Construction and control analysis of biochemical network models

Habilitationsschrift
zur Erlangung der Lehrbefähigung
für das Fach Theoretische Biophysik

vorgelegt dem Fakultätsrat der Mathematisch–Naturwissenschaftlichen Fakultät 1
der Humboldt-Universität zu Berlin

von

Dr. Wolfram Liebermeister

geboren am 25. Juli 1972 in Tübingen

Prof. Dr. Jan-Hendrik Olbertz
Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Stefan Hecht, PhD
Dekan

Berlin, den 16. 5. 2012

Gutachter:

1. Prof. Dr. Jens Timmer
2. Prof. Dr. Wolfgang Marwan
3. Prof. Dr. Hanspeter Herzel

Contents

Summary

Simulating living cells in the computer is a great promise. However, before large dynamic models can become reality, the details of biochemical networks have to be revealed and brought to agreement with available metabolome, proteome, and flux data. Flux analysis and kinetic modelling exist as common frameworks, but model building and data integration remain complex and are often governed by intuition. As models become larger, more and more routine steps need to be supported or executed by software. This automation requires theory about model validity, safe formalisms that capture the intuition of modellers, and modelling approaches that can compensate for missing data. Moreover, semantic information encoded within the models can enable software to process the biological meaning of models automatically.

In the past eight years, I have developed concepts and methods to support a dynamic modelling of large metabolic systems yet to come. In this habilitation treatise, I present these methods and the publications in which they were developed. The first two sections introduce basic concepts of biochemical modelling with illustrations from my chapters of *Systems Biology - a Textbook*. The following three sections extend manual kinetic modelling in three different directions: first, approaches to fill given networks with standard rate laws and to determine their kinetic constants. Second, ways to compute the control coefficients in large metabolic systems and to describe the deterministic or stochastic dynamic close to steady state. Third, concepts for capturing and processing the biochemical meaning of models, which enable automatic model search, validity checks, and model combination. All these methods contribute to an efficient construction of biochemical models and to the integration of various types of biochemical data.

1 Introduction: dynamic in biochemical networks

1.1 The processes of life emerge from biochemistry

All processes of life are based on biochemistry. If we could zoom deep into a living cell, we would find a highly structured machinery of molecules and macromolecular complexes. Their stochastic but ordered movements, governed by the laws of chemistry and thermodynamics, constitute life. This materialistic view of biology has its roots in the second half of the 19^{th} century and was strongly promoted, for instance, by Rudolf Virchow. The old distinction between living and non-living matter was further challenged when Eduard Buchner proved that yeast extracts can perform fermentation [Buchner, 1897]. Thus, biological processes could take place outside the context of a living cell just like other chemical processes. Today, biologists are convinced that the special nature of living matter lies in the way it is structured and in the dynamic processes that emerge from it. Revealing the mechanistic details remains the big challenge for molecular biology. In his book "What is life" [Schrödinger, 1944], Erwin Schrödinger summarised this scientific programme in the question "How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?".

From genotype to phenotype

Biochemical processes in cells are much more complex and organised than chemical reactions in a test tube, and to understand living systems, we need to consider various levels of organisation aside from mere chemistry. First of all, macromolecules like proteins or RNA can exert very specific functions, and their blueprints are inherited via gene sequences and adapted during evolution. However, although many genome sequences are known and the layout of cellular networks becomes clearer, it is still hard to see how these sequences give rise to dynamic, self-sustained biochemical processes and to the distinct phenotypes we observe.

Biochemical systems are complex

In the view of systems biology, biomolecules form the material basis, but life is constituted by the dynamic behaviour of their interactions, movements, and transformations. The macromolecules in a cell form biochemical reaction systems, which show global dynamic properties like bifurcations, robustness and sensitivity [Stelling et al., 2004]. Further specific features of cells are their spatial structure and the regulation of chemical reactions, which allows biological systems to behave stably and flexibly at the same time. Since Monod's seminal studies on the transcriptional regulation Lac operon [Jacob and Monod, 1961], it has become clear that information encoded in substance concentrations is crucial for cells to orchestrate their internal biochemical processes and that regulation – e.g.. transcriptional adjustment of enzyme levels – allows them to adapt to challenges by the environment.

Dynamics of biochemical networks

The dynamic in cells is determined by network structures and by kinetic properties of macromolecules, which are predefined by the genome sequence. However, the phenotype, which is the target of natural selection, only emerges in the living cell and through the biochemical dynamic. The aim of systems biology is to understand this connection – how global dynamic arises from individual compounds and their interactions – through massive data collection and models. In the long run, theory should also explain how cellular networks were shaped by natural selection, how they are adapted to specific environments, how they deal with unpredictable events and biochemical noise, and how processes within single cells determine the behaviour of cell populations and multicellular organisms. The first step towards this aim is to understand how complex biochemical systems are structured and how they work mechanistically. Tracing all interactions one by one would hardly be possible, but mathematical models allow us to simulate the dynamic of complex biochemical systems in time, and sensitivity analyses like Metabolic Control Analysis [Heinrich and Schuster, 1996] can further help to understand the global dynamics.

Outline of the text

This habilitation treatise summarises concepts and methods for metabolic modelling that I developed during the past years (for an overview, see Table 1). After some general words about metabolic models in this section and an overview of modelling methods in section 2, section 3 is concerned with the translation of given metabolic

Method	Section	References
Network component analysis	3.1	[Buescher et al., 2012]
Standard rate laws	3.2	[Liebermeister and Klipp, 2006a, Liebermeister et al., 2010]
Parameter dependence scheme	3.3	[Liebermeister and Klipp, 2005]
Parameter uncertainties	3.4	[Liebermeister and Klipp, 2005]
Priors obtained by machine learning	3.4	[Liebermeister, 2005a]
Parameter balancing	3.6	[Liebermeister and Klipp, 2006b, Lubitz et al., 2010] [Borger et al., 2007]
Elasticity sampling	4.3	[Liebermeister et al., 2010]
Sampling of interaction effects	4.3	[Liebermeister et al., 2010]
Response to periodic perturbations	4.4	[Liebermeister, 2005b]
Propagation of chemical noise	4.5	[Liebermeister, 2005b]
Model reduction	4.6	[Liebermeister et al., 2005]
Model annotation	5.2	[Krause et al., 2010]
Model merging	5.2	[Liebermeister, 2008, Schulz et al., 2006] [Krause et al., 2010, Schulz et al., 2011]
Model search	5.3	[Schulz et al., 2011]
Model alignment	5.3	[Schulz et al., 2011]
Model validity criteria	5.4	[Liebermeister, 2008]

Table 1: List of methods described in this habilitation treatise

networks into kinetic models. Parameter balancing, an approach based on standardised rate laws and dependencies between kinetic constants, allows to determine model parameters by massive data integration. Section 4 describes how reaction elasticities in large models can be sampled and used to study the dynamic of small perturbations. In section 5, I present methods for automatic model processing that employ semantic annotations and facilitate model search, validity checks, and model merging. Further background information about thermodynamic flux analysis, parameter estimation, or stochastic biochemical models can be found in my chapters of *Systems Biology – a Textbook* [Klipp et al., 2009].

1.2 Metabolic networks and their regulation

Metabolism is a central function of all cells. The life cycle of macromolecules starts and ends with small molecules, which are interconverted by thousands of enzymatic reactions. Like all chemical reactions, these reactions are governed by thermodynamics, but catalysis by enzymes makes them efficient and controllable. Elements like compounds, enzymes, and chemical reactions can be depicted in biochemical networks: Figure 1 shows, as an example, the central carbon metabolism of the soil bacterium *Bacillus subtilis* and its transcriptional and allosteric regulation. Reconstructions of metabolic, signalling, and gene regulation networks can be much larger, comprising thousands of metabolites, chemical reactions, genes, or proteins. The network of enzyme-catalysed reactions resembles a technical device: many enzyme activities are regulated by small compounds, either by direct binding or indirectly via transcription factors that adjust the production of enzymes to the current needs. Aside from metabolic systems, which enable a continuous flow of matter, there are signalling systems which transmit information. Their elements are usually molecules that can assume different states – e.g., bound or unbound, phosphorylated or not – and, in turn, can influence the state of other substances. Metabolic networks and signalling systems are interlinked and form a complex feedback system that allows cells to adapt to external challenges.

Biochemical networks

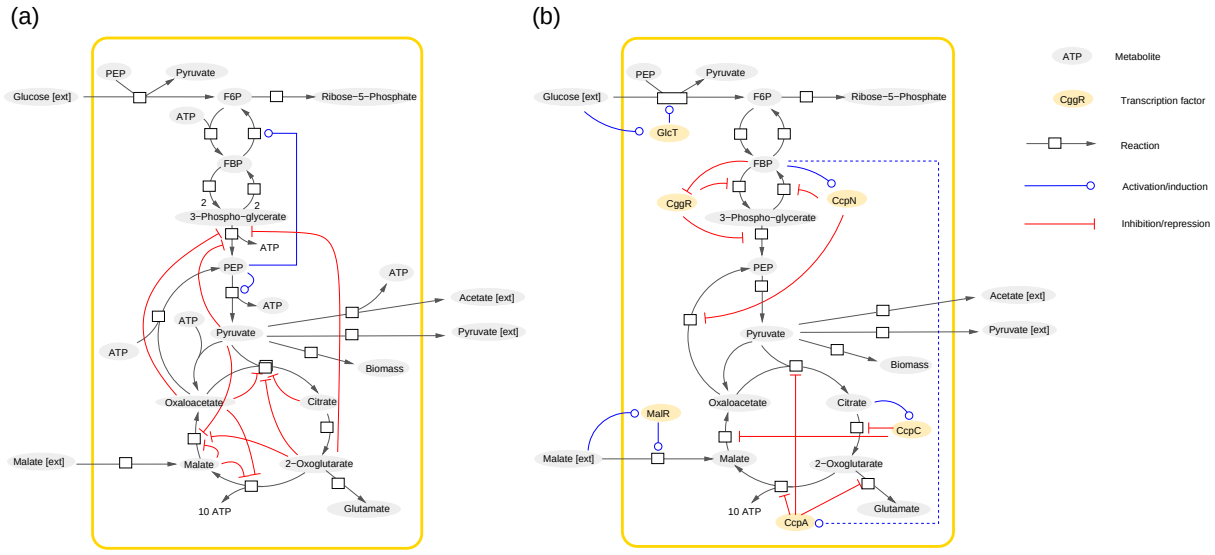

Figure 1: Scheme of central metabolism in the soil bacterium *Bacillus subtilis*. (a) Metabolic reactions and allosteric enzyme regulation. The network comprises a simple version of glycolysis, several side branches, and the citric acid cycle. Lumped reactions between small metabolites (grey ellipses) are shown as arrows with squares. Allosteric regulation allows enzyme activities to respond rapidly to changing concentrations of small molecules (blue arrows: activation; red arrows: inhibition). (b) The same network, with transcriptional regulation (blue: effective activation; red: effective repression). The transcription factors (yellow ellipses) are regulated by small molecules, which close regulatory feedback loops.

1.3 Mathematical models of biochemical networks

Biochemical models

The interplay of molecules in cells is invisible and much too complex to be imagined in all details. Although we cannot grasp it in its full complexity, we may focus on some relevant subsystems and ask what dynamic behaviour will arise in them. Mathematical models can help us to do this. In mechanistic models, we depict processes as we imagine them to happen and translate these pictures into a mathematical form that can be clearly communicated and used for calculations. The way from a qualitative scheme to mathematical equations and computer simulations is shown in Figure 2. Mechanistic models are based on physical laws, but due to the necessary simplifications, they often describe hypothetical "proxy" cells containing just a well-stirred mixture of chemicals. In fact, these are two of the big challenges in systems biology: finding the right degree of simplification and designing experiments for which this simplification is justified.

Networks and quantitative dynamic

Biological networks are a good starting point for building quantitative models. By analysing their structure, we can understand important features of a metabolic system, for instance, which biochemical conversions it can perform. However, biological networks alone do not suffice to understand the biochemical dynamic. If we try to see how the system in Figure 1 would respond to certain perturbations, e.g., the sudden addition of a nutrient, we easily get lost. On the one hand, it is difficult to trace all the possible routes of action in a complex system; on the other hand, the dynamic behaviour depends on quantitative details that are not part of the picture. To create quantitative models, we need to fill the network with rate laws describing the specific kinetic properties of enzymes. By "assigning numbers to the arrows" [Ronen et al., 2002], we can turn networks into dynamic mathematical models and use them for quantitative simulations.

Complex cell models

Mathematical models have a long tradition in biology and biochemistry (see [Lotka, 1925] for an early example). Since the seventies, metabolic pathways have been studied intensively with the help of flux analysis and metabolic control analysis. However, comprehensive cell models became a realistic vision when high-throughput technologies started to provide data on a genome-wide scale. Since the success of DNA microarrays, technology advanced and

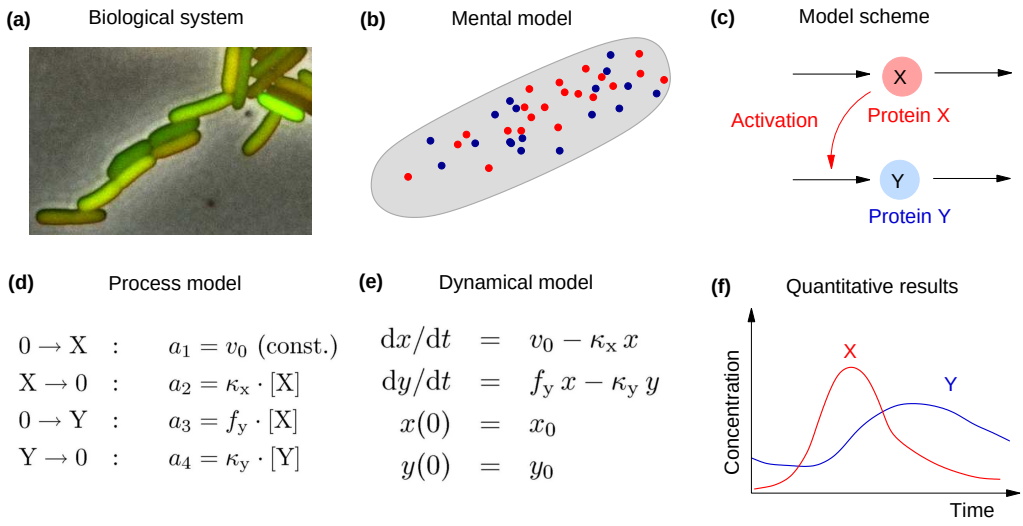

Figure 2: The process of modelling from biological systems to numerical simulation. To describe the expression of two genes in bacteria (a), we consider their transcript numbers as relevant variables (b) and depict their production and degradation in a simple scheme (c). This scheme can be translated into rules defining a stochastic process (d) or into differential equations describing the average concentrations of transcripts (e). Numerical simulation yields time courses (f) which can then be compared or fitted to experimental data. Figure from *Systems Biology – a Textbook* with kind permission.

metabolome, proteome, and flux data became available. The emergent field of systems biology had to join two traditions, small prototypical models from mathematical biology and statistical analyses of high-throughput data. Eventually, efforts from both sides may converge into detailed models, which explain high-throughput data within a common frameworkand as the output of an understandable biological system. Such models will serve several purposes:

Purposes of models

1. Models provide a language to describe how biological systems are structured and how they work. In contrast to verbal descriptions or graphical schemes, they also allow to simulate their dynamic. In the future, simulations may replace some experiments, allowing us to test the effects of drugs or genetic modifications and to determine variables that cannot directly be measured.

2. By comparing model predictions to data, we can test the mental picture that a model relies on. As long as a model has not been falsified, we can assume that it captures all relevant processes and describes them reasonably well.

3. Models are becoming increasingly important for data analysis because they can integrate different types of data – among others, metabolite and enzyme concentrations, metabolic fluxes, and kinetic or thermodynamic constants – and to check them for consistency. Moreover, parameter fitting and model selection permit to fill gaps, e.g. to compute non-measurable quantities by solving inverse problems, or to design new experiments that can provide the missing information.

In summary, model simulations can shed light on the dynamic of biochemical systems and bring out the specific roles of physical laws, the network structure, enzyme regulation, and the quantitative properties of enzymes or metabolites. The following section summarises established modelling methods, their mathematical concepts, and the considerations behind them.

2 Biochemical network models

2.1 Modelling approaches for biochemical networks

Types of mathematical models

A biochemical network can be realised by various types of models, representing different pictures of reality. As shown in Figure 2, the compounds in a cell may be described as single molecules, by molecule numbers, or by real-valued concentrations. Moreover, models can describe not only what *does* happen in a cell, but also what *may* happen (e.g., which metabolic fluxes are physically possible), what *is likely* to happen, and what *is supposed* to happen (e.g., in order to maximise growth rate). Different mathematical formalisms can express these different perspectives. In metabolic models, spatial and stochastic aspects are usually neglected: the main modelling approaches are *stoichiometric models*, which capture the metabolic fluxes in large networks, and *kinetic models*, which describe the mechanisms and quantitative dynamic of biochemical pathways.

