Cutting the Wires: Modularization of Cellular Networks for Experimental Design

Moritz Lang,[†*] Sean Summers,[‡] and Jörg Stelling[†]
[†]Department of Biosystems Science and Engineering, ETH Zürich, and Swiss Institute of Bioinformatics, Basel, Switzerland; and [‡]Automatic Control Laboratory, ETH Zürich, Zurich, Switzerland

ABSTRACT Understanding naturally evolved cellular networks requires the consecutive identification and revision of the interactions between relevant molecular species. In this process, initially often simplified and incomplete networks are extended by integrating new reactions or whole subnetworks to increase consistency between model predictions and new measurement data. However, increased consistency with experimental data alone is not sufficient to show the existence of biomolecular interactions, because the interplay of different potential extensions might lead to overall similar dynamics. Here, we present a graph-based modularization approach to facilitate the design of experiments targeted at independently validating the existence of several potential network extensions. Our method is based on selecting the outputs to measure during an experiment, such that each potential network extension becomes virtually insulated from all others during data analysis. Each output defines a module that only depends on one hypothetical network extension, and all other outputs act as virtual inputs to achieve insulation. Given appropriate experimental time-series measurements of the outputs, our modules can be analyzed, simulated, and compared to the experimental data separately. Our approach exemplifies the close relationship between structural systems identification and modularization, an interplay that promises development of related approaches in the future.

INTRODUCTION

Our knowledge of the species and their interactions in most cellular networks is incomplete. For many networks, a consensus already exists about a set of core species and high-confidence reactions, but this cannot yet explain all experimental data. To close the gap between model predictions and experimental results, one might extend the network with hypothetical, low-confidence reactions. However, if several (sets of) competing model extensions exist that are consistent with already available data, new experiments have to be designed to either confirm or discard the individual extensions. The corresponding tasks of model discrimination and experimental design for model structure identification pose substantial theoretical and computational challenges.

In model discrimination (see Kirk et al. (1) for a recent review), each possible combination of low-confidence reactions defines a so-called candidate model M_i, $i \in 1,\ldots, m$, typically with unknown parameterization (2–6). The goal is to identify the most probable model $\widehat{M}$, respective to estimating the posterior probabilities $p(M_i|D)$ in Bayesian approaches, given the experimental data D. However, the number of candidate models typically increases exponentially with the number of low-confidence reactions, and each model has to be evaluated in potentially high-dimensional parameter spaces. In our opinion, the computational complexity resulting from these two effects constitutes one of the most important challenges in constructing larger biomolecular models. Consequently, model discrimination has focused on networks with typically only up to 20 or 30 parameters, and with only a small number of mutually compatible hypothetical reactions (5) or candidate models (2,6).

These approaches assume that the existence of a given low-confidence reaction cannot be detected directly, but only through its influence on the dynamics of measurable species in the network. The decision of which species to measure is thereby often not given a priori, but rather as the subject of experimental design (5,7). For large, highly interconnected networks including several—often only poorly understood—low-confidence reactions, it becomes challenging to identify which species concentrations should be measured such that the experimentally observable dynamics can be explained only by the existence or non-existence of specific low-confidence reactions, and not by a combination of other hypothetical reactions.

Here, we propose to solve this problem of experimental design for model discrimination by modularization: with an adequate definition of modularization, a module—which is interpretable as a subnetwork—and the low-confidence reactions therein can be (structurally) identified separately from the rest of the network in a divide-and-conquer strategy (see below for a discussion of identifiability issues).

Modularity has been hypothesized early to be a key feature of biological network organization (8). Albeit modularization is an important tool for analyzing large-scale networks, surprisingly few conceptually different modularization methods exist (9,10). Most approaches determine highly connected cliques or communities of species, subnetworks with relatively few connections to other modules, or a combination thereof (11–15). Depending on the specific

Submitted July 17, 2013, and accepted for publication November 12, 2013.

*Correspondence: moritz.lang@bsse.ethz.ch

Editor: Reka Albert.

