in any modelling project, but remain remarkably cumbersome. A complete list of reactions and parameter values is in principle sufficient to reconstruct a model, but only with considerable effort. As an alternative, code can be shared, but only with people using the same software package. Systems biology markup language (SBML) is being developed as a universal XML-compliant standard for exchanging modelling data, but XML is not well suited to describing algebraic relationships among variables. SBML lacks 'roundtrip capability'[42] and remains a work in progress. An alternative approach is to exchange rules rather than models, thereby avoiding code all together. A complementary approach, whose potential has been illustrated in genomic applications such as BLAST, is to provide models as web services in which simulation, sensitivity analysis, etc. are possible (although this does not help in the creation of model derivatives). VirtualCell and JSim both allow models to be accessed via the web.

**Table 1** Glossary

| | |
|---|---|
| **Biochemical master equation (CME)** | Similar to a Markov process, the CME is a description of the stochastic state of a set of reactions and is specified as follows: $\frac{\partial p(x,t)}{\partial t} = \sum_{u=1}^{r}[-p(x,t)a_u(x) + p(x-v_u,t)a_u(x-v_u)]$ where $x$ is the state vector listing the concentrations of each species, $p$ is the probability that the species concentrations will be $x$ at time $t$, $r$ is the number of reactions, $a_u$ is the probability that reaction u will take place and $v$ is the change in $x$ due to reaction u[44,45]. |
| **Bifurcation analysis** | An analysis that shows how the qualitative behaviour of a model changes as a function of a few free parameters about the steady state. |
| **Bistability** | The ability of a model to rest in two different stable states (usually an 'on' and 'off' state). |
| **Calibration** | The process of adjusting parameter values so that model responses are as close as possible to experimental data — sometimes referred to as model regression or training |
| **Decision tree analysis** | Decision tree analysis creates a series of classifications that define sequential decision points based on a selected property or outcome of a model or empirical data. |
| **Finite-time Lyapunov exponents** | Also known as direct Lyapunov exponents (DLEs), these are a measure of parameter sensitivity analysis as multiple parameters change simultaneously. To find highly sensitive regions of phase space, at a finite (chosen) time, DLEs measure the distance between neighbouring trajectories, whose initial conditions were similar. DLEs are particularly useful in analysis of transient signalling. |
| **Free parameters** | A parameter that can be changed when fitting the model (for example, rate constants, initial conditions and other algebraic constants). |
| **Hill exponent** | An exponent (h) that traditionally quantifies the extent of cooperative binding of multiple proteins, but is also used to describe sigmoidal steepness: $f(x) = \frac{x^h}{p + x^h}$ |
| **Indeterminable parameter** | A parameter that is difficult to fit with confidence and can result when a model is underconstrained. |
| **Lumped parameter** | A free parameter that represents a combination of parameters associated with elementary equations. Used to reduce the detail in a model |
| **Mass-action kinetics** | Mass action kinetics define chemical reaction rates as a product of a rate constant and the concentrations of the reactants. Both forward and reverse rates can be specified (see Box 1 for examples). |
| **Mesoscale models** | Models of an intermediate size that are neither purely empirical, nor contain complete mechanistic detail. |
| **Michaelis-Menten kinetics** | An approximation of mass action kinetics typically used for enzyme–substrate interactions when the concentration of the substrate is in excess of the enzyme. Conservation of mass and applying the equilibrium assumption on the intermediate complex reduces the number of equations (see Box 1 for an example). |
| **Model order reduction** | Any mathematical procedure that reduces the number of equations and free parameters in a model with retaining the scope |
| **Model parameters** | Numbers in a model that supply the specific values for variables in equations and that are constant during a single execution, but that can change between executions. |
| **Objective function** | A function used to measure the 'goodness' of a model. A common example is deviation from an ideal behaviour, which we attempt to minimize for parameter estimation. |
| **Parameter estimation** | The regression process by which parameters are estimated by comparing model output to experimental data |
| **Prediction** | An outcome obtained by executing a model and that has not yet been obtained experimentally or used in calibration |
| **Prior knowledge** | Information that is available at the start of the modelling process and comprises part of the initial assumptions |
| **Robustness** | Variously, the insensitivity of a model output to changes in parameter values or to noise |
| **Sensitivity analysis** | Determining the change in model output associated with changes in parameter values (see Box 2). |
| **Singular perturbation theory** | Basis of principles used to separate model reactions into fast and slow reactions. This is especially useful in reducing model complexity and evaluating transient signals. |
| **Simulation** | The process of using mathematical models to study the responses and properties of a system under differing conditions — usually different parameter values and occasionally different model structures |
| **Transfer function** | A function that maps an input to an output value. |
| **Web service** | A computational resource than can be accessed via the world wide web, typically using a web browser or other local client |

**ACKNOWLEDGEMENTS**

We thank G. Danuser, J. Gunawardena, B. Schoeberl and W. Fontana for critical reading of this manuscript. This work was funded by a National Institutes of Health (NIH) grant P50-GM68762 and a Deparment of Energy (DOE) Computational Science Graduate Fellowship to B.B.A. (DE-FG02-97ER25308).

**COMPETING FINANCIAL INTERESTS**

The authors declare that they have no competing financial interests.