Flux analysis and kinetic models Stoichiometric models, combined with flux analysis methods like flux balance analysis (FBA) [Orth et al., 2010], elementary mode analysis [Schuster et al., 1999], energy balance analysis [Beard et al., 2002], or the principle of minimal fluxes [Holzhütter, 2004, Holzhütter, 2006], predict metabolic fluxes from reaction stoichiometries and from a number of physical and biochemical assumptions. Given a list of reactions, they allow us to find flux distributions that are stationary (or "steady"), resembling the flow in a river in which water is nowhere accumulated nor depleted. In contrast to chemical equilibria, such stationary states or *Fließgleichgewichte* ("flow equilibria"), as they are tellingly called in German [von Berthalanffy, 1932, von Berthalanffy, 1953], contain non-zero fluxes, but the production and consumption of all substances is balanced. An exception are the *external* substances, which participate in additional reactions not covered by the model. The balanced production and consumption of metabolites is expressed by the stationarity condition

Stationary state

$$N\,v = 0, \tag{1}$$

where v is the flux vector and N is the stoichiometric matrix, restricted to the internal metabolites. Within the space of potential flux vectors, Eq. (1) defines the subspace of stationary flux distributions. Additional conditions (fixed flux directions, upper and lower limits, and optimality criteria such as maximal biomass production) can be used to narrow down the flux distributions even further. Other constraints come from thermodynamic analysis [Beard et al., 2002, Qian and Beard, 2005, Kümmel et al., 2006, Hoppe et al., 2007]: owing to the second law of thermodynamics, all fluxes have to lead from higher to lower chemical potentials. Flux distributions that cannot satisfy this condition – e.g., certain flux cycles – are physically impossible and can be excluded. Moreover, thermodynamics links the reaction directions to metabolite concentrations and to Gibbs free energies of formation. These constraints do not specifically depend on enzyme kinetics, which makes them particularly useful.

Thermodynamic constraints

Kinetic models Stoichiometric models can cover large networks and successfully predict biomass yield, viability, and growth defects of deletion mutants. However, to describe the dynamic interplay between enzyme levels, reaction rates, and metabolite levels, we need to consider reaction kinetics. By expressing the reaction rates v_l as functions of metabolite and enzyme concentrations c_i and u_l, we obtain a system of ordinary differential equations for the internal metabolite concentrations:

$$\frac{\mathrm{d}c}{\mathrm{d}t} = N\,v(c,u). \tag{2}$$

Such kinetic models describe the production and consumption of compounds by kinetic rate laws and can be used for dynamic simulations. They are not restricted to metabolism, but can capture all kinds of biochemical reaction

systems, including metabolic and signalling pathways and the slower transcriptional adaption of enzyme levels.

The main difficulty in building kinetic models is to choose the rate laws and their kinetic parameters. If rate laws are unknown, they can be replaced by standard rate laws representing simple reaction mechanisms. In a well-mixed chemical solution, reaction rates are proportional to the reactant concentrations or, more precisely, to the frequency of reactant molecules hitting one another. Since reaction events occur in both directions, we obtain forward and reverse terms. For a bimolecular reaction A + B $\leftrightharpoons$ 2 C, the resulting reversible mass-action rate law reads

Rate laws

$$v = k^{\mathrm{f}}\, a\, b - k^{\mathrm{r}}\, c^2. \tag{3}$$

with substance concentrations a, b, and c and rate constants k^{f} and k^{r}. Enzyme-catalysed reactions are more complicated: they comprise several subprocesses (substrate binding, catalysis, product release) and possibly conformation changes of the enzyme. Enzymatic reactions can further be affected by effector molecules or enzyme modifications. Most models do not resolve such subprocesses in detail, but hide them in effective rate laws which involve the concentrations of reactants, effectors, and the enzyme. The most common choice for a unimolecular reaction A $\leftrightharpoons$ B is the Michaelis-Menten kinetics

Michaelis-Menten kinetics

$$v = u\, \frac{k^{\mathrm{cat}} a}{a + k_{\mathrm{A}}^{\mathrm{M}}} \tag{4}$$

or its reversible variant

$$v = u\, \frac{k^{\mathrm{cat}+} a/k_{\mathrm{A}}^{\mathrm{M}} - k^{\mathrm{cat}-} b/k_{\mathrm{B}}^{\mathrm{M}}}{1 + a/k_{\mathrm{A}}^{\mathrm{M}} + b/k_{\mathrm{B}}^{\mathrm{M}}}. \tag{5}$$

The symbols in this formula denote the enzyme concentration (u), the reactant concentrations (a and b), the Michaelis constants ($k_{\mathrm{A}}^{\mathrm{M}}$ and $k_{\mathrm{B}}^{\mathrm{M}}$), and the catalytic constants ($k^{\mathrm{cat}+}$ and $k^{\mathrm{cat}-}$). Both rate laws are saturable: for increasing substrate concentrations, their rates approach a maximal value which is proportional to the enzyme concentration.

Thermodynamics in kinetic models Like all chemical processes, enzymatic reactions obey the laws of thermodynamics. If we ignore this in kinetic models, we run the risk to obtain models that effectively describe a *perpetuum mobile*. The most important thermodynamic requirements are as follows: each compound i has a chemical potential μ_i, defined as the derivative of the system's Gibbs free energy by the compounds' mole number. For ideal mixtures, the chemical potentials read $\mu = \mu^{(0)} + RT \ln c$ with gas constant R, absolute temperature T, and concentration c. The negative difference $A = -\Delta\mu_l$ of chemical potentials along a reaction l, called the reaction affinity, can be seen as a thermodynamic forcing driving the reaction. Any chemical reaction is, at least in principle, reversible and consists of a difference of forward and backward fluxes. Their ratio $v^{\mathrm{f}}/v^{\mathrm{r}}$ obeys the formula $v^{\mathrm{f}}/v^{\mathrm{r}} = \exp(A/RT)$ [Beard and Qian, 2007]. The reaction affinity thus determines whether a reaction is close to equilibrium ($A \approx 0$, $v^{\mathrm{f}} \approx v^{\mathrm{r}}$) or nearly irreversible (A large, $v^{\mathrm{f}} \gg v^{\mathrm{r}}$). In chemical equilibrium, the net flux will vanish; in all other cases, it leads from higher to lower chemical potentials. Every biochemical system has chemical equilibrium states in which all reaction affinities and net reaction rates vanish. The mass-action ratios for all equilibrium states are identical and therefore called equilibrium constants. Their logarithms satisfy the formula

Reversibility

Equilibrium constant

$$\ln k^{\mathrm{eq}} = -\frac{1}{RT}\, \left[\mu_{\mathrm{P}}^{(0)} - \mu_{\mathrm{S}}^{(0)}\right] \tag{6}$$

where $\mu_{\mathrm{S}}^{(0)}$ and $\mu_{\mathrm{P}}^{(0)}$ are the standard chemical potentials, summed over substrates or products. If the network contains loops, this formula can lead to mathematical dependencies between the equilibrium constants. As an

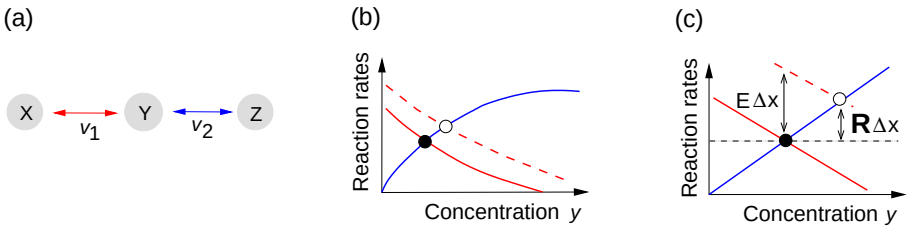

Figure 3: Stationary states in a metabolic system. (a) In a simple two-reaction pathway, the concentration of X and Z are externally controlled, while the concentration of the intermediate Y arises from the biochemical dynamic. (b) The rates of both reactions (solid lines, red and blue) vary with the concentration of Y (x-axis). In a steady state, both rates must be identical; this determines both the concentration of Y and the flux (solid circle). An increased concentration of X shifts the rate of reaction 1 (dashed red curve) and thereby the steady state (open circle). (c) A close-up compares the immediate rate change directly after the perturbation (left arrow) to the long-term, steady rate change (right arrow). In a linear approximation, these changes are proportional, respectively, to the elasticity E and to the flux response coefficient R. Figure redrawn from *Systems Biology – a textbook*.

example, consider three uni-molecular reactions in a circle: if we multiply all equilibrium constants along the circle, the result has to be 1. This dependence is an example of a *Wegscheider condition* [Wegscheider, 1902, Schuster and Schuster, 1989]. Wegscheider conditions can arise for all quantities that represent differences along chemical reactions and interlink them between reactions. Another type of constraints, the Haldane relationships [Haldane, 1930], relate the kinetic constants in a rate law to the equilibrium constant of the reaction. To derive such a relationship, we consider a chemical equilibrium, set the reaction rate to zero, and solve the equation for the mass-action ratio. For the mass-action rate law (3), we obtain the simple Haldane relationship $k^{\mathrm{eq}} = k^{\mathrm{f}}/k^{\mathrm{r}}$.

Wegscheider conditions

Haldane relationships

Models that violate one of the conditions above will be thermodynamically incorrect. Sometimes this does not matter, but considering thermodynamics can have several advantages: it helps to obtain a sound and general formalism in which irreversibility will not be assumed, but explained by the model; it avoids models that describe a *perpetuum mobile*; it permits to relate reaction rates to thermodynamic forces, to link kinetic models and thermodynamic flux analysis, and to make better use of kinetic data.

Metabolic control theory A central question in systems biology is how cells respond to perturbations like genetic modifications, the action of drugs, or changes in the environment. Often, the interesting point is not the dynamic right after a perturbation, but its long-term effect on the steady state. To understand the effect of perturbations, we cannot just consider the perturbed reaction, but have to account for its interaction with the entire network. The inhibition of an enzyme, for instance, will provoke a depletion of substrate and an accumulation of product; this will have secondary effects and may eventually affect many reactions and metabolites, which all contribute to the overall effect of the inhibition. In the example in Figure 3, the system dynamic dampens the initial perturbation, and the new stationary flux is lower than the initial flux right after the perturbation. In large metabolic systems, tracing such global effects may be complicated, but kinetic models can help. After setting the left-hand side of Eq. (2) to zero, we can solve for the concentrations and fluxes and obtain the stationary concentrations as functions of the enzyme levels and other system parameters. A sensitivity analysis of these stationary concentrations and fluxes can tell us about the global effects of small local perturbations. Metabolic Control Analysis (MCA) [Heinrich and Schuster, 1996, Hofmeyr, 2001] allows us to compute such sensitivities directly from the network structure and the rate laws. It interrelates local and global responses and can also cover second-order effects [Höfer and Heinrich, 1993], perturbations of dynamic time courses [Ingalls and Sauro, 2003], and periodic parameter perturbations [Ingalls, 2004, Liebermeister, 2005b]. In my work, I have used MCA for parameter estimation [Liebermeister and Klipp, 2006b] and to study the propagation of chemical noise [Liebermeister, 2005b], variability and uncertainties [Liebermeister and Klipp, 2005], and

Perturbations in biochemical systems

Metabolic Control Analysis

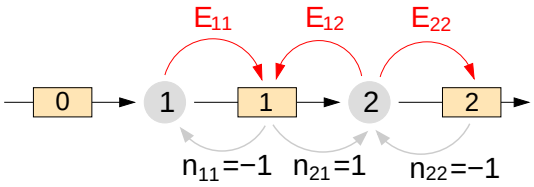

Figure 4: Interplay between metabolite concentrations and reaction rates in kinetic models. The scheme shows a metabolic pathway (circles: metabolites; boxes: reactions). A small concentration change will move the system out of its original steady state and lead to rate changes which can be linearly approximated by the elasticities E_{li} (top). The perturbed rates will increase or decrease the concentrations according to the stoichiometric coefficients n_{il} (bottom). The interplay of both processes leads to dynamic changes of concentrations and rates. Figure redrawn from *Systems Biology – a textbook*.

the optimal regulation of enzymes [Liebermeister et al., 2004].

Reaction elasticities

MCA expresses local and global changes by two kinds of sensitivities, the elasticities and the response or control coefficients. The *elasticities* describe the enzyme in isolation and are defined as the derivatives of a kinetic rate law with respect to the concentrations or kinetic constants appearing in the rate law directly. A variant of them, the scaled elasticities, are derivatives on logarithmic scale and refer to relative rather than absolute changes. As shown in Figure 4, the elasticities and stoichiometric coefficients together determine the local dynamic around a reference steady state: in a linear approximation, the fate of a metabolic perturbation vector Δc is described by a linear differential equation system $\mathrm{d}\Delta c/\mathrm{d}t = M\Delta c$ with the Jacobian matrix $M = N\,E$ where E is the elasticity matrix. These equations describe how small metabolic perturbations will propagate within the network.

Response and control coefficients

The second type of sensitivities, the response and control coefficients, describe the steady state response of concentrations and fluxes to an initial perturbation. Instead of tracing individual causal chains or relying on numerical simulations, MCA treats the perturbations in a linear approximation, which leads to relatively simple formulae. The *response coefficients* describe how fluxes or concentrations will respond to parameter perturbations anywhere in the system. The concentration control coefficients C^{S}_{il} are a normalised version of the response coefficients: they are defined as $C^{\mathrm{S}}_{il} = R^{\mathrm{S}}_{il}/E_l$ where the l^{th} enzyme parameter affects the l^{th} reaction with the elasticity E_l. Sum rules for metabolic control coefficients, called summation and connectivity theorems [Heinrich and Schuster, 1996], bring out relations between the influences of enzymes along stationary flux modes or around a common metabolite.

Global system dynamic and causality

A main reason to use mathematical models – and not just statistical data analysis and biological intuition – is that the dynamic of complex systems is hard to grasp. Analytical reasoning would consider all bits and pieces one at a time and trace their interactions via causal chains. However, this hardly helps to understand global phenomena in complex systems – not only because all their details and quantitative properties are hard to keep in mind, but because their dynamics are so closely entangled. If a system contains feedback loops or just reversible reactions, an infinite number of causal chains would be needed to explain the simple adjustment of a steady state exactly. Metabolic control theory, in contrast, accounts for all dynamic interactions at the same time and directly describes stationary changes, allowing us to to simulate complex global behaviour.

2.2 How kinetic models are built

Model building

Building a kinetic model entails a few big decisions - for instance, choosing the mathematical formalism - and a lot of work on the details. A typical modelling project comprises the following steps: choosing the biochemical elements and setting up the network; determining the formulae, for instance, the kinetic rate laws; determining the parameter values from literature, by fitting, or by optimisation; and possibly, repeating parts of this process for sub-sampled data sets (for cross-validation) or model variants (for model selection). For the different steps,

data on network structures, thermodynamic and kinetic constants, as well as transcript, protein, metabolite, or flux data are collected from the literature, from databases, or from new experiments.

Modelling software

In contrast to traditional paper-and-pencil models, most models today are constructed with the help of computers. Numerical calculations for single enzyme kinetic started as soon as computers were available [Chance, 1943, Chance et al., 1952]. Today, tools like COPASI [Hoops et al., 2006] or CellDesigner [Funahashi et al., 2008] and formats like SBML (Systems Biology Markup Language [Hucka et al., 2003]) help to build, edit, and exchange models efficiently. There are two contrary modelling approaches: in bottom-up modelling, we prepare parts of models, for instance, individual rate laws, and put them together. In the top-down approach, we start from a global, coarse-grained description and gradually refine it – for instance, by filling a metabolic network with kinetic constants. Both approaches could be supported by automatic workflows, which permit to execute modelling steps one after the other. For automating more and more of these steps, many models are published in computer-readable formats and enriched with information about the biochemical details. This makes it easier to build complex models graphically, to modify existing models, and to fill them with data retrieved from databases or publications on the internet.

Bottom-up and top-down modelling

2.3 Elements of biochemical network modelling

With the kinetic modelling formalism at hand, setting up model equations and solving them numerically should be routine work. However, creating meaningful models may take a lot of trial and error. Even if models look simple, they are based on many unexpressed considerations which ensure that the models make sense biologically and meet their purposes. Building a model requires physical and biological knowledge, sensible compromises between accuracy and limited data, and ideas about the biochemical details that are purposefully neglected because they do not play a role under the circumstances considered. While all this is important for manual modelling, it becomes even more important in computer-assisted modelling. Here, all considerations need to be made explicit and safe modelling formalisms have to be developed to compensate for the lack of intuition. Let us now step back for a moment and see what kinds of knowledge and thoughts go into the development of models.