0006-3495/14/01/0321/11 $2.00

http://dx.doi.org/10.1016/j.bpj.2013.11.2960

definitions of the terms "highly connected" and "modularity", they utilize various automatic or semiautomatic approaches including hierarchical clustering (11,16), graph cuts based on the number of shortest paths through network edges (i.e., "betweenness"; see Girvan and Newman (14)), aggregation of smaller elementary modules based on size (12), greedy algorithms (17), or hybrid approaches based on eigenvectors of a modularity matrix combined with fine-tuning by local optimization (13). A recent approach punishes only bidirectional connections between modules to eliminate minimal, nontrivial feedback loops that establish retroactivity (18,19). These and similar approaches are especially valuable when trying to initially understand larger networks because the elements in a given module typically have common—sometimes surprising—properties that might help to explain the structure of the network (14).

Other module definitions aim at facilitating the analysis of specific network properties (20). Pioneering work on monotone network decompositions (21,22), however, appears to have limited practical applicability because analytical simplicity gained by monotone modules is often compensated for by complex interfaces between modules. For metabolic (mass-balanced) networks, elementary flux modes are (minimal) sets of reactions that can operate at steady state (23). This and similar concepts (24,25) can be seen as modules describing functionally related sets of reactions in steady state. Network motifs (26)—that is, small subnetworks with a wiring that appears statistically more often in networks than expected—constitute another approach; their practical value lies in coarse-graining and identification of small functional units, rather than in the modularization of large networks.

Nearly all available modularization approaches have in common what we consider an inherent design problem (pathway-based methods such as from Stelling et al. (23) are an exception): they will (nearly) always return at least one network modularization. Therefore, the absolute significance of a given modularization typically cannot be quantified because scores and similar metrics provide relative assessments only. Note that this holds also for approaches like those of Newman (13) and Clauset et al. (17), which might return a single module containing the whole network, but for which only slightly worse performing modularizations with more than one module might exist. Often, this inherent design problem results from weakly defined goals of modularization approaches beyond visualization, data mining, and similar.

Here, we propose that modularization methods should provide experimental designs for structural systems identification, and succeed if and only if a given type of experimental design is possible. Our graph-based modularization approach, called insulating modularization, aims at identifying groups of outputs that should be measured simultaneously such that each output defines a module with exactly one low-confidence reaction, and such that the models of each module can be simulated and analyzed separately, specifically without knowledge on the existence of all other low-confidence reactions.

It is an intuitive idea to define modules such that (at least parts of) the interactions in a given module can be analyzed in isolation from the other modules. However, this idea is rarely considered as part of the definition of a module. As mentioned explicitly in Bowsher (12), the isolated identification of most of the parameters in a module should be feasible by measuring all (and probably also a subset of) the species in that module. Different to Bowsher (12), we do not consider parameter identification but model selection, based on a minimal numbers of species to be measured for a specific model selection task. Thus, our proposed method is conceptually different to other modularization approaches in that we consider modularity of a specific question on a (biological) network, rather than modularity of the network as such (27,28). Consequently, the modularizations proposed by our method can change completely with the (biological) question; similarities between modularizations proposed by our method and by other methods would be rather coincidental.

Our method has the main advantage of attenuating the combinatorial explosion both in the number of competing model structures and in the required number of parameter samples (because of lower-dimensional parameter spaces for each module). Other experimental design methods are typically concerned with determining dynamic trajectories of input signals, sampling times, suitable gene knockouts, and similar (29–33) to improve model identification or selection. Albeit some methods (33) can also identify (optimal) sets of outputs, the majority of these approaches are complementary to our method: for a given modularization, they can further specify the experimental design, for example, by determining (optimal) dynamic trajectories for external inputs.

RESULTS

Overview

We analyze relatively well-known networks in which the number of high-confidence reactions is (significantly) higher than the number of hypothetical, low-confidence reactions. Furthermore, albeit not strictly necessary, we assume that at least a few reactions are irreversible. The goal of an insulating modularization is to separate the network into the same number of subnetworks (modules) as the number of low-confidence reactions. Each module contains exactly one low-confidence reaction, and it is insulated from the rest of the network by choosing an adequate set of outputs such that the dynamics of all species in the interface of a module to the rest of the network are experimentally measured. Each output is furthermore defined such that it can be used for the evaluation of models of exactly

one module, that is, for testing whether the associated low-confidence reaction exists (Fig. 1).

We achieve the insulation between the modules by utilizing that the representation of applied input signals and measured output signals is essentially identical for the mathematical analysis; both are typically given as a set of time-value pairs. Thus, when measuring the sets of outputs proposed by our method, one can use the measurement data of the output of one module (respectively a spline fitted to the data) as a virtual input for the simulation of all others (compare Fig. 1 *D*).