1. Nurse, P. A long twentieth century of the cell cycle and beyond. *Cell* **100,** 71–78 (2000).
2. Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A. & Palsson, B. O. Metabolic pathways in the post-genome era. *Trends Biochem. Sci.* **28,** 250–258 (2003).
3. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4,** R54 (2003).
4. Bhalla, U. S., Ram, P. T. & Iyengar, R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297,** 1018–1023 (2002).
5. Hoffmann, A., Levchenko, A., Scott, M. L. & Baltimore, D. The IκB–NF-κB signaling module: temporal control and selective gene activation. *Science* **298,** 1241–1245 (2002).
6. Huang, C. Y. & Ferrell, J. E., Jr. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl Acad. Sci. USA* **93,** 10078–10083 (1996).
7. Markevich, N. I., Hoek, J. B. & Kholodenko, B. N. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.* **164,** 353–359 (2004).
8. Schoeberl, B., Eichler-Jonsson, C., Gilles, E. D. & Muller, G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnol.* **20,** 370–375 (2002).
9. Oda, K., Matsuoka, Y., Funahashi, A. & Kitano, H. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* **1,** 0010 (2005).
10. Gardiner, C. W. *Handbook of Stochastic Processes* (Springer, New York, 2005).
11. Danuser, G. & Waterman-Storer, C. M. Quantitative fluorescent speckle microscopy of cytoskeleton dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 361–387 (2006).
12. Mallavarapu, A. & Mitchison, T. Regulated actin cytoskeleton assembly at filopodium tips controls their extension and retraction. *J. Cell Biol.* **146,** 1097–1106 (1999).
13. Odde, D. J. & Buettner, H. M. Time series characterization of simulated microtubule dynamics in the nerve growth cone. *Ann. Biomed. Eng.* **23,** 268–286 (1995).
14. Ponti, A., Machacek, M., Gupton, S. L., Waterman-Storer, C. M. & Danuser, G. Two distinct actin networks drive the protrusion of migrating cells. *Science* **305,** 1782–1786 (2004).
15. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94,** 814–819 (1997).
16. Paulsson, J. Summing up the noise in gene networks. *Nature* **427,** 415–418 (2004).
17. Conzelmann, H., Saez-Rodriguez, J., Sauter, T., Kholodenko, B. N. & Gilles, E. D. A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. *BMC Bioinformatics* **7,** 34 (2006).
18. Blinov, M. L., Faeder, J. R., Goldstein, B. & Hlavacek, W. S. A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems* **83,** 136–151 (2006).
19. Hlavacek, W. S. *et al.* Rules for modeling signal-transduction systems. *Sci. STKE* re6 (2006).
20. Tolle, D. P. & Le Novere, N. Particle-Based Stochastic Simulation in Systems Biology. Current Bioinformatics **1,** 1–6 (2006).
21. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403,** 339–342 (2000).
22. Tyson, J. J., Chen, K. C. & Novak, B. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* **15,** 221–231 (2003).
23. von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406,** 188–192 (2000).
24. Conrad, E. D. & Tyson, J. J. in *System Modeling in Cellular Biology* (eds. Szallasi, Z., Stelling, J. & Periwal, V.) 97–123 (MIT Press, Cambridge, 2006).
25. Farrow, L. A. & Edelson, D. The steady-state assumption: fact or fiction? *Int. J. Chem. Kin.* **1,** 309–322 (1974).
26. Flach, E. H. & Schnell, S. Use and abuse of the quasi-steady-state approximation. *IEE Proc. Syst. Biol.* **153,** 187–191 (2006).
27. Segel, L. A. On the validity of the steady state assumption of enzyme kinetics. *Bull. Math. Biol.* **50,** 579–593 (1988).
28. Balci, O. in *Proceedings of the 29th conference on Winter simulation* 135–141 (ACH Press, Atlanta, 1997).
29. Sargent, R. G. in *2005 Proceedings of the Winter Simulation Conference* 14 (ACH Press, New York, 2005).
30. van Riel, N. A. W. & Sontag, E. D. Parametric estimation in models combining signal transduction and metabolic pathways: the dependent input approach. *IEE Proc.Syst. Biol.* **153,** 263–274 (2006).
31. Geva-Zatorsky, N. *et al.* Oscillations and variability in the p53 system. *Mol. Syst. Biol.* **2,** 0033 (2006).
32. Aldridge, B. B., Haller, G., Sorger, P. K. & Lauffenburger, D. A. Direct Lyaponov exponent analysis enables parametric study of transient signalling governing cell behaviour. *IEE Proc. Syst. Biol.* **153,** (2006).
33. Bentele, M. *et al.* Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. *J. Cell Biol.* **166,** 839–851 (2004).
34. Frey, D. & Li, X. in *Engineering Systems 2004 Symposium* (MIT Engineering Systems Division, Cambridge, 2004).
35. Wiggins, S. in *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (eds. Marsden, J. E., Sirovich, L. & Antman, S. S.) 356–xxx (Springer-Verlag, New York, 2003).
36. Hoppenstaedt, F. C. *Analysis and Simulation of Chaotic Systems* (Springer-Verlag, New York, 2000).
37. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19,** 524–531 (2003).
38. Merks, R. M. H. & Glazier, J. A. A cell-centered approach to developmental biology. *Physica A* **352,** 113–130 (2005).
39. Dyson, F. A meeting with Enrico Fermi. *Nature* **427,** 297 (2004).
40. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnol.* **23,** 961–966 (2005).
41. Moles, C. G., Mendes, P. & Banga, J. R. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* **13,** 2467–2474 (2003).
42. Alves, R., Antunes, F. & Salvador, A. Tools for kinetic modeling of biochemical networks. *Nature Biotechnol.* **24,** 667–672 (2006).
43. Le Novere, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol* **23,** 1509–1515 (2005).
44. Gillespie, D. T. A Rigorous Derivation of the Chemical Master Equation. *Physica A* **188,** 404–425 (1992).
45. Roussel, M. R. & Zhu, R. Reducing a chemical master equation by invariant manifold methods. *J. Chem. Phys.* **121,** 8716–8730 (2004).