Considerations in modelling

1. **Biochemical knowledge and data.** Biochemistry tells us about enzymatic rate laws, about the elements of biochemical networks like small compounds, enzymes, mRNA, genes, transcription factors, or kinases, and about their ways to interact. To model specific pathways, we need to know their structure (e.g., a stoichiometric matrix and regulation arrows) and kinetic constants (e.g., from databases like Brenda [Schomburg et al., 2004]), and experimental data can be used to fit the model parameters or to test model predictions.

2. **Laws of physics.** Mechanistic models employ physical principles like mass balance or thermodynamics, which are often directly implemented in the formalism. General physical laws are of great help because they provide information without requiring any specific measurements.

3. **Model assumptions.** Reality is continuous, dynamic, complex, and not exactly known, while our language and reasoning – which allow us to build models – are discrete, and assertive. Accordingly, models are not meant to describe cells in all details, but to depict specific processes the way we see them. Compared to what is known about cells, models are extremely simplified: when building a model, we select a system in question and isolate it from other processes that exist, but are not modelled. Through simplifying assumptions, we can delimit the system in space (e.g., neglect a cell's environment), focus on specific time scales (e.g., neglect faster and slower processes), and disregard other biochemical processes.

4. **Uncertainty.** Models do not describe reality as it is, but what we know about reality. Since our data are limited, our mental pictures of cells remain uncertain. Subjective uncertainty can be quantified by

probability distributions and Bayesian statistics provides a rigorous framework to handle them: instead of insisting on a single model, we assess how well different models are supported by data, and a statistics over such models can tell us which aspects of models are likely to be trustworthy.

5. **Noise, variability, sensitivity, and robustness.** Unlike their idealised pictures in textbooks, living cells are full of randomness and due to chemical noise, genetic variation, and changing environments, every cell is different. In models, randomness can be captured by stochastic models. Such models do not predict a single, deterministic behaviour, but allow for deviations, which are quantified by probabilities. Randomness, sensitivity, and robustness are closely related, and control coefficients and other measures of sensitivity can give a sense of how randomness will spread and affect measurable quantities.

6. **Biological function.** Mechanistic models describe how a biochemical system behaves and explain how this behaviour is implemented biochemically. But we can also ask *why* biochemical systems are set up the way they are. Why do many organisms share the same metabolic pathways? Is there a reason for the specific values of enzyme parameters? Such questions about the function of biological systems are actually questions about evolution. One way to address them in models is to assume that biochemical systems are optimised to perform certain actions – e.g., to grow fast – under certain biochemical and physical constraints. This leads to optimality-based models like flux balance analysis and, on the level of interacting cells, to evolutionary game theory [Pfeiffer and Schuster, 2005].

7. **Sustainable modelling.** For a long time, published models tended to be incomplete and hard to re-implement. Since models are a way to communicate knowledge, the form in which they are provided is important. To contribute to future large-scale models, researchers should make sure that their models can be easily reusable, modified, and incorporated into more comprehensive models.

2.4 Methods for computer-assisted modelling

Genome-scale metabolic models

One of the big challenges in systems biology is to develop large kinetic models of metabolism. Genome-scale network reconstructions (e.g., [Herrgård et al., 2008]) are produced routinely and almost automatically from genome sequences. The next step is to turn such networks into kinetic models, i.e., to determine their rate laws and kinetic constants. Genome-scale kinetic models would help to integrate high-throughput data, to check their consistency, and to predict the global effects of local perturbations by mutations, drugs, or differential expression. Constraint-based models, which are already obtained from network reconstructions, cannot fully answer these questions.

Currently, building genome-scale kinetic models is difficult because many rate laws and corresponding parameters are unknown; because parameter estimation is numerically hard; because the pathways of interest are surrounded by larger environments, which cannot be easily modelled; and because the combination of existing models can lead to inconsistencies. These problems need to be solved before large kinetic models can be built routinely. To develop methods and software for model building and combination, we need to understand the concepts and the intuition of modellers. We have to consider how modellers would solve certain tasks, what kind of information they would need, and how these procedures could be formalised. Contemplating this can help to avoid mistakes that a researcher would easily recognise, but that are still hard to spot for software.

How do modellers think?

During the past years, I have developed solutions for some of the issues above (see Table 1). As a guiding principle, the approaches were meant to be generic, scalable to large systems, and safe, to make them easily usable in automatic workflows and large-scale kinetic modelling. The methods fall into three main categories: kinetics and data integration (section 3); elasticity sampling and metabolic control analysis (section 4); and automatic modelling (section 5).

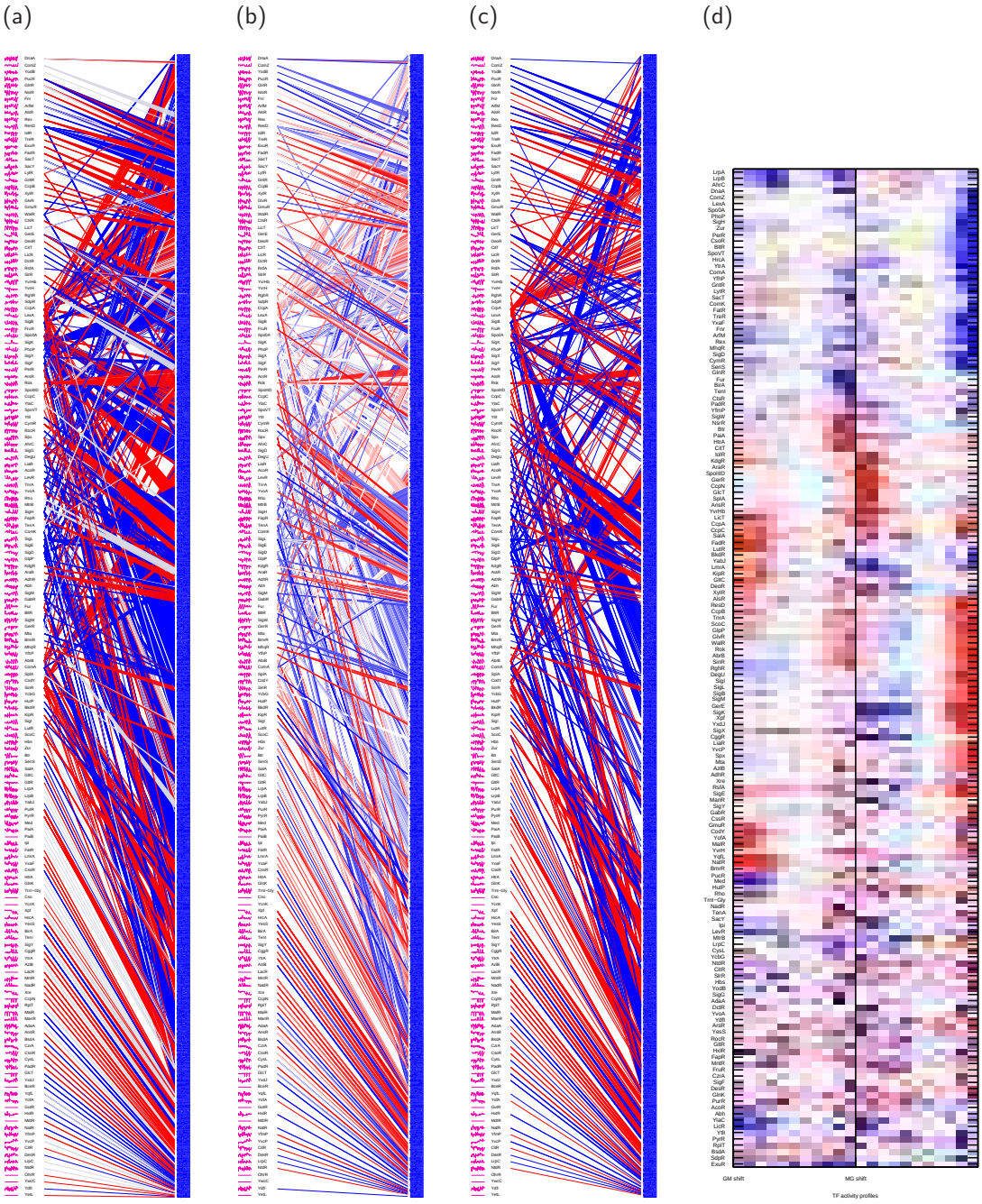

Figure 5: Model of transcriptional regulation in *Bacillus subtilis*. (a) The transcription network contains 158 transcription factors (left) and 1754 genes (right), connected by 2900 regulation arrows, indicating activation (blue), repression (red), or unknown modes of action (grey). I built the network based on information from [Goelzer et al., 2008], DBTBS [Sierro et al., 2008], SubtiWiki [Florez et al., 2009], and ChIP/chip experiments [Buescher et al., 2012]. (b) Influence weights (blue: positive, red: negative, shaded: weak influences) were assigned to the arrows by Network Component Analysis. The expression data stem from metabolic shift experiments (addition of malate to glucose-grown bacteria cultures and *vice versa*). (c) Confirmed part of the network. By omitting edges that contribute little to the data fit, I obtained a relevant subnetwork for this experiment. (d) Temporal activities of transcription factors determined by Network Component Analysis. For each transcription factor (rows), activity profiles from both shift experiments are shown. Figures from [Buescher et al., 2012].

3 From kinetic data to dynamic metabolic models

3.1 Assigning numbers to the arrows

Network component analysis

To build genome-scale biochemical models, we need to define a network structure, fill it with rate laws, and determine their parameters. An example of this approach is Network Component Analysis (NCA), a method to parametrise large transcription networks [Liao et al., 2003]. NCA is based on a simple model of transcriptional regulation in which each gene promoter has an input function of the power-law form $x_i(b) \sim \prod_l b_l^{a_{il}}$. After converting the expression data to logarithmic scale, this becomes a linear model $\ln x_i(b) = \sum_l a_{il} \ln b_l + \text{const.}$ of transcriptional regulation. I used NCA within an experimental study of metabolism in *B. subtilis* [Buescher et al., 2012] to build a large-scale model of transcriptional regulation (Figure 5). In the experiments, the carbon source malate was added to glucose-grown bacteria cultures and *vice versa*. I estimated the influence weights a_{il} between 158 transcription factors and 1754 genes from expression patterns measured during two metabolic shifts. Most of the expression data were well explained by the transcription network and about half of the edges could be discarded because they contributed little to the data fit. Furthermore, the signs of some edges (activation or repression) were initially unknown and could be determined by NCA. Among the estimated transcription factor profiles (see Figure 5 (d)), the most prominent activity changes were found in regulators of central metabolism (compare the scheme in Figure 1). Their activities and their roles in regulating glycolysis, gluconeogenesis, and the citric acid cycle are shown in Figure 6.

From networks to models

Since this approach works well for transcription networks, we may be tempted to apply it to metabolic networks as well. However, this brings some additional challenges. First, metabolic models do not only cover one type of data – expression data – but various kinetic and thermodynamic constants, metabolite concentrations, fluxes, and protein levels. Second, metabolic networks carry global fluxes, which have to satisfy mass balance and thermodynamic laws and as a consequence, different model parameters will constrain each other. To account for these constraints, we may use a stoichiometric model as a frame, determine consistent sets of equilibrium constants or other parameters in this network, and impose them on the rate laws. Third, to obtain sensible models, we may need to include reactions for which no data are available, so we have to deal with missing knowledge and with uncertainties.

Modelling workflow

A possible workflow for filling metabolic networks with rate laws consists of four steps [Borger et al., 2007]: (i) choose a metabolic network structure; (ii) choose rate laws; (iii) collect data about kinetic constants; (iv) determine a consistent parameter set that agrees with the data, and assess its uncertainties. Models created in this way will contain reasonable parameter values, satisfy a number of safety and quality criteria, and may serve as a starting point for further detailed modelling. To realise this workflow, I developed a number of concepts and tools, which will be explained in the following.

3.2 Standard kinetic rate laws

Unknown rate laws

Every enzyme has its individual rate law. Its details depend on the binding affinities between enzyme and reactants, on the enzyme mechanism, and possibly on activation or inhibition by small molecules. To translate a metabolic network into a kinetic model, we need to determine individual rate laws for all reactions. Many enzymes have been studied in detail, but for all others, standard rate laws have to be assumed. Standard rate laws should capture a variety of cases, for instance, different stoichiometries and different combinations of allosteric effectors. Since existing rate laws were too specialised (Michaelis-Menten rate for uni-uni reactions) or do not include enzyme saturation (mass-action, power-law [Savageau, 1969], or lin-log kinetics [Visser and Heijnen, 2003]), I developed the convenience kinetics [Liebermeister and Klipp, 2006a], which is saturable and applicable to any reaction stoichiometry. By including more kinetic variants, I later obtained the modular rate laws [Liebermeister et al., 2010],

Standard rate laws

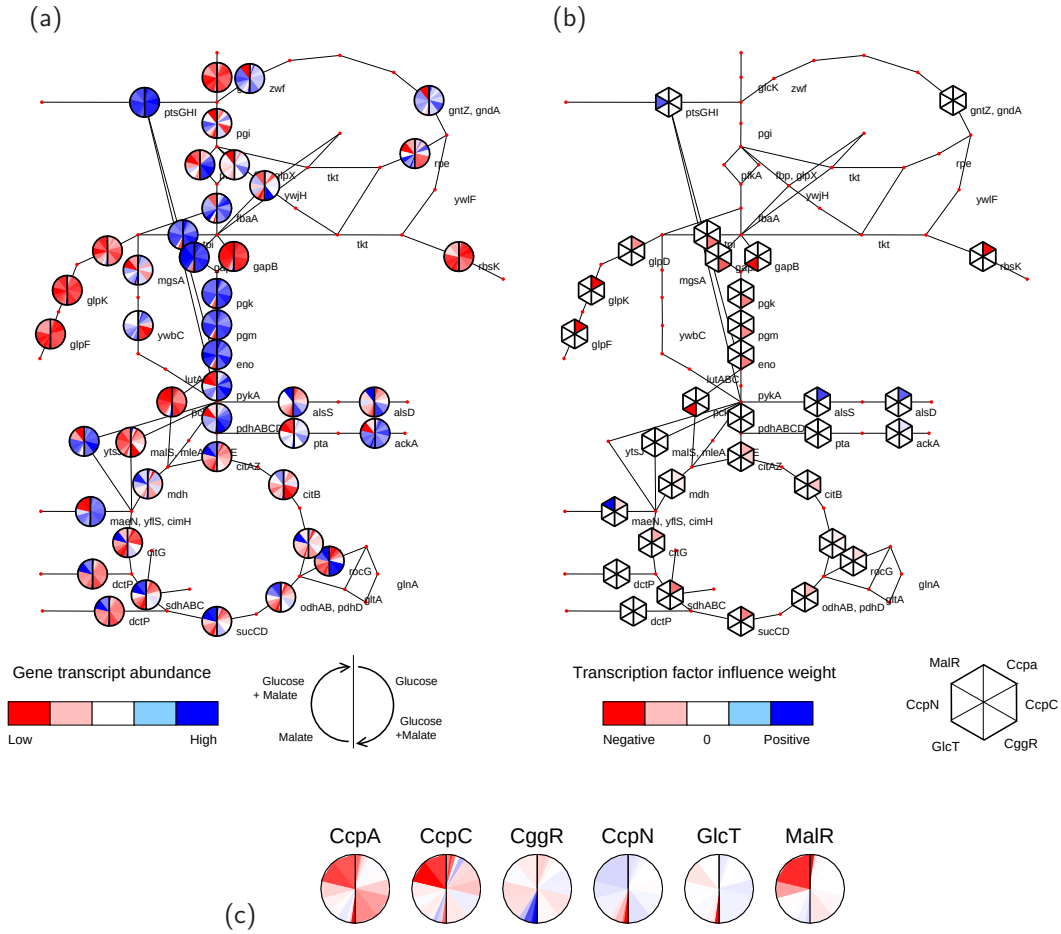

Figure 6: Transcriptional regulation of central metabolism in *B. subtilis*. (a) Gene expression during metabolic shifts, measured using tiling arrays. The dynamics during opposite shifts (addition of malate to glucose-grown cultures and *vice versa*) are shown in the halves of each circle. Time runs in clockwise direction. (b) Influences of the transcription factors CcpA, CcpC, CggR, CcpN, Glct, and MalR (compare Figure 1) determined by Network Component Analysis. (c) Temporal activities of the transcription factors. During growth on malate, lower glycolysis is repressed by CggR. As soon as glucose becomes available, the repression is relieved and CcpN starts to suppress gluconeogenesis via gapB. Figure from [Buescher et al., 2012].

a family of rate laws that can be flexibly adjusted to many situations. They are simple and biochemically plausible, thermodynamically sound, and can be automatically inserted into SBML models.