For strict modularizations, two alternative models of each module (with and without the respective low-confidence reaction) can be created. The strictness property guarantees that both models only depend on measured outputs and high-confidence reactions (see the Supporting Material), such that their agreement with the experimental results can be directly compared using, for instance, Bayesian inference (5). Specifically, if the strict modules do not overlap, the probability for the existence of one low-confidence reaction becomes conditionally independent of the existence of all other low-confidence reactions, given the experimental measurements of the proposed outputs (see the Supporting Material). Overlapping modules, as in Fig. 1 *C*, may contain reactions assigned to more than one module (here: R_2). If the corresponding reaction rates depend on parameters that are adjusted by a model discrimination algorithm, independent adjustments in different modules may lead to incompatibilities. Thus, for overlapping modularizations, one should check at least a posteriori if such a case appeared. Such cases should be taken seriously, the more so as they may indicate certain flaws in the structure of the original model. We will not discuss details because corrections depend on the specific modeling technique and the specific model discrimination method (e.g., including fixing of parameter values or coupled discrimination approaches), and because in our experience these overlaps occur rarely in real-world modularizations problems, might not occur in all possible modularizations, and are commonly small in size (concerning only a few reactions; see examples in sections Example Network and JAK2/STAT5 signaling).

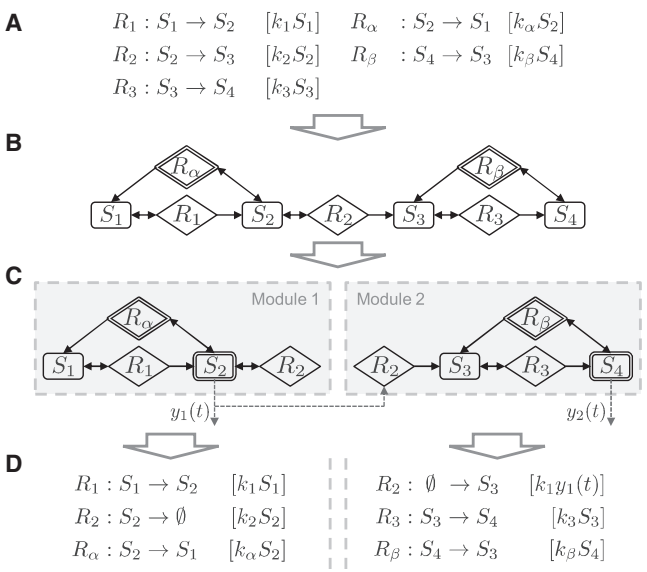

FIGURE 1 Workflow of our modularization approach. (*A*) Mass-action kinetics model with four species (S_1–S_4), three high-confidence reactions (R_1–R_3), and two low-confidence reactions R_α and R_β that should be experimentally verified or discarded. (*B*) Species-reaction graph of the network, where box-shaped vertices represent species, diamond-shaped ones reactions, and low-confidence reaction vertices are marked by two borders. A directed edge is drawn from a species to a reaction if the respective reaction rate depends on the species' concentration, and a directed edge from a reaction to a species if turnover of the reaction changes the concentration of the species. (*C*) By measuring S_2 and S_4, the network can be split into two modules, with each module including exactly one low-confidence reaction. The time series data y_1 of species S_2, the output of the first module, serves as a virtual input for the second module. (*D*) Models of the two identified modules can be extracted, simulated separately given the experimental time-series data $y_1(t)$ and $y_2(t)$, and used to assess the existence of each low-confidence reaction independently.

For simple modularizations, only models of the modules representing the network without the respective low-confidence reactions are guaranteed to be independent of all other low-confidence reactions. Thus, models pertaining to one module cannot necessarily be directly compared without knowledge on the existence of other low-confidence reactions. However, following the principle of parsimony, which also underlies Bayesian inference (35), one should favor the network without the hypothetical reaction if it is in sufficient agreement with all data. If details on the interactions of the species in the low-confidence reactions are unknown, one might, for example, merge whole subnetworks into one black-box, low-confidence reaction to represent the hypothesis that some species might influence—by a yet unknown mechanism—the concentrations of some other species.