Common modular rate law

The idea behind these rate laws is that many detailed rate laws – like ping-pong or fixed-order bi-bi-kinetics – share a mathematical form that resembles reversible Michaelis-Menten kinetics Eq. (5); they just differ by their denominators, which depend on the enzyme mechanism. To obtain a simple generic rate law, I chose a new denominator that corresponds to a simple random-order enzyme mechanism. The resulting formula for the common modular rate law [Liebermeister et al., 2010] generalises Michaelis-Menten kinetics to bimolecular and other reactions with only a few kinetic constants. For a bimolecular reaction A + B $\leftrightharpoons$ 2 C without allosteric regulation, it reads

$$v = u\,\frac{k^{\mathrm{cat}+}(a/k_{\mathrm{A}}^{\mathrm{M}})(b/k_{\mathrm{B}}^{\mathrm{M}}) - k^{\mathrm{cat}-}(c/k_{\mathrm{C}}^{\mathrm{M}})^2}{(1+a/k_{\mathrm{A}}^{\mathrm{M}})(1+b/k_{\mathrm{B}}^{\mathrm{M}}) + (1+c/k_{\mathrm{C}}^{\mathrm{M}})^2 - 1}. \tag{7}$$

Like in the Michaelis-Menten kinetics (5), there is a Michaelis constant k_x^{M} for every reactant and a catalytic constant $k^{\mathrm{cat}\pm}$ for each reaction direction. Additional terms for regulation (not shown) will contain activation or

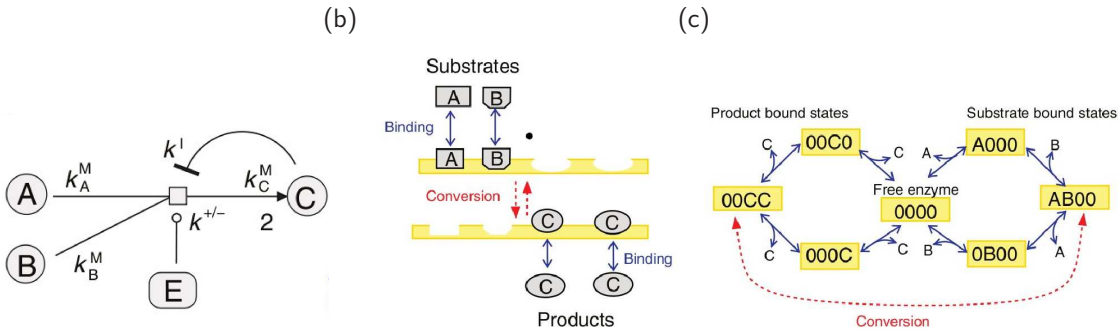

Figure 7: Common modular rate law. (a) Reaction scheme for a reaction A + B ⇋ 2 C catalysed by an enzyme E. The kinetic constants used in the common modular rate law are shown. (b) The enzyme mechanism behind the common modular rate law includes substrate binding (top), chemical conversion (centre), and product release (bottom). All steps are reversible; substrate and product molecules cannot be bound at the same time. (c) Binding states of the enzyme. Letter codes show which binding sites are occupied in each state (yellow boxes). The seven states correspond to the seven summands arising in the denominator of Eq. (7). Figure from [Liebermeister et al., 2010].

inhibition constants for allosteric effectors. The reactant C has a stoichiometric coefficient of 2, which appears as an exponent and leads to a sigmoid rate law similar to a Hill equation [Hill, 1910]. The formula can be easily generalised to arbitrary stoichiometries [Liebermeister et al., 2010]:

$$v = u\,\frac{k^{\mathrm{cat}+}\prod_i \left(\frac{c_i}{k_i^{\mathrm{M}}}\right)^{m_i^+} - k^{\mathrm{cat}-}\prod_i \left(\frac{c_i}{k_i^{\mathrm{M}}}\right)^{m_i^-}}{\left(1+\frac{c_i}{k_i^{\mathrm{M}}}\right)^{m_i^+} + \left(1+\frac{c_i}{k_i^{\mathrm{M}}}\right)^{m_i^-} - 1}. \tag{8}$$

Here the exponents m_i^+ and m_i^- are the (positive) stoichiometric coefficients of substrates and products, multiplied by a factor h. This factor resembles a Hill exponent, can be chosen individually for each reaction, and has a default value of 1. The enzyme mechanism behind Eq. (7) is shown in Figure 7 (b). Substrates and products bind rapidly, in random order, and independently; substrate and product molecules cannot be bound at the same time. The Michaelis constants k^{M} in Eq. (7) arise from the dissociation constants of the elementary binding steps, while the catalytic constants $k^{\mathrm{cat}\pm}$ are rate constants for the conversion of substrates into products. Of course, if unknown rate laws are replaced by this standard rate law, there will be approximation errors. However, replacing a ping-pong and fixed-order mechanism [Liebermeister and Klipp, 2006a] would yield errors of similar size, so if no specific enzyme mechanism is known, using the common modular rate law will be just as plausible as using more detailed rate laws.

Enzyme mechanism

Considering other binding mechanisms, I obtained a number of alternative rate laws (see Table 3.2) following the general scheme

$$v = f\,\frac{T}{D + D^{\mathrm{reg}}}. \tag{9}$$

Different mathematical expressions to be inserted for the terms f, T, D, and D^{reg} lead to different modular rate laws adapted to various situations. They cover several types of allosteric regulation and five types of denominators which correspond to different mathematical simplifications of the enzyme mechanism (see Figure 8 and Table 3.2). Finally, each rate law can be turned into a sigmoid kinetics by choosing an exponent h different from 1. However, all modular rate laws have the same thermodynamic properties, which are determined by the numerator T and make them fit into the Thermodynamic-Kinetic Modelling (TKM) formalism [Ederer and Gilles, 2007]. In particular, they can be split into a product of factors related to kinetics and thermodynamics [Liebermeister et al., 2010].

Thermodynamic properties

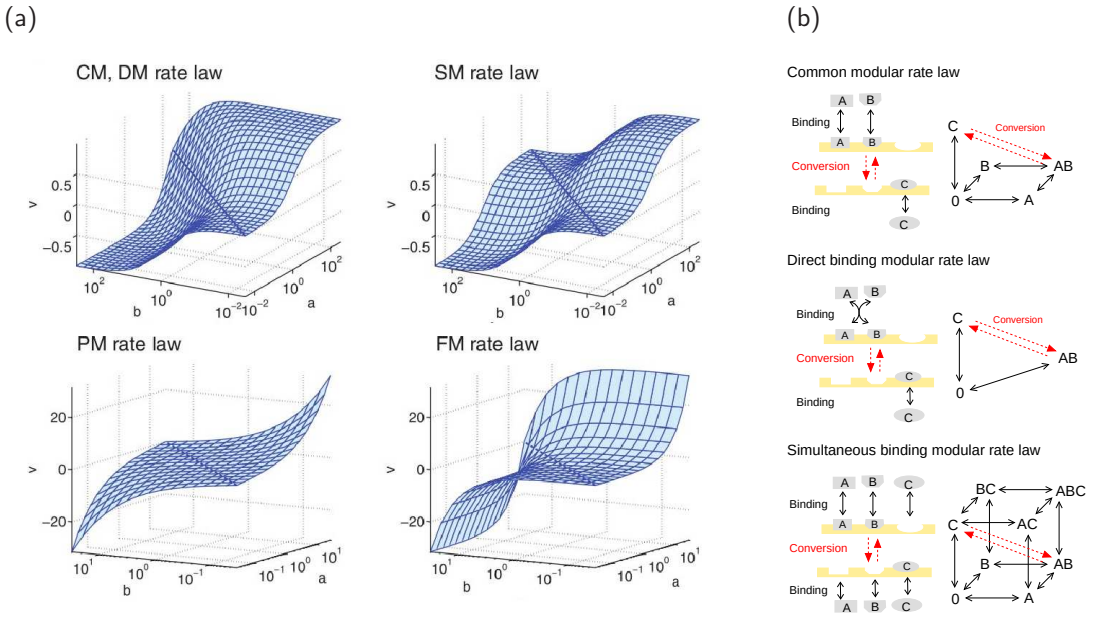

Figure 8: Modular rate laws and their saturation properties. (a) Modular rate laws for a reaction A $\leftrightharpoons$ B. The rate (vertical axis) depends on the substrate and product concentrations a and b. For details about rate laws and their full names, see Table 3.2. (b) Enzyme mechanisms behind modular rate laws shown as in Figure 7. Figures from [Liebermeister et al., 2010].

Name	Description
Convenience kinetics	Generalised reversible Michaelis-Menten kinetics
Common modular (CM) rate law	Generalised reversible Michaelis-Menten kinetics
Power-law modular (PM) rate law	Generalised mass action or power law
Direct binding modular (DM) rate law	Single-step substrate binding (and product release)
Simultaneous binding modular (SM) rate law	Reactant contribute multiplicative factors
Force-dependent modular (FM) rate law	Reaction rate depends solely on the thermodynamic force

Table 2: Types of modular rate laws and their distinguishing properties. For more details, see [Liebermeister et al., 2010]

A factor $\sinh\left(\frac{hA}{2RT}\right)$ describes the effect of the thermodynamic force (with reaction affinity A, Boltzmann gas constant R, and absolute temperature T), while the kinetic factor represents substrate availability, enzyme saturation, and regulation. Thus, the thermodynamic force does not only determine the flux direction, but has a quantifiable effect on the rate.

Usage in SBML

The name *modular rate law* emphasises the fact that the terms in Eq. (9) can be chosen individually and that the laws themselves act as building blocks for metabolic models. Compared to other standard rate laws like lin-log or power-law kinetics, the modular rate laws are biochemically more plausible and can be better adjusted to specific situations. Because of their generic form, the modular rate laws can be created directly from the stoichiometric matrix or from a graphic scheme of the network as in Figure 7 (a). In practice, they can easily be inserted into SBML models (Systems Biology Markup Language [Hucka et al., 2003]) with the tool SBMLsqueezer [Dräger et al., 2008] and with my tool semanticSBML [Krause et al., 2010] (`www.semanticsbml.org`). Most of the necessary information is directly extracted from the SBML elements, their attributes, and SBO terms (Systems Biology Ontology [Le Novère, 2006]) in the model code. Some decisions (e.g., the type of rate law used or specific types of allosteric regulation) are left to the user or taken by default rules. Known or estimated kinetic constants can be provided in the standard table format SBtab, which I developed for this purpose (see `www.semanticsbml.org`). Aside from their direct usage in models, the modular rate laws also provide simple

formulae for reaction elasticities, which are useful for sampling algorithms. This will be the topic of section 4.

3.3 Constraints between model parameters

Parameter dependencies

When parametrising a transcription network, we can choose and change the input functions of different genes independently. In metabolic networks, in contrast, the kinetic constants in different rate laws may be dependent due to physical laws. Whenever we change, for instance, a Michaelis constant, other parameters will be affected because they are coupled by Haldane relationships (see paragraph 2.1). Furthermore, the equilibrium constants in a model will have to satisfy Wegscheider conditions [Wegscheider, 1902]. If a kinetic model is parametrised by Michaelis constants and catalytic constants, these constants are effectively coupled through the equilibrium constants, and such dependencies have to be respected when fitting, optimising, or sampling the parameters. This is a challenge for modelling.

For automatic model building, I had to address these dependencies in detail. The first step is to state them in an explicit and simple form. The Haldane relationships are a good example: by introducing the velocity constant $k^{\mathrm{V}} = \sqrt{k^{\mathrm{cat}+}\, k^{\mathrm{cat}-}}$ and using the Haldane relationships and Eq. (6), the catalytic constants can be expressed in terms of the Michaelis constants, the velocity constants, and the standard chemical potentials. If we convert all kinetic constants to logarithmic scale, the formula is linear. In fact, many relevant parameters display such linear dependencies and can be therefore be arranged in dependence schemes [Liebermeister and Klipp, 2005] and [Lubitz et al., 2010]: one group of basic parameters can be freely chosen, while all remaining derived parameters are computed from them using linear equations; some of the parameters need to be expressed on logarithmic scale. If we use such a scheme to describe a thermodynamically feasible state, it will contain standard chemical potentials and concentrations as basic parameters; and equilibrium constants, chemical potentials, and reaction affinities as derived parameters. To obtain a full kinetic parameter set, we can further include Michaelis constants and velocity constants (basis parameters) and catalytic rate constants (derived parameters). In any case, the dependence scheme can be written in the compact form

Dependence schemes

$$\begin{pmatrix} \theta^{\mathrm{ind}} \\ \theta^{\mathrm{dep}} \end{pmatrix} = \begin{pmatrix} I \\ R \end{pmatrix} \theta^{\mathrm{ind}} \tag{10}$$

with vectors θ^{ind} (basic parameters) and θ^{dep} (derived parameters). The dependence matrix R depends only on the network structure [Liebermeister and Klipp, 2006a, Lubitz et al., 2010] and on the choice of parameters to be described. The dependence scheme drastically simplifies parameter estimation because it allows us to use linear regression and multivariate normal distributions for the entire parameter set. As we shall see in a moment, this helps to trace parameter uncertainties and correlations in complex metabolic models and to integrate various kinds of data by using Bayesian parameter estimation.

3.4 Uncertain or variable model parameters

Cells handle variability

Variability and uncertainty are ubiquitous in biology. What counts in evolution is not only how cells succeed in an ideal constant environment, but also how they deal with unforeseeable events. Accordingly, cells are much more adaptive than man-made machines and are prepared to deal with failures or impreciseness inside their own machinery. Sensitivity and robustness can be implemented biochemically by special network structures or by regulation systems. In models, uncertainty and variability play similar roles and can be described by random distributions [Wang et al., 2004, Liebermeister and Klipp, 2005]. Kinetic constants and other parameters will be uncertain if measured values contain errors or are not available. If parameters vary within a cell population, the parameters for a randomly picked cell will also be uncertain. To describe uncertainty mathematically, we may replace the parameters in our model (e.g., the level of a certain enzyme) by random variables (e.g. describing the

Uncertainty and variability

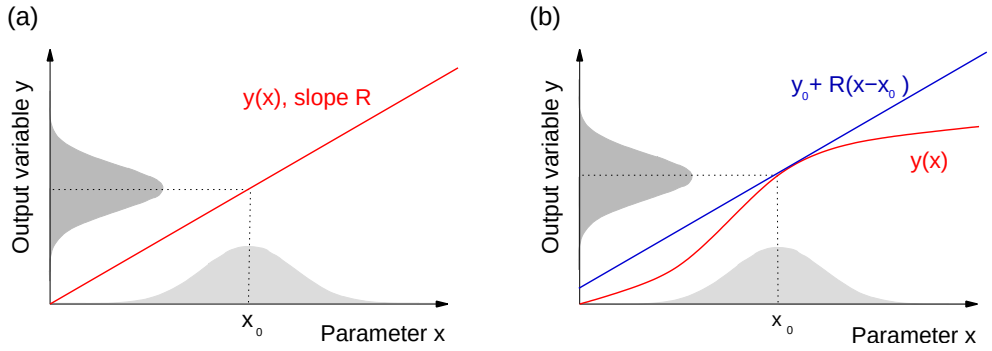

Figure 9: Propagation of parameter uncertainties. (a) A parameter y ("output variable") depends on a parameter x, which follows a normal distribution. If the dependence between them is linear, y will be normally distributed as well. (b) Nonlinear dependencies can be linearly expanded around a reference value x_0. Figure from *Systems Biology – a Textbook*.

variability of this enzyme level in the cell population), defined by probability distributions.

Propagation of uncertainty

For dealing with uncertainty and variability, and for later use in Bayesian parameter estimation, I characterised these distributions in more detail. Since the kinetic constants are dependent, the parameter vector has to be described by a multivariate distribution. With Eq. (10), this is not difficult: if the basic parameters in the vector θ^{ind} (on logarithmic scale) follow a multivariate normal distribution $\mathcal{N}(\bar{\theta}, C_\theta)$, the elements of θ^{dep} will also be normally distributed and follow the distribution $\mathcal{N}(R\,\bar{\theta}, R\,C\,R^\top)$. As shown in Figure 9 (a), the variance of a dependent parameter increases with its sensitivity to the basic parameter. The same principle also applies to other quantities, e.g., steady-state concentrations computed from the model, which may depend on the parameters in a non-linear way. If these response functions are Taylor-expanded around a typical reference parameter set, we can approximate their mean values and variances from the slopes and curvatures of the response function. The covariance matrix of a steady-state quantity vector y reads, to first order,

$$\mathrm{cov}(y) = R^y_x \,\mathrm{cov}(x)\, {R^y_x}^\top \tag{11}$$

where $\mathrm{cov}(x)$ is the covariance matrix of the parameter vector x and the matrix R^y_x contains the response coefficients between x and y. For nonlinear response curves, second-order formulae explain how the mean values of quantities in a cell population differ from the corresponding values obtained from the mean parameter set [Liebermeister and Klipp, 2005].

Correlated metabolite levels

Sensitivities and variability in biochemical systems are tightly connected, and this fact has some remarkable consequences. On the one hand, the propagation of variability by dependence schemes can explain correlated metabolite levels [Steuer et al., 2003]. In a linearised model, the metabolite correlations will depend on three factors: the network structure and the reaction elasticities, which together determine the response matrix R^y_x, and the covariance matrix $\mathrm{cov}(x)$ of the perturbation parameters. On the other hand, cells may evolve specific sensitivities to deal with variability: if an input x varies strongly and an output y needs to remain stable, the response coefficient between them should be small, which will favour certain network structures. Due to trade-offs between different robustness properties, natural selection will only realise those that are most important. This may leave its traces in the structure of biochemical pathways, their regulation systems, and the resulting response matrices. To understand how variances, robustness, and metabolite correlations arise in a given network structure, we need to understand how the network structure determines the response coefficients. The response coefficients do not arise from the network structure alone, but also on the reaction elasticities, which depend on the kinetic rate laws and on the metabolic state in question. However, the stoichiometric and allosteric structures predetermine the possible response coefficients to a good extent. In Section 4, I shall present methods to assess

Uncertainty and robustness

this in detail.

3.5 Parameter distributions for metabolic models

Choosing the parameter distribution

We saw that uncertain model parameters follow a multivariate joint distribution and that correlations between them arise from the dependence scheme. However, how should such distributions be chosen in practice? If we disregard for a moment all parameter dependencies, we just have to determine a distribution for each parameter individually. In Bayesian statistics, a probability distribution summarises what we can know about a certain quantity based on prior knowledge and data, which are both uncertain. If the shape of the distribution is unknown, we may choose it by the principle of minimal information [Jaynes, 1957]. This principle states that our distribution should represent as little information as possible in the sense of Shannon entropy [Shannon, 1948] and for distributions with predefined mean value and variance, it leads to the normal distribution. If we consider the kinetic constants on a logarithmic scale and treat them by normal distributions, their dependencies will be linear (given by Eq. (10)) and their linear combinations will again be normal; on non-logarithmic scale, such parameters will follow log-normal distributions.

Priors from machine learning

How can we determine the mean values, variances, and covariances of these distributions? First of all, they can represent known values and their uncertainties in data. If we have no data for some parameter, we may consider a statistics over parameters of this type. For an unknown Michaelis constant, for instance, we may consider the distribution of Michaelis constants in the Brenda database [Schomburg et al., 2004], possibly broken down to specific enzyme classes. If we have additional information about out model element, e.g. about molecule structures, we may use it to estimate more specific values, which will narrow down the distributions even further. Along these lines, I used machine learning, based on linear regression and analysis of variance, to determine distributions of individual kinetic constants [Borger et al., 2006] and physiological metabolite concentrations [Liebermeister, 2005a].

Given all this, how can we now account for parameter dependencies? Most simply, we could determine independent distributions for the basic parameters and compute the distributions of all derived parameters by using the dependence scheme. However, this will not permit to use data for the derived parameters, e.g., measured catalytic constants, which can provide useful information also about other parameters. In fact, to determine a joint distribution of all model parameters, based on evidence from all kinetic data collected for a model, we may use Bayesian estimation. A practical procedure, called "parameter balancing", is described in the following paragraph.

3.6 Parameter balancing

Kinetic data are inconsistent and incomplete

Kinetic constants for a metabolic model can be collected from the literature, obtained from databases like Brenda [Schomburg et al., 2004] or Sabio-RK [Wittig et al., 2006], fitted to dynamic data, optimised for other objectives, or simply be guessed. If measured values are directly inserted into a model, the parameter set can be incomplete, redundant, and contradictory. Thus, the parameters have to be adjusted to comply with the constraints in our dependence scheme. While this seems to complicate modelling at first sight, parameter constraints also have an advantage because they reduce the number of free parameters and permit to derive unknown parameters from known ones.

Parameter balancing

To put this idea into practice, I developed parameter balancing, a method to determine complete, consistent parameter sets that resemble known kinetic and thermodynamic data [Liebermeister and Klipp, 2006b, Lubitz et al., 2010]. Parameter balancing combines all methods described in this section: the modular rate laws, the dependence scheme, and parameter distributions. To obtain balanced parameters, the dependence scheme is translated into a linear regression model, which is then treated using Bayesian statistics. General expectations

about the parameter ranges are formulated as prior distributions. Since priors cannot be formulated for derived parameters, for instance, the equilibrium constants, these parameters are controlled by pseudo-values, which we introduced for this purpose in [Lubitz et al., 2010]. Priors and pseudo-values help to keep the results in meaningful ranges even if few data are available. Due to the linear dependence scheme, the method is applicable to relatively large models.

Model ensembles

Parameter balancing does not only yield a fixed parameter vector, but a posterior probability distribution, which is multivariate normal (for logarithmic parameters) and describes the typical parameter values, their uncertainties, and the correlations between different parameters. As I proposed above, the posterior describes exactly what can be known about the parameters from data, constraints, and prior expectations. Aside from matching model parameters to data and typical distributions, it may also be important to restrict some of them by lower or upper bounds. Parameter balancing can include bounds on single parameters and linear inequalities for parameter combinations. With such constraints, our multivariate normal posterior becomes restricted to a region in parameter space defined by the inequality constraints.

Accounting for flux data

The main limitation of parameter balancing is that, due to the mathmatical form of the the rate laws, metabolic fluxes do not fit into the dependence scheme. Therefore, parameter balancing cannot be used to adjust a model directly to given steady state fluxes. However, there are two possible solutions:

1. Even if the flux values do not fit into the scheme, we can ensure that the balanced parameters will at least comply with the flux directions. Each non-zero flux defines the sign of a reaction affinity, which can be prescribed in parameter balancing, and the balanced parameters and concentrations will realise the correct flux directions. Once this has been accomplished, the enzyme levels and velocity constants can be rescaled to match exactly the predefined fluxes; this is guaranteed to work if the flux distribution is thermodynamically feasible.

2. Alternatively, a posterior distribution obtained from parameter balancing can be reused as a prior for a subsequent Bayesian parameter estimation, for instance, to fit a full kinetic model to flux and concentration data [Liebermeister and Klipp, 2006b]. However, the nonlinear rate laws will lead to non-Gaussian posterior distributions and finding the posterior mode becomes a nonlinear optimisation problem, which can be hard to solve for larger models.

When inserting balancing parameters into kinetic models, we can either use the posterior mode – representing a consistent and most probable parameter set – or parameter sets randomly sampled from the posterior. By creating an ensemble of such random models and assessing their dynamic properties, we can learn about the potential behaviour of such models, given all quantitative information we used during parameter balancing.

Tool for parameter balancing

Along with the publication [Lubitz et al., 2010], we implemented the entire process of parameter balancing as a free software in python. Given an annotated SBML model and a table of known thermodynamic, kinetic, and dynamic data, we can run parameter balancing interactively at `www.semanticsbml.org`. In summary, parameter balancing is a convenient way to integrate large amounts of kinetic, thermodynamic, and dynamic data, but it does not directly permit to fit flux data. Moreover, the numerical effort for genome-scale models would be large. Thus, to build larger kinetic models with predefined steady states, we need a different kind of approach in which fluxes and concentrations can be predefined and the sampled parameters agree with them by construction. Such an approach is presented in the following section.

4 From flux distributions to dynamic in large metabolic networks

4.1 Flux analysis and kinetic models are converging

Network reconstructions

A variety of metabolic networks, comprising thousands of metabolites and reactions, have been reconstructed semi-automatically from genome sequences [Henry et al., 2010]. Turning these networks into kinetic models would help to put high-throughput data into context and to understand the global effects of differential gene expression, enzyme inhibitors, and genetic modifications. So far, kinetic models are much smaller and cover only a tiny fraction of the available data, but the gap between network reconstructions and kinetic models is diminishing. Using thermodynamic analysis, we can find consistent flux distributions, metabolite concentrations, and equilibrium constants. When building kinetic models, we may choose kinetic constants that realise this predefined steady state. The dependence schemes from section 3 show that this is possible: we may first choose the Michaelis constants and equilibrium constants and then use parameter balancing to determine catalytic constants that satisfy the Haldane relationships. Finally, the catalytic constants can be rescaled to yield the required flux.

Dynamic close to steady state

On the contrary, some analyses of kinetic models do not even require the full rate laws, but solely the reaction elasticities. If we linearise a model around a stationary reference state, the resulting linear model will describe the dynamic for small perturbations such as periodic perturbations of enzyme levels, varying nutrient supply, or chemical noise. Another potential application is the analysis of optimal differential expression [Liebermeister et al., 2004]. The linearised model, defined by the stoichiometric matrix and the elasticity matrix, is much easier to handle mathematically than the original model. In particular, there is a powerful theory for periodic perturbations, noise, control, and model reduction.

Elasticities and stoichiometries

Thus, if the steady state fluxes, concentrations, and reaction elasticities were known, MCA and many other analyses could be applied even to very large systems. Unfortunately, to determine the elasticities, we need a kinetic model. However, some information about them can be obtained from the network structure alone: in models without allosteric regulation, the reaction rates increase with the substrate level and decreases with the product level, and so a given flux distribution will define the signs of all elasticities. As I will explain below, the elasticities are further biased by the thermodynamic forces. Thus, even if the elasticities of a model are not exactly known, the network structure and thermodynamics determine them to some extent, and this allows us to explore the dynamic possibilities of metabolic systems even ithout knowing their exact kinetics.

4.2 Model ensembles and structural kinetic modelling

Monte Carlo sampling

The dynamic of biochemical systems close to steady state including their responses to differential gene expression can be predicted by simulations or sensitivity analyses of kinetic models. The results will depend on the network structure, the kinetic rate laws, and the particular steady state. But what is the contribution of each of these factors? To study this, we may vary the system parameters and create an ensemble of models, each with the same model structure, different kinetic constants drawn from a distribution; for each instance of the model, qualitative or quantitative features of interest are evaluated. Features that are constant throughout the simulations can be attributed to the network structure, while variable features obviously depend on the parameters. Together with colleagues, I used this Monte Carlo approach to study the signs and values of control coefficients [Klipp et al., 2004] and the conditions for spontaneous oscillations [Borger et al., 2005]. With the same approach, one could also compare different versions of a model (e.g., variants of the network structure) and check them for systematic differences that stand out from the variability caused by the random parameters.

Parameter sampling

Unfortunately, the models in such ensembles will have different steady states, which may be hard to compare and also hard to compute. This problem is addressed by Structural Kinetic Modelling (SKM, [Steuer et al., 2006]),

in which stationary fluxes and concentrations are predefined and the scaled elasticities are sampled from random distributions. However, a basic problem of SKM is that the elasticities are sampled independently. We saw above that this assumption leads to thermodynamically incorrect models. As an example, let us consider a reaction A $\rightleftharpoons$ B with reversible mass-action kinetics $v = k^{\mathrm{f}}\, a - k^{\mathrm{r}}\, b$. The scaled reaction elasticities read $E_{\mathrm{A}} = k^{\mathrm{f}}\, a/v$ for substrate A and $E_{\mathrm{B}} = k^{\mathrm{f}}\, b/v$ for product B. No matter if the elasticities are large or small, their difference will be the same: $E_{\mathrm{A}} - E_{\mathrm{B}} = 1$. If we sample the elasticities independently, this relationship will be violated and the sampled values will not correspond to a correct mass-action model. Similar restrictions hold for any reversible rate law.

Structural kinetic modelling

4.3 Elasticity sampling

To overcome this problem, we have to acknowledge that elasticities are dependent and account for it in the sampling algorithm. Moreover, since the specific dependencies vary from rate law to rate law, we have to choose specific rate laws whose elasticities we understand. In fact, the elasticity formulae for the modular rate laws [Liebermeister et al., 2010] lead to a thermodynamically consistent algorithm for elasticity sampling: the scaled elasticities (see Section 3.2) can be split into three terms:

Elasticity formulae

$$E_{c_i} = E_{c_i}^T - E_{c_i}^D + E_{c_i}^f. \tag{12}$$

The thermodynamic term $E_{c_i}^T$ can be computed directly from the reaction Gibbs free energy, while the kinetic term $E_{c_i}^D$ and the regulation term $E_{c_i}^f$ depend on how strongly the enzymes are saturated with reactant or effector molecules. This can be expressed by saturation parameters which compare the concentrations of metabolites to the corresponding Michaelis, activation, or inhibition constants. The saturation parameters can be chosen without any constraints. By sampling them independently and computing the elasticities by Eq. (12), we obtain an elasticity matrix that corresponds to a specific, thermodynamically consistent kinetic model.

Thus, we obtain an algorithm for thermodynamically correct elasticity sampling: after choosing the network structure (comprising both stoichiometries and enzyme regulation), we determine the steady-state fluxes, concentrations, and reaction Gibbs free energies by a thermodynamic flux analysis. Afterwards, the saturation values are sampled independently from the range $]0, 1[$ and the scaled reaction elasticities are computed by Eq. (12). Using the fluxes and concentrations, we can further compute the unscaled elasticities, the Jacobian matrix, the response coefficients, and various dynamic properties of the system.

Algorithm for elasticity sampling

The modular rate laws also provide formulae for the curvatures [Liebermeister et al., 2010]: the second-order elasticities, defined by second derivatives of the rate laws, describe interactive effects that two arguments of a rate law – parameters or concentrations – have on the rate. An interactive effect is not the immediate effect of a double perturbation, but the difference of this effect and the effect that would be expected from the effects of separate single perturbations. The second-order elasticities permit to compute second-order response and control coefficients for stationary [Höfer and Heinrich, 1993] and periodic perturbations [Liebermeister, 2005b]. First-order response coefficients show how a small increase of each single enzyme level would increase or decrease the rate of biomass production. The second-order control coefficients, in contrast, approximate their interactions, for instance, synergisms or antagonisms between enzyme-inhibiting drugs.

Second-order elasticities

Thus, elasticity sampling allows us to apply Metabolic Control Analysis to models of unknown kinetics. All input information can be obtained from the network structure and from thermodynamic analysis, while the remaining uncertainties can be determined from the spread of the sampled results. In the rest of this section, I shall discuss three applications that go beyond usual MCA: the response to periodic perturbations, the propagation of noise, and the construction of effective dynamic models by linear model reduction.

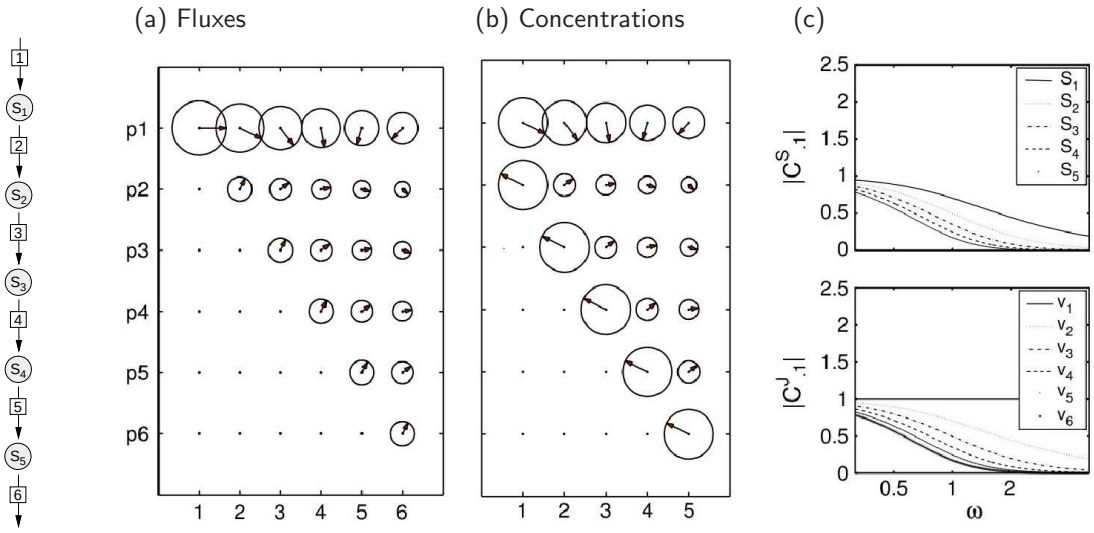

Figure 10: Driven oscillations in a linear metabolic pathway (shown on the left). (a) A periodic enzyme level leads to oscillations in the catalysed reaction, which then affect all downstream reactions. Each row of circles shows the effect of one oscillating enzyme (p1 through p6). The reaction rates (from left to right) display harmonic oscillations of different amplitudes (circle radii) and phase angles (arrows), forming a damped wave. Due to the irreversible rate laws, the perturbations spread only downstream. (b) Simultaneous oscillations of metabolite concentrations. For each enzyme perturbed, the substrate of the perturbed reaction shows the highest amplitude and almost the opposite phase. (c) Amplitudes of concentrations (top) and reaction rates (bottom), in response to oscillations of the first enzyme, at varying circular frequencies ω (x-axis). With increasing frequency, the oscillations become smaller and wash out earlier in the pathway: the pathway acts as a low-pass filter. Figure from [Liebermeister, 2005b].