It is important to note that our algorithm only guarantees—given adequate measurement data with sufficient temporal and quantitative resolution—that the modules can be virtually insulated from each other. More specifically, our method guarantees that there exists a directed path between the low-confidence reaction and the corresponding output in a given module. This is necessary but not sufficient for identification of the low-confidence reaction using its corresponding output. Symmetries in the network might, for example, lead to structural nonidentifiability of a low-confidence reaction given the outputs of a modularization. In addition, practical identifiability can be prevented, for example, by a low sensitivity of the output of a module to the respective low-confidence reaction, or by high model uncertainties. We therefore recommend the application of an adequate method (e.g., Sedoglavic (36) for deterministic models consisting of ordinary differential equations) to check identifiability of the low-confidence reactions for a given modularization, and potentially to choose a different

modularization. Measurements of additional species in the modules corresponding to the nonidentifiable low-confidence reactions may alleviate these problems.

Insulating modularizations

In this section, we provide the mathematic definitions and theorems used by our modularization approach. We illustrate each definition by discussing the small example given in Fig. 1. We recommend that readers not interested in all details of our method focus on Definitions 1–4, given below, and the respective examples. In the following, we assume that standard graph-theoretical notations are known to the reader (see, e.g., Cormen et al. (37)).

Definition 1 (binary-labeled species-reaction graph)

A species-reaction (SR) graph is a directed bipartite graph $G_{SR} = (V_S, V_R, E)$. The vertices in $V_S = \{S_1,\ldots,S_n\}$ and in $V_R = \{R_1,\ldots,R_r\}$ represent the species and the reactions of a biomolecular network, respectively. The directed edges of the network are defined as $(S_i,R_j) \in E$ if the reaction rate of R_j depends on the concentration of S_i, and $(R_j,S_i) \in E$ if turnover of reaction R_j changes the concentration of species S_i.

An SR graph is binary-labeled (BLSR graph) if each of its vertices has a binary label (weight) of 0 or 1. These labels are assigned in the following manner: (low-confidence reaction label function) $\eta_R(R_j)$ is 1 (0), if R_j is in the low-confidence (high-confidence) set of reactions, and (output label function) $\eta_S(S_i)$ is 1 (0), if the concentration of S_i is measured (not measured).

For notational convenience, let $V_{R,\eta} = \{R_j \in V_R : \eta_R(R_j) = 1\}$ and $V_{S,\eta} = \{S_i \in V_S : \eta_S(S_i) = 1\}$ be the set of low-confidence reactions and the set of measured outputs, respectively.

Fig. 1 *A* shows a small example of a mass-action reaction network with three high-confidence (R_1–R_3) and two low-confidence reactions (R_α and R_β), and Fig. 1 *B* its corresponding BLSR graph. In this example, none of the species S_1–S_4 is measured yet, but measuring any combination of species would also correspond to a BLSR graph. Note that the definition is not limited to mass-action models, and that it is compatible with reactions that involve more than two species and with reaction rates of (nearly) arbitrary form.

Definition 2 (reachability cost of a BLSR graph)

The reachability cost (length) $\omega(p)$ of a directed path p in a given BLSR graph is the sum of labels of the vertices in the path. The reachability cost $\omega_{\min}(u,v)$ is the minimal reachability cost of all directed paths from vertex u to another vertex $v \neq u$, or $+\infty$ if no path from u to v exists. For $u = v$, the reachability cost equals the label of the vertex.

In our example network (Fig. 1 *B*), the reachability costs from S_1 to R_1 and from S_1 to S_4 are zero because directed paths exist that do not contain R_α or R_β. On the other hand, the reachability cost from S_2 to S_1 is 1, because the only simple directed path contains R_α. Without a directed path between S_4 and S_1, the corresponding reachability cost is $+\infty$. Finally, by definition the reachability cost from S_1 to itself is 0 (S_1 is not measured), and from the low-confidence reaction R_α to itself it is 1.

Definition 3 (insulating modularization)

A BLSR graph is a simple insulating modularization, if and only if the number of low-confidence reactions is equal to the number of measured species, and there exists a unique, bijective function μ mapping low-confidence reactions to measured outputs, such that the reachability cost from a low-confidence reaction R_η to a measured output S_η is 2 if and only if $\mu(R_\eta) = S_\eta$.

A BLSR graph is a strict insulating modularization, if and only if it is a simple insulating modularization, and the reachability cost between two distinct low-confidence reactions is larger than 2.