4.4 Biochemical systems under periodic perturbations

Periodic dynamics

Periodic phenomena are common in nature. Spontaneous oscillations have been observed in various biochemical systems and in mathematical models, they are often explained by stable limit cycles. A prominent example are the glycolytic oscillations in yeast, which appear spontaneously and can become synchronised between cells. If a cell contains a biochemical oscillator or lives in a periodic environment, the driven oscillations will spread like waves in the biochemical network. To complement the existing works about spontaneous biochemical oscillations, I studied the fate of such driven oscillations based on the first- and second-order elasticities [Liebermeister, 2005b]. The basic phenomenon, the transmission of oscillations from reaction to reaction, is shown in Figure 10. In a metabolic pathway with mass-action kinetics, an alternating enzyme level leads to a periodic accumulation and depletion of the product and, as a consequence, to phase-shifted oscillations of all downstream rates and metabolite concentrations.

Driven oscillations

The mathematical description resembles metabolic control analysis. If we consider small perturbations around a steady state, the model dynamic can be linearised and we obtain a system of linear equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = A\,x + B\,u \tag{13}$$

with system variables x (deviations of metabolite concentrations from their reference values) and perturbation parameters u (deviations of other parameters, e.g., external concentrations). The matrices A and B can be computed from the stoichiometric matrix and the elasticity matrices. Similar linear equations appear in mechanics and electrical engineering and are studied in control engineering. In the frequency domain, their dynamic is described by a response function, the Fourier transform of the pulse-response function, which in our case can be seen as a generalised metabolic response coefficient [Ingalls, 2004, Liebermeister, 2005b]. Since the perturbations

Spectral response coefficients

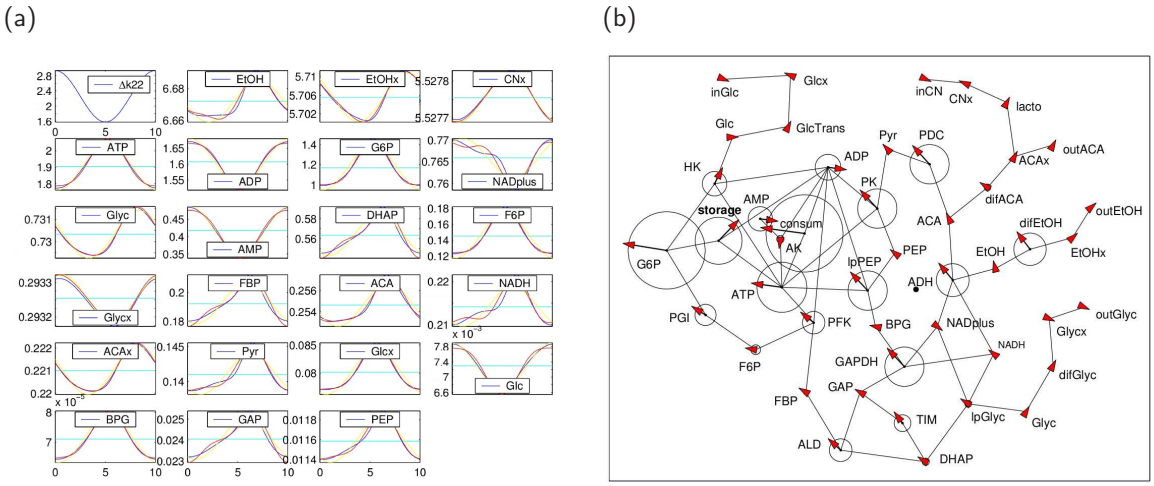

Figure 11: Driven oscillations in a model of yeast glycolysis [Hynne et al., 2001], caused by periodic changes of the external glucose concentration. (a) Periodic concentration changes, approximated using spectral response coefficients. Metabolites are shown in separate boxes, the time (on the x-axis) covers one oscillation period. The curves were obtained from numerical integration (blue) and from approximations of order zero (cyan), one (yellow), and two (red). (b) Amplitudes (arrow lengths) and phase shifts (angles) of concentration changes (first-order response coefficients) shown on the network. Figure redrawn from [Liebermeister, 2005b].

are now periodic, the responses can show phase shifts, the spectral response coefficients are complex numbers and depend on the driving frequency. In [Liebermeister, 2005b], I also introduced second-order response coefficients to capture interaction effects between pairs of harmonic perturbations.

Once we know how harmonic oscillations spread in the linearised model, we can also predict the fate of other perturbations (Figure 11) and noise (Figure 12) using Fourier transformations: we transform the original perturbation profile into the frequency domain, multiply it with the spectral response coefficient, and transform it back to the time domain. For systems with nonlinear rate laws, first- and second-order spectral response coefficients yield good approximations of the true oscillatory dynamic. Figure 11 shows an example from [Liebermeister, 2005b], a model of yeast glycolysis with an alternating supply of glucose. The driven oscillations, approximated using first- and second-order spectral response coefficients, agree well with the curves obtained by direct numerical integration. However, the quality of the approximation decreases as the amplitude of the perturbations becomes larger.

Simulation by Fourier synthesis

4.5 Propagation of chemical noise

The spectral response coefficients can not only describe driven oscillations, but also the propagation of noise. Since reactions between molecules are random events, the dynamic of macroscopic concentrations is not strictly deterministic as assumed in kinetic models. On the contrary, the molecule numbers jiggle around a mean curve which – in the case of a linear system dynamic – coincides with the deterministic curve from our kinetic model. As shown by Gillespie [Gillespie, 2000], this stochastic dynamic can be approximated by a stochastic Langevin equation, which is basically a kinetic model with white noise terms added to the forward and reverse reaction rates. Since the noise level grows proportionally with the square root of the mean reaction rate (in units of reaction events per second), the relative impact of noise will increase for smaller particle numbers; if the mean rates become too small, the approximation breaks down.

Chemical noise

Chemical Langevin equation

In [Liebermeister, 2005b], I studied the fate of temporal fluctuations in metabolism in more detail. In the steady state of a Langevin equation model, the particle numbers will not be fixed, but fluctuate around the deterministic

Spreading of noise

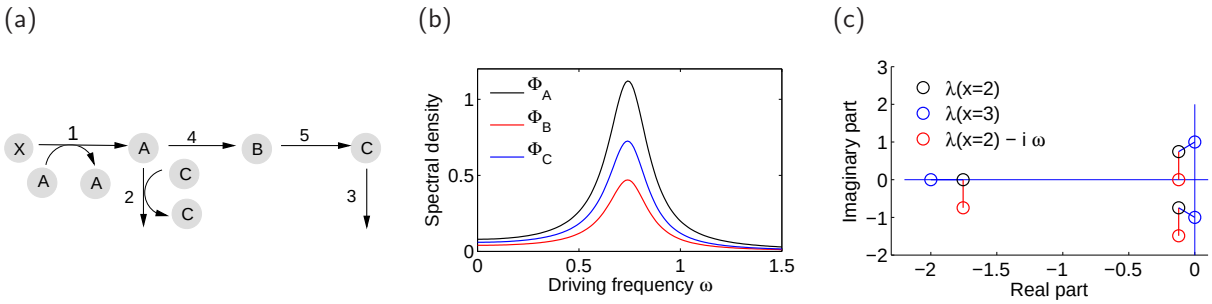

Figure 12: Propagation of noise in a small biochemical model, the minimal biochemical system with Hopf bifurcation [Wilhelm and Heinrich, 1995]. (a) Network scheme (circles: compounds; arrows: reactions). As the level of external substrate X exceeds a critical value, the levels of A, B, and C start to oscillate spontaneously. At subcritical values, oscillations can appear in the form of intermediate noise fluctuations. (b) Spectral densities of the metabolite fluctuations caused by chemical noise. Although the system does not oscillate spontaneously, the noise spectrum shows already a resonance peak close to the oscillation frequency. (c) Eigenvalues of the Jacobian matrix shown in the complex plane. Solutions for three values of the bifurcation parameter are marked by different colours. The complex eigenvalues close to the imaginary axis are related to the resonance. Figure from *Systems Biology – a Textbook*.

steady state of the kinetic model. The fluctuations arise from white noise in individual reactions, spread in the network, and are damped depending on their frequencies. To quantify the resulting noise in metabolite concentrations, I described it by its frequency-dependent spectral density. If the spectrum of the original chemical noise is flat, as implied by the white-noise term, the spectral density of the metabolite concentrations arises from the spectral response coefficients. These coefficients decrease at high frequencies, and so the systems acts as a low-pass filter. Moreover, if a system has a tendency to oscillate, resonances may amplify the noise at specific frequencies [Liebermeister, 2005b] (see Figure 12). This approach does not only apply to chemical noise, but also to small fluctuations caused by a cell's environment.

4.6 Black-box models built by linear model reduction

Feedback through the environment

Biological systems are always embedded in larger environments that cannot be fully captured by models. A biochemical pathway, for instance, is part of a metabolic network. An increased flux through the pathway will affect fluxes in the network and these perturbations will feed back into the pathway. It would be desirable to capture such effects in models, but including the entire network just to provide dynamic boundaries for a small pathway may be too much of an effort. One would rather like to replace the network by a simple effective model that mimics its interactions with the central pathway. Such a model can be a black box, i.e., its mathematical equations may have any shape, as long as it shows the right input-output behaviour.

Reduced network models

How can we obtain such effective models? One option is to derive empirical relationships from data fits. In [Liebermeister et al., 2005], I developed an alternative approach: if there exists a kinetic models of the surrounding network, we can approximate it by a linear dynamic model. Using model reduction methods such as balanced truncation [Moore, 1981], such models can then be replaced by low-dimensional linear models with a similar input-output behaviour. The interface between the pathway of interest and surrounding environment model consists of communicating metabolites and reactions, which are shared by both subsystems. A sensible subdivision, with communicating metabolites on the side of the pathway and communicating reactions on the side of the environment, is shown in Figure 13.

In practice, employing an effective environment model will only make sense if it improves the simulations. In my tests with small models, this was the case, even if it was build from purposefully perturbed environment models [Liebermeister et al., 2005]. But what if only a network structure, but no kinetic model of the environment, is available? In this case, sampled elasticities could be used again to create instances of the environment model. Since

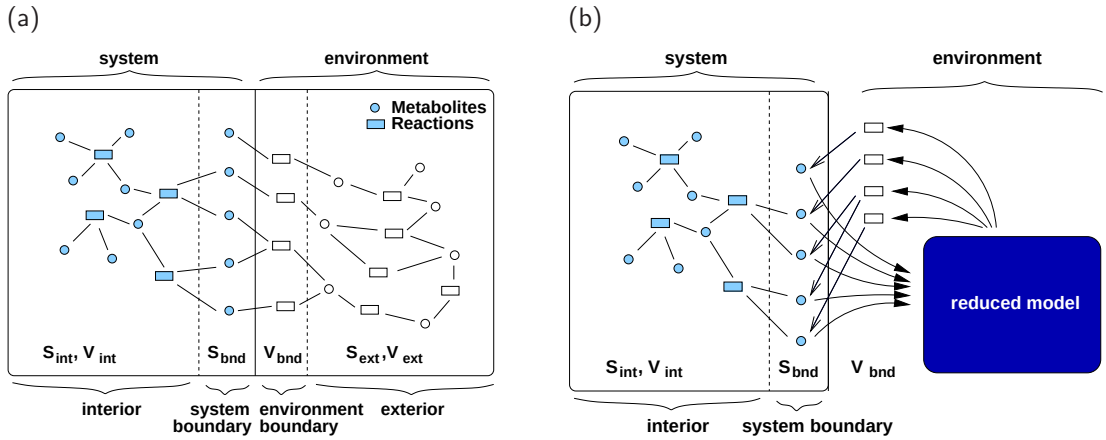

Figure 13: Model reduction applied to biochemical network models. (a) A metabolic system (schematic example) is split into a pathway of interest (left half) and its environment (right half). Both subsystems interact via communicating metabolites (belonging to the pathway of interest) and reactions (belonging to the environment). (b) To facilitate numerical simulations, the environment is replaced by a reduced black-box model. Figure from [Liebermeister et al., 2005].

the linearisation by itself causes a relatively large approximation error, the error from using sampled elasticities may play a minor role. Further improvements can be expected from nonlinear model reduction, but this is still a big challenge in control engineering. Using the methods described so far, we may build pathway models of various size, determine their parameters from data, and connect them to larger models. But how can models be combined in practice if they stem from different sources and have not been specifically prepared? This practical side of model construction is the topic of the following section 5.

5 Sustainable model building

5.1 Automatic model processing

With increasing amounts of high-throughput data, systems biology models are becoming larger and more precise and may eventually cover all central cell functions. However, the final aim is not necessarily a single comprehensive model for each type of cell: quite the contrary, different questions and applications require models of different resolution and extension and with a focus on different processes. Furthermore, various models and model variants are needed to study the metabolism of newly sequenced species or the effects of genetic variants. All this requires that models be flexibly built, modified, and recombined by software. To move from traditional paper-and-pencil models to computer-assisted modelling, the systems biology community needs to establish three things:

Flexible model combination

1. **A technical infrastructure** for processing models automatically, comprising standard formats for models and data; central repositories for curated models; biochemical annotations of model elements; biochemical databases and ontologies, i.e., formal representations of biological concepts and their relationships; methods to process models automatically; and software tools that make all this easy to use.

2. **Theoretical concepts** to ensure that automatically built models will be biochemically meaningful. This includes concepts for the automatic processing of semantic information, but also safe model formulations that capture the knowledge and intuition of modellers about the mathematical, physical, and biochemical aspects of models [Liebermeister, 2008].

3. **A culture of sustainable modelling**. Standard formats help to edit models, to avoid double efforts of reimplementing and debugging models, and to build models that are easily reusable. Such sustainable modelling will be a key virtue in future systems biology modelling, as we pointed out in [Krause et al., 2011]. Modellers should anticipate that their models will be reused, possibly as parts of larger cell models. This is not only a matter of technology, but also of personal awareness and organisation. It concerns modelling itself, the design of experiments, and collaboration within research projects. For instance, standard operating procedures, established at the start of a project, can make experiments and data and models compatible and therefore more useful.

Community efforts have resulted in standard formats like the Systems Biology Markup Language (SBML) and software like CellDesigner [Funahashi et al., 2008] or COPASI [Hoops et al., 2006]. SBML is developed as a community effort and supported by more than 200 software tools. Within the past ten years, it has become the standard format for systems biology models and its success has triggered other developments like the MIRIAM guidelines [Le Novère et al., 2005] (Minimum Information Requested In the Annotation of Models), the model repository BioModels Database [Novère et al., 2006], and the Systems Biology Graphical Notation [Novère et al., 2009], a standard for depicting biochemical network models.

SBML and annotations

SBML represents a model neither as a pure biochemical networks nor as a pure mathematical equation system, but as a mixture of both. While some biochemical facts are encoded in the XML structure (e.g., the links between reactions and their reactants), the meaning of the elements (e.g. the identity of the compounds) is specified by additional semantic annotations. Such annotations point to entries in public web resources, e.g., the ChEBI database for biochemical compounds [Degtyarenko et al., 2008] and allow software to process models by their biochemical meaning. Modelling tools provide users with simple and abstract views of their models, allowing them to access further details only when required. Future software tools could allow users to retrieve data and models from the internet and to treat them as building blocks for other models. In this section, I describe concepts and software that I developed to support such a modular modelling.

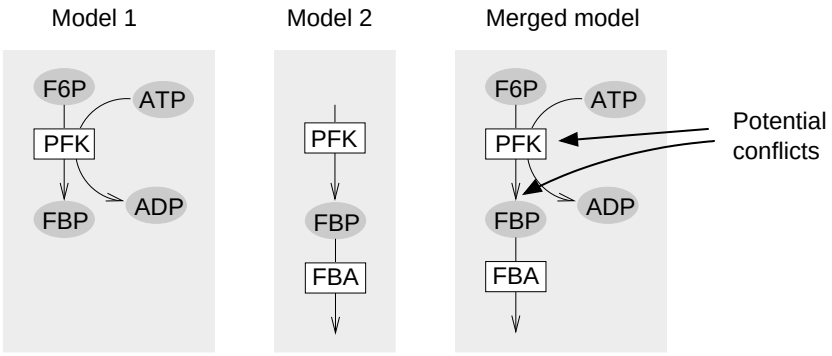

Figure 14: Model merging. Two simple network models (left and centre) share the elements phosphofructokinase (PFK) and fructose 1,6-bisphosphate (FBP). When merging the models, the duplicate elements have to be joined to avoid redundancies in the merged model (right). If duplicate elements have different features (e.g., concentrations for a compound, rate laws for a reaction), these conflicts need to be resolved either manually or by applying default rules. Figure from [Liebermeister, 2008].

5.2 Model merging

Model merging

Model merging is not only important in practice, but also exposes some of the general problems in automatic modelling. We do not even have to consider models that are based on different mathematical formalisms; even joining models that belong to one formalism can be a challenge. Biochemical models from the literature, which we intend to merge, may overlap in complicated ways, so we need to find out how model structures can be made to agree. I outlined some basic problems of model merging in [Liebermeister, 2008]. In principle, we can regard each model as a list of compounds and reactions, each one with a set of properties: stoichiometries and rate laws, for instance, can be listed as properties of the reactions. To merge the models, we join their element lists, find duplicate elements, replace them by single elements, and translate the resulting list into a syntactically valid model.

Problems in model merging

This looks relatively straightforward. However, what is simple for a human modeller may be difficult for software. First, models may use different naming schemes, so elements should not be matched by their names, but by their biological meaning as declared by formal annotations. Second, there may be duplicate elements with conflicting descriptions (see Figure 14), and resolving them automatically can be difficult. Other problems arise from conflicting model assumptions or systems described at different levels of resolution.

Software semanticSBML

Automatic model merging is an important long-term goal, but still hard to achieve. As a step in this direction, my colleagues and I developed an editor for interactive merging of SBML models. SBMLmerge [Schulz et al., 2006], which later became a part of semanticSBML [Krause et al., 2010], automatically aligns models, detects inconsistencies between model elements, allows the user to change the alignment and to resolve the conflicts, and returns the merged model in SBML format. Relatively early on, it became clear that the models have to contain semantic annotations to be safely merged. Since assigning these annotations is tedious, we developed the tool *SBMLannotate*, which allows users to easily edit annotations. Further tools for automatic searches, alignments, clustering, and merging of SBML models were added [Krause et al., 2010, Schulz et al., 2011] (`www.semanticsbml.org`). SemanticSBML is freely available at www.semanticsbml.org as a stand-alone python software and as a web application.

5.3 Comparisons between biochemical models

Element comparison

Element comparisons are a basic task in automatic model processing and crucial to align models, to cluster them by similarity, and to search for models resembling or overlapping a given model. Figure 15 shows an example, the

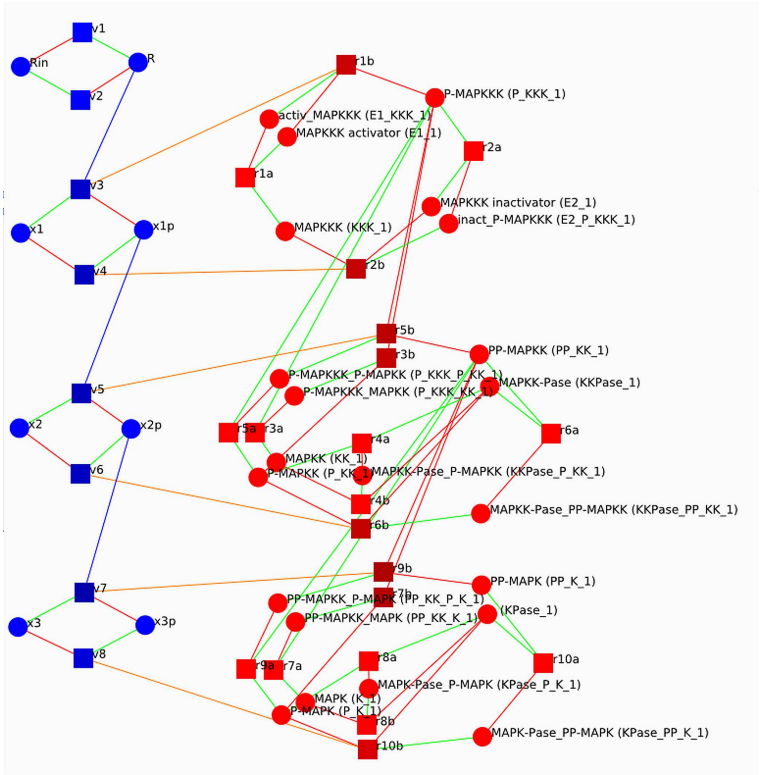

Figure 15: Alignment between two mitogen-activated kinase (MAPK) models from [Hornberg et al., 2005] (BioModel 84; left, in blue) and [Huang and Ferrell, 1996] (BioModel 9; right, in red). MAP kinase cascades, a common type of signal transduction pathways, consist of proteins that transmit signals by phosphorylating the following protein in the pathway. Elements were matched automatically between models by semanticSBML. Network nodes stand for species (circles) and reactions (squares). Orange lines connect corresponding elements between models. Figure from [Schulz et al., 2011].

alignment between two MAP kinase cascade models. After matching similar elements, the resemblance between both models is easy to see despite their different resolution. Together with colleagues, I developed a number of similarity scores for models and individual model elements [Schulz et al., 2011]. The key question – which is characteristic for clustering and information retrieval in general – is what aspects of a thing are considered in the similarity measure. Since my main aim was to align overlapping pathway models, model had to be similar if they shared biochemical elements. Other aspects of the models, like the mathematical formulation or numerical parameter values, were disregarded, but they could become relevant in other contexts.

Similarity measures

Similarities between model elements are defined as follows: we look up the biological concepts mentioned in their annotations, use ontologies to connect them by paths, and compute a similarity score from the semantic relationships along these paths. The similarity scores for models fall into three groups, depending on how much information is taken into account: some scores only count which biological concepts are shared between models; other scores compare individual elements by their annotations and align them between models; more advanced scores use also the network structure to infer similarities between non-annotated elements. In an evaluation with benchmark models, already simple model comparisons by feature vectors – which only count the appearance of biochemical concepts in a model – yielded very good results. To make these methods applicable, we included the similarity scores into semanticSBML, where they are now used for model search, model clustering, and for initial model alignments in model merging. All these tasks can be performed online at `www.semanticsbml.org`. As examples, a model clustering and the results of a model search are shown in Figure 16.

(a) (b)

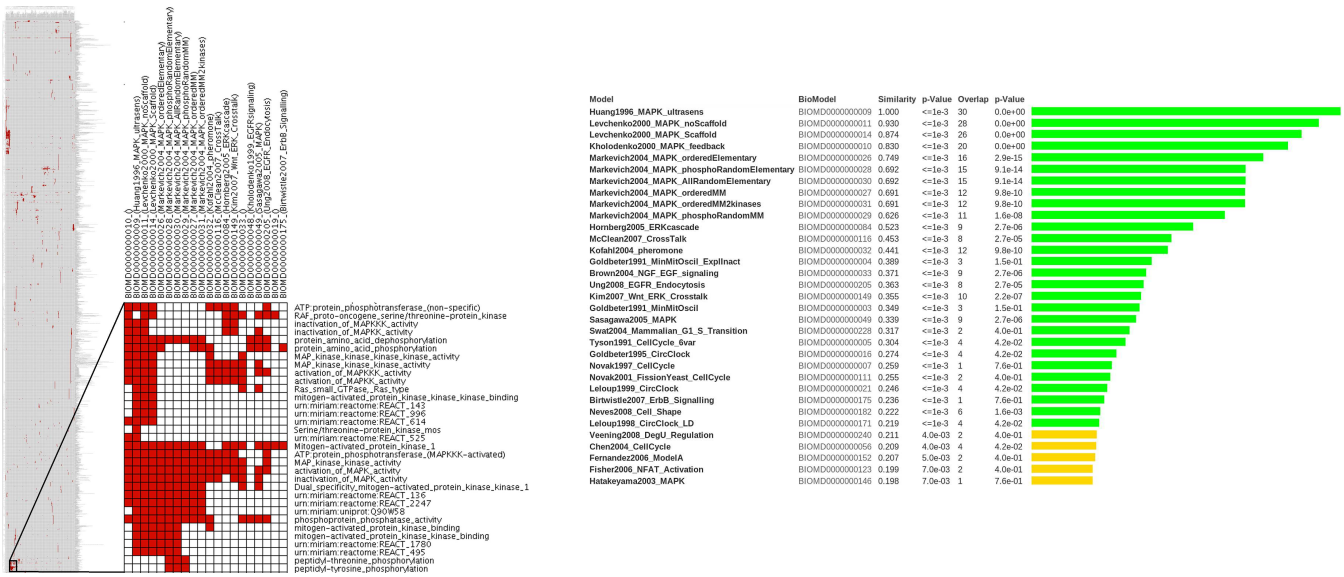

Figure 16: Clustering and ranking of systems biology models from BioModels Database. (a) Clustering of models and biochemical concepts. The annotation matrix (left) shows which biological concepts (rows) appear in which models (columns). After sorting the matrix by hierarchical two-way clustering, similar concepts or models appear close to each other. The close-up shows a selection of MAP-kinase models and the concepts typically associated with them. (b) Semantic model search. Starting with the MAP-kinase model from [Huang and Ferrell, 1996] (BioModel 9) as a query model, all models from BioModels Database were ranked by similarity, shown by bar sizes. On the top hits are shown. The first hit is the query model itself; the following high-scoring models also describe MAP kinase cascades. Figure from [Schulz et al., 2011], created using the online tools at `www.semanticsbml.org`.

5.4 What is a valid model?

Model validity

Validity criteria

Detecting conflicts between models can be difficult, especially if the models are large. To develop automatic checks and safe model formalisms, we need to state clearly what we mean by conflicts and by incorrect models. Models are never correct in the sense that they depict reality in each and every detail. George Box formulated this as "All models are wrong, but some are useful." [Box and Draper, 1987]. In practice, a model will be acceptable if it meets its purposes and the modeller's expectations, which vary widely from case to case: energetic balancing, for instance, may be necessary for some metabolic models, but not so relevant in signal transduction models. Therefore, we do not need an absolute criterion for model validity, but a catalogue of criteria that can be applied depending on the circumstances and that can be tested by software. As a first step, I classified such criteria in [Liebermeister, 2008]: not surprisingly, conflicts can arise on various levels, each requiring specific validity criteria, and testing them automatically requires different types of information (for an overview, see Table 3).

Criteria	Test requires:	Example conflict
Syntax	SBML syntax rules	References to undefined model elements
Computation	Equation system encoded in model	Equations under-determined
Semantics	Semantic annotations and ontologies	Negative Michaelis-Menten constants
Physics	Checks for model equations	Mass balances violated
Biochemistry	Data about realistic values	Unrealistically high concentrations
Facts about organism	Databases / network reconstructions	Genes not present in organism

Table 3: Classification of validity criteria for systems biology models. Making validity criteria explicit can help to develop tests for meaningful models and to devise safe algorithms for model merging. For a detailed description, see [Liebermeister, 2008].

Once validity criteria have been defined, we may attempt to ensure them by safe model formulations and merging procedures. Translating the criteria into a formal representation will enable software to process not only the models themselves, but also the users' expectations about models. This can be used for sanity checks and for suggesting ways to resolve conflicts. On the one hand, conflicts can be detected after models have been modified or recombined. On the other hand, they may be avoided from the start if the original models have been built from common schemes, for instance, from a cellular network at a certain standard resolution. Some of the validity criteria – for instance, compliance with physical laws – can be implemented in the mathematical formulation: for instance, parametrising models by chemical standard potentials, as proposed in section 3, will keep them thermodynamically correct even if parameter values are changed or models are combined. In fact, some basic validity criteria are already guaranteed by the structure of SBML: certain mistakes (e.g., a missing initial concentration value) can be detected as syntax errors. Thus, a clear understanding of model validity will help not only to design safe merging algorithms, but also to develop data formats whose syntax ensures model validity.

Model checks

Safe model formulations

6 Outlook

6.1 Biochemical models in the future

One decade after systems biology has been announced as a new field of research, most of the complexity in cells is still unexplored. However, high-throughput experiments and network reconstructions are advancing fast and models are becoming larger, more detailed, and more reliable. Comprehensive models of cells and organisms, which are envisaged for the future, would open a broad range of applications. Simulations could help to predict the effects and interactions of drugs and to account for genetic variants in personalised therapies. In genetic engineering, models could help to find genetic modifications that will improve the production of commodity chemicals.

Increasingly complex models

How are we advancing towards this goal? With increasing computer power and more kinetic and dynamic data to come, flux analysis and kinetic models may finally converge. Thermodynamic and stoichiometric models may serve as blueprints for combining kinetic rate laws into genome-scale, dynamic models of metabolism. In such models, different pathways may be modelled to different resolution and quality depending on their importance for the model and on the available data. The economic use of enzymes, a key aspect in flux prediction, will be described more quantitatively, and observed enzyme levels and their regulation may be explained by economic principles. At the same time, microscope images, time-lapse movies, and other single cell data will bring new challenges for data analysis. Stochastic and spatial models will not only be used to model the experimental results, but also to interpret raw data and to extract hidden information from them. Generally, modelling will become an integral part of data analysis and experimental design and will evolve from a separate research field into a standard technique of biological research.

Towards automated modelling

The methods presented here can simplify model building and help to automate some of its steps. In particular, they allow us (i) to translate metabolic networks into dynamic models, integrating various kinds of data; (ii) to insert plausible rate laws and kinetic constants that agree with thermodynamic laws; (iii) to determine elasticities for large metabolic networks and to assess the global system response to single and double perturbations, periodic perturbations, and noise; (iv) to study the roles of network structure, thermodynamics, and enzyme kinetics for dynamic behaviour and control properties; (v) to construct effective models by applying model reduction to low-quality, large-scale models; and (vi) to search, match, and combine existing models. The theoretical concepts help to design good model formalisms and data structures, and this can change the way modellers work. An extension of kinetic modelling to large networks and an easy mapping between data and models through annotations will enable modellers to take advantage of high-throughput data and will bridge the gap between data analysis and dynamic simulation. Automated routine steps and workflows can help modellers to abstract from model details and to develop a high-level perspective on models, allowing them to concentrate on the more creative aspects of their work.

6.2 How models can help to understand life

Mechanistic biochemical models

"How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" Schrödinger's question remains puzzling. If our aim is to find faithful pictures of the biochemical processes inside cells, mathematical models can help to test and complete these pictures, for instance, by extracting biochemical constants from experimental data. With their focus on molecular and cellular processes, biochemical models cover the mesoscopic scale between nanometres and micrometres, between the macroscopic world of cell populations and the microscopic world of molecule physics, to which they are connected through parameters like molecular binding energies. Thus, biochemical models have their place in the hierarchy of physical theories. But is there also something specifically "biological" to them, something that

distinguishes them from other physical models?

Biology is complex and historical

As a result of evolution, cellular pathways and much more diverse and "historical" than other physical systems like atoms, stars, or the weather. The essence of physical theories is often condensed in a simple set of equations, like the Maxwell equations in electrodynamics. Although biochemical models are also built on general formalisms – e.g., the kinetic modelling paradigm and the laws of thermodynamics – their essence lies much more in the specific network structures or numerical values, which differ from organisms to organism and have to be collected one by one through experiments. Biologists need to integrate various details and facts before a comprehensive picture can emerge and, accordingly, understanding in biology is often reached through examples and analogies rather than through general laws. With high-throughput data and the ability to build complex models of different organisms, this may possibly change. The paradigm shift towards complex models and the huge accumulation of data also creates unprecedented challenges for data analysis and modelling. And like in research on other complex systems, one of the key aspects of a model is actually what is *not* in the model, what could be left out without compromising the results.

Essentially biological questions?

On the contrary, there are views and principles that are really specific to biology. Prominent examples are genetic information, recognition between molecules, self-replication, the control and usage of noise, and, of course, evolution. Even if these principles do not explicitly appear in biochemical models, they have an impact on how models are made. The different aspects of a biochemical system – network structure, kinetics, dynamic, function, and evolution are closely entangled. To understand the structure and details of biochemical networks, we have to consider evolution; to understand how evolution proceeds, we need to see how a genotype – defining network structures, transcriptional regulation, and kinetic constants of enzymes – is translated into a biochemical dynamic and further into cellular traits and behaviour.

Eventually, all these perspectives on biochemical systems should converge into a consistent picture. As I argued here, this picture will not be a monolithic replica of the cell, but a collection of models of different resolution and scopes, which can be combined and adjusted according to the situation. Flexible modelling methods will enable us to translate our understanding of biochemistry and cell physiology into computer simulations, to face the consequences of our mental pictures, and to improve them based on experimental data.

7 References

[Beard and Qian, 2007] Beard, D. and Qian, H. (2007). Relationship between thermodynamic driving force and one-way fluxes in reversible processes. *PLoS ONE*, 2(1):e144.

[Beard et al., 2002] Beard, D. A., Liang, S., and Qian, H. (2002). Energy balance for analysis of complex metabolic networks. *Biophysical Journal*, 83(1):79–86.

[Borger et al., 2005] Borger, S., Liebermeister, W., and Klipp, E. (2005). Distribution of a bifurcation parameter in a genetic network with uncertain parameters. In *Proceedings of the 4th workshop on computation of biochemical pathways and genetic networks*, pages 95–101. Logos-Verlag, Berlin.

[Borger et al., 2006] Borger, S., Liebermeister, W., and Klipp, E. (2006). Prediction of enzyme kinetic parameters based on statistical learning. *Genome Informatics Series*, 17(1).

[Borger et al., 2007] Borger, S., Uhlendorf, J., Helbig, A., and Liebermeister, W. (2007). Integration of enzyme kinetic data from various sources. *In Silico Biology*, 7(S1):09.

[Box and Draper, 1987] Box, G. and Draper, N. (1987). *Empirical model-building and response surfaces*. Wiley.

[Buchner, 1897] Buchner, E. (1897). Alkoholische Gärung ohne Hefezellen. *Berichte der Deutschen Chemischen Gesellschaft*, 30:117–124.

[Buescher et al., 2012] Buescher, J., Liebermeister, W., Jules, M., Uhr, M., Muntel, J., Botella, E., Hessling, B., Kleijn, R., Chat, L. L., Lecointe, F., Mäder, U., Nicolas, P., Piersma, S., Rügheimer, F., Becher, D., Bessièeres, P., Bidnenko, E., Denham, E., Dervyn, E., Devine, K., Doherty, G., Drulhe, S., Felicori, L., Fogg, M., Goelzer, A., Hansen, A., Harwood, C., Hecker, M., Hubner, S., Hultschig, C., Jarmer, H., Klipp, E., Leduc, A., Lewis, P., Molina, F., Noirot, P., Peres, S., Pigeonneau, N., Pohl, S., Rasmussen, S., Rinn, B., Schaffer, M., Schnidder, J., Schwikowski, B., van Dijl, J., Veiga, P., Walsh, S., Wilkinson, A., Stelling, J., Aymerich, S., and Sauer, U. (2012). Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. *Science*, 335(6072):1099–1103.

[Chance, 1943] Chance, B. (1943). The kinetics of the enzyme-substrate compound of peroxidase. *J. Biol. Chem.*, 151:553577.

[Chance et al., 1952] Chance, B. et al. (1952). The mechanism of catalase action. II Electric analog computer studies. *Arch. Biochem. Biophys.*, 37:322339.

[Degtyarenko et al., 2008] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database issue):D344.

[Dräger et al., 2008] Dräger, A., Hassis, N., Supper, J., Schröder, A., and Zell, A. (2008). SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Systems Biology*, 2:39.

[Ederer and Gilles, 2007] Ederer, M. and Gilles, E. (2007). Thermodynamically feasible kinetic models of reaction networks. *Biophys. J.*, 92:1846–1857.

[Florez et al., 2009] Florez, L., Roppel, S., Schmeisky, A., Lammers, C., and Stülke, J. (2009). A community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki. *Database*, 1:bap012.

[Funahashi et al., 2008] Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8):1254–1265.

[Gillespie, 2000] Gillespie, D. (2000). The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306.

[Goelzer et al., 2008] Goelzer, A., Brikci, F., Martin-Verstraete, I., Noirot, P., Bessières, P., Aymerich, S., and Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Systems Biology*, 2(20).

[Haldane, 1930] Haldane, J. (1930). *Enzymes*. Longmans, Green and Co., London. (republished in 1965 by MIT Press, Cambridge, MA).

[Heinrich and Schuster, 1996] Heinrich, R. and Schuster, S. (1996). *The Regulation of Cellular Systems*. Chapman & Hall.

[Henry et al., 2010] Henry, C., DeJongh, M., Best, A., Frybarger, P., Linsay, B., and Stevens, R. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28:977–82.

[Herrgård et al., 2008] Herrgård, M., Swainston, N., Dobson, P., Dunn, W., Arga, K., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Li, P., N, L. N., , Liebermeister, W., Mo, M., Oliveira, A., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasil, I., Weichart, D., Brent, R., Broomhead, D., Westerhoff, H., Kirdar, B., Penttilä, M., Klipp, E., Palsson, B. O., Sauer, U., Oliver, S., Mendes, P., Nielsen, J., and Kell, D. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26:1155–1160.

[Hill, 1910] Hill, A. (1910). The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *The Journal of Physiology*, 40:Supplement.

[Höfer and Heinrich, 1993] Höfer, T. and Heinrich, R. (1993). A second-order approach to metabolic control analysis. *J. Theor. Biol.*, 164:85–102.

[Hofmeyr, 2001] Hofmeyr, J.-H. (2001). Metabolic control analysis in a nutshell. In *ICSB 2001 Online Proceedings, http://www.icsb2001.org/toc.html*, page 291.

[Holzhütter, 2004] Holzhütter, H. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.*, 271(14):2905–2922.

[Holzhütter, 2006] Holzhütter, H. (2006). The generalized flux-minimization method and its application to metabolic networks affected by enzyme deficiencies. *BioSystems*, 83:98–107.

[Hoops et al., 2006] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI–a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074.

[Hoppe et al., 2007] Hoppe, A., Hoffmann, S., and Holzhütter, H. (2007). Including metabolite concentrations into flux-balance analysis: Thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst. Biol*, 1(1):23.

[Hornberg et al., 2005] Hornberg, J., Bruggeman, F., Binder, B., Geest, C., de Vaate, A., Lankelma, J., Heinrich, R., and Westerhoff, H. (2005). ERK phosphorylation and kinase/phosphatase control. *FEBS Journal*, 272(1):244–258.

[Huang and Ferrell, 1996] Huang, C. and Ferrell, J. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *PNAS*, 93(19):10078.

[Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., Arkin, A., Bornstein, B., Bray, D., Cornish-Bowden, A., Cuellar, A., Dronov, S., Gilles, E., Ginkel, M., Gor, V., Goryanin, I., Hedley, W., Hodgman, T., Hofmeyr, J., Hunter, P., Juty, N., Kasberger, J., Kremling, A., Kummer, U., Novère, N. L., Loew, L., Lucio, D., Mendes, P., Minch, E., Mjolsness, E., Nakayama, Y., Nelson, M., Nielsen, P., Schaff, T. S. T. J., Shapiro, B., Shimizu, T., Spence, H., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and the SBML Forum (2003). The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[Hynne et al., 2001] Hynne, F., Danø, S., and Sørensen, P. (2001). Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophys. Chem.*, 94:121–163.

[Ingalls, 2004] Ingalls, B. (2004). A frequency domain approach to sensitivity analysis of biochemical systems. *J Phys Chem B*, 108:1143–1152.

[Ingalls and Sauro, 2003] Ingalls, B. and Sauro, H. (2003). Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. *J. Theor. Biol.*, 222(1):23–36.

[Jacob and Monod, 1961] Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356.

[Jaynes, 1957] Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 106:620–630.

[Klipp et al., 2004] Klipp, E., Liebermeister, W., and Wierling, C. (2004). Inferring dynamic properties of biochemical reaction networks from structural knowledge. *Genome Informatics*, 15(1):125–137.

[Klipp et al., 2009] Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009). *Systems Biology - A Textbook*. Wiley-VCH.

[Krause et al., 2011] Krause, F., Schulz, M., Swainston, N., and Liebermeister, W. (2011). *Methods in Systems Biology*, volume 500 of *Methods in Enzymology*, chapter Sustainable model building: the role of standards and biological semantics, pages 371–395.

[Krause et al., 2010] Krause, F., Uhlendorf, J., Lubitz, T., Klipp, E., and Liebermeister, W. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics*, 26(3):421–422.

[Kümmel et al., 2006] Kümmel, A., Panke, S., and Heinemann, M. (2006). Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Molecular Systems Biology*, 2:2006.0034.

[Le Novère, 2006] Le Novère, N. (2006). Model storage, exchange and integration. *BMC Neuroscience*, 7(Suppl 1):S11.

[Le Novère et al., 2005] Le Novère, N., Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E., Mendes, P., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23(12):1509–1515.

[Liao et al., 2003] Liao, J., Boscolo, R., Yang, Y., Tran, L., Sabatti, C., and Roychowdhury, V. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA*, 100(26):15522–15527.

[Liebermeister, 2005a] Liebermeister, W. (2005a). Predicting physiological concentrations of metabolites from their molecular structure. *J. Comp. Biol.*, 12(10):1307–1315.

[Liebermeister, 2005b] Liebermeister, W. (2005b). Response to temporal parameter fluctuations in biochemical networks. *J. Theor. Biol.*, 234(3):423–438.

[Liebermeister, 2008] Liebermeister, W. (2008). Validity and combination of biochemical models. In *Proceedings of 3rd International ESCEC Workshop on Experimental Standard Conditions on Enzyme Characterizations*.

[Liebermeister et al., 2005] Liebermeister, W., Baur, U., and Klipp, E. (2005). Biochemical network models simplified by balanced truncation. *FEBS Journal*, 272(16):4034 – 4043.

[Liebermeister and Klipp, 2005] Liebermeister, W. and Klipp, E. (2005). Biochemical networks with uncertain parameters. *IEE Proc. Sys. Biol.*, 152(3):97–107.

[Liebermeister and Klipp, 2006a] Liebermeister, W. and Klipp, E. (2006a). Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Mod.*, 3:41.

[Liebermeister and Klipp, 2006b] Liebermeister, W. and Klipp, E. (2006b). Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor. Biol. Med. Mod.*, 3:42.

[Liebermeister et al., 2004] Liebermeister, W., Klipp, E., Schuster, S., and Heinrich, R. (2004). A theory of optimal differential gene expression. *BioSystems*, 76:261–278.

[Liebermeister et al., 2010] Liebermeister, W., Uhlendorf, J., and Klipp, E. (2010). Modular rate laws for enzymatic reactions: thermodynamics, elasticities, and implementation. *Bioinformatics*, 26(12):1528–1534.

[Lotka, 1925] Lotka, A. (1925). *Elements of Physical Biology*. reprinted by Dover in 1956 as Elements of Mathematical Biology.

[Lubitz et al., 2010] Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010). Parameter balancing for kinetic models of cell metabolism. *J. Phys. Chem. B*, 114(49):16298–16303.

[Moore, 1981] Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEETransAC*, AC-26:17–32.

[Novère et al., 2006] Novère, N. L., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., and Hucka, M. (2006). Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34:Database Issue:D689–91.

[Novère et al., 2009] Novère, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M., Wimalaratne, S., Bergman, F., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D., Sander, C., Sauro, H., J.L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741.

[Orth et al., 2010] Orth, J., Thiele, I., and Palsson., B. (2010). What is flux balance analysis? *Nature Biotechnology*, 28:245–248.

[Pfeiffer and Schuster, 2005] Pfeiffer, T. and Schuster, S. (2005). Game-theoretical approaches to studying the evolution of biochemical systems. *Trends in biochemcial sciences*, 30(1):20–25.

[Qian and Beard, 2005] Qian, H. and Beard, D. (2005). Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. *Biophysical Chemistry*, 114(2-3):213–220.

[Ronen et al., 2002] Ronen, M., Rosenberg, R., Shraiman, B., and Alon, U. (2002). Assigning numbers to the arrows: parametrizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci USA*, 99(16):10555–10560.

[Savageau, 1969] Savageau, M. (1969). Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J. Theor. Biol.*, 25(3):370–379.

[Schomburg et al., 2004] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32:Database issue:D431–433.

[Schrödinger, 1944] Schrödinger, E. (1944). *What is life?* Cambridge University Press.

[Schulz et al., 2011] Schulz, M., Krause, F., Novère, N. L., Klipp, E., and Liebermeister, W. (2011). Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular Systems Biology*, 7:512.

[Schulz et al., 2006] Schulz, M., Uhlendorf, J., Klipp, E., and Liebermeister, W. (2006). SBMLmerge, a system for combining biochemical network models. *Genome Informatics Series*, 17(1).

[Schuster et al., 1999] Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60.

[Schuster and Schuster, 1989] Schuster, S. and Schuster, R. (1989). A generalization of Wegscheider's condition. Implications for properties of steady states and for quasi-steady-state approximation. *J. Math. Chem.*, 3:25–42.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal,*, 27:379–423 and 623–656.

[Sierro et al., 2008] Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Research*, 36 (Database issue):D93–D96.

[Stelling et al., 2004] Stelling, J., Sauer, U., Szallasi, Z., Doyle, F., and Doyle, J. (2004). Robustness of Cellular Functions. *Cell*, 118(6):675–685.

[Steuer et al., 2006] Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006). Structural kinetic modeling of metabolic networks. *Proc Natl Acad Sci USA*, 103(32):11868–11873.

[Steuer et al., 2003] Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomics networks. *Bioinformatics*, 19(8):1019–1026.

[Visser and Heijnen, 2003] Visser, D. and Heijnen, J. (2003). Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab Eng*, 5(3):164–176.

[von Berthalanffy, 1932] von Berthalanffy, L. (1932). *Theoretische Biologie I. Band : Allgemeine Theorie, Physikochemie, Aufbau und Entwicklung des Organismus*. Gebrüder Borntraeger, Berlin.

[von Berthalanffy, 1953] von Berthalanffy, L. (1953). *Biophysik des Fließgleichgewichts. Einführung in die Physik offener Systeme und ihre Anwendung in der Biologie*. Vieweg & Sohn, Braunschweig.

[Wang et al., 2004] Wang, L., Birol, I., and Hatzimanikatis, V. (2004). Metabolic control analysis under uncertainty: Framework development and case studies. *Biophysical Journal.*, 87(6):3750–3763.

[Wegscheider, 1902] Wegscheider, R. (1902). Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reactionskinetik homogener Systeme. *Z Phys Chem*, 39:257–303.

[Wilhelm and Heinrich, 1995] Wilhelm, T. and Heinrich, R. (1995). The smallest chemical reaction systems with Hopf-bifurcation. *J. Math. Chem.*, 17:1–14.

[Wittig et al., 2006] Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., and Rojas, I. (2006). SABIO-RK: Integration and curation of reaction kinetics data. *Proceedings of the 3rd International workshop on Data Integration in the Life Sciences 2006 (DILS'06) Hinxton, UK. Lecture Notes in Computer Science*.

8 Publications submitted for the habilitation

[1] J.M. Buescher, **W. Liebermeister**, M. Jules, M. Uhr, J. Muntel, E. Botella, B. Hessling, R.J. Kleijn, L. Le Chat, F. Lecointe, U. Mäder, P. Nicolas, S. Piersma, F. Rügheimer, D. Becher, P. Bessièeres, E. Bidnenko, E.L. Denham, E. Dervyn, K.M. Devine, G. Doherty, S. Drulhe, L. Felicori, M.J. Fogg, A. Goelzer, A. Hansen, C.R. Harwood, M. Hecker, S. Hubner, C. Hultschig, H. Jarmer, E. Klipp, A. Leduc, P. Lewis, F. Molina, P. Noirot, S. Peres, N. Pigeonneau, S. Pohl, S. Rasmussen B. Rinn, M. Schaffer, J. Schnidder, B. Schwikowski, J.M. van Dijl, P. Veiga, S. Walsh, A.J. Wilkinson, J. Stelling, S. Aymerich, and U. Sauer. Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism *Science*, 335(6072):1099–1103.

[2] M. Schulz, F. Krause, N. Le Novère, E. Klipp, and **W. Liebermeister**. Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular Systems Biology*, 7:512, 2011.

[3] T. Lubitz, M. Schulz, E. Klipp, **W. Liebermeister**. Parameter balancing for kinetic models of cell metabolism. *J. Phys. Chem. B*, 114(49):16298–16303, 2010.

[4] **W. Liebermeister**, J. Uhlendorf, and E. Klipp. Modular rate laws for enzymatic reactions: thermodynamics, elasticities, and implementation. *Bioinformatics*, 26(12):1528–1534, 2010.

[5] F. Krause, J. Uhlendorf, T. Lubitz, E. Klipp, and **W. Liebermeister**. Annotation and merging of SBML models with semanticSBML. *Bioinformatics*, 26(3):421–422, 2010.

[6] **W. Liebermeister**. Validity and combination of biochemical models. *Proceedings of 3rd International ESCEC Workshop on Experimental Standard Conditions on Enzyme Characterizations*, 2008.

[7] S. Borger, J. Uhlendorf, A. Helbig, and **W. Liebermeister**. Integration of enzyme kinetic data from various sources. *In Silico Biology*, 7(S1):09, 2007.

[8] M. Schulz, J. Uhlendorf, E. Klipp, and **W. Liebermeister**. SBMLmerge, a system for combining biochemical network models. *Genome Informatics Series*, 17(1), 2006.

[9] **W. Liebermeister** and E. Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Mod.*, 3:41, 2006.

[10] **W. Liebermeister** and E. Klipp. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor. Biol. Med. Mod.*, 3:42, 2006.

[11] **W. Liebermeister**, U. Baur, and E. Klipp. Biochemical network models simplified by balanced truncation. *FEBS Journal*, 272(16):4034 – 4043, 2005.

[12] **W. Liebermeister**. Predicting physiological concentrations of metabolites from their molecular structure. *J Comp Biol*, 12(10):1307–1315, 2005.

[13] **W. Liebermeister**. Response to temporal parameter fluctuations in biochemical networks. *J Theor Biol*, 234(3):423–438, 2005.

[14] **W. Liebermeister** and E. Klipp. Biochemical networks with uncertain parameters. *IEE Proc. Sys. Biol.*, 152(3):97–107, 2005.

9 Acknowledgments

I like to collectively thank all the people who helped and supported me during those years with discussions and their presence, privately or at work. My special thanks goes to Prof. Dr. Dr. Edda Klipp for our long-lasting cooperation, to Prof. Dr. Reinhart Heinrich (†) and Prof. Dr. Uri Alon, who greatly inspired my thoughts, and to Dr. Ron Milo, who kindly hosted me in 2011 at the Weizmann Institute of Science.