Intuitively, in a simple insulating modularization, for every low-confidence reaction there exists exactly one measured species, such that the corresponding two vertices are connected by a directed path that does not contain any other measured species or low-confidence reactions. A strict insulating modularization additionally requires that every path between two different low-confidence reactions contains at least one measured species.

In our example network (Fig. 1 *B*), we obtain an insulating modularization if we decide to measure species S_2 and S_4 (that is, change the corresponding labels to 1). In this case, the reachability costs from R_α to S_2 and from R_β to S_4 are 2, whereas the reachability cost from R_α to S_4 is 3 (the shortest path contains S_2), and the reachability cost from R_β to S_2 is $+\infty$ (no directed path exists). This insulating modularization is strict: the shortest path from R_α to R_β has reachability cost 3, and there exists no directed path from R_β to R_α (reachability cost $+\infty$). Alternatively, we could also decide to measure S_2 and S_3, which leads to another strict insulating modularization. However, we would not obtain an insulating modularization by measuring S_1 and S_3 because the reachability costs from R_β to S_3 and from R_α to S_3 would be 2, indicating that S_1 fails to insulate S_3 from the influence of R_α.

Definition 4 (module)

A module $M(S_\eta)$ in a simple insulating modularization is the set of all vertices that have a reachability cost of 1 to a given measured output S_η:

$$M(S_\eta) = \{V_i \in V_S \cup V_R : \omega_{min}(V_i, S_\eta) = 1\}. \quad (1)$$

The interface $I(S_\eta)$ of a module is the set of all vertices for which a directed edge to a vertex in the corresponding module exists:

$$I(S_\eta) = \{V_i \in V_S \cup V_R : \omega_{\min}(V_i, S_\eta) = 2 \wedge \exists V_M \in M(S_\eta) : (V_i, V_M) \in E\}. \quad (2)$$

Fig. 1 *C* shows the two resulting modules when choosing to measure S_2 and S_4 in our example network (Fig. 1 *B*). The first module contains species S_1 and S_2 as well as reactions R_1 and R_2; its interface consists only of the low-confidence reaction R_α. The second module encompasses S_3 and S_4 as well as R_2 and R_3; the interface is constituted by the low-confidence reaction R_β and the measured output of S_2 ($y_1(t)$) of the first module. Note that in this example the two modules overlap because reaction R_2 pertains to both modules.

Lemma 5

A module in a simple insulating modularization contains exactly one measured species, and no low-confidence reaction:

$$M(S_\eta) \cap V_{S,\eta} = \{S_\eta\}, \quad (3a)$$

$$M(S_\eta) \cap V_{R,\eta} = \varnothing. \quad (3b)$$

The interface of a module only contains vertices that correspond to measured species, and exactly one vertex corresponding to a low-confidence reaction:

$$I(S_\eta) \subseteq V_{S,\eta} \cup V_{R,\eta}, \quad (4a)$$

$$I(S_\eta) \cap V_{R,\eta} = \{R_\eta\}, \text{ with } \mu(R_\eta) = S_\eta. \quad (4b)$$

Proof

A module $M(S_\eta)$ is defined by the set of all vertices for which a path to S_η with a reachability cost of one exists. S_η is part of $M(S_\eta)$ because $\omega_{\min}(S_\eta, S_\eta) = \eta_S(S_\eta) = 1$. No other vertex $V_i \in V_{S,\eta} \cup V_{R,\eta}$, $V_i \neq S_\eta$, being an output species or low-confidence-reaction, can be in the module $M(S_\eta)$, because every path from it to S_η at least contains itself and S_η ($\omega_{\min}(V_i, S_\eta) \geq \eta(V_i) + \eta_S(S_\eta) = 2$), or else no path exists between them at all ($\omega_{\min}(V_i, S_\eta) = \infty$).

That R_η, $\mu(R_\eta) = S_\eta$, is the only element of $V_{R,\eta}$, which is part of the interface $I(S_\eta)$, follows directly from the definition of a simple insulating modularization.

The shortest path between any element in the interface and S_η has, by definition, a reachability cost of 2 ($\forall V_i \in I(S_\eta)$ $\exists p = (V_i, V_j, \ldots, S_\eta) : V_j \in M(S_\eta) \wedge \omega(p) = 2$). Because, by definition, V_j is part of the module, it is true that $\omega_{\min}(V_j, S_\eta) = 1$, and $\omega(p) = \eta(V_i) + \omega_{\min}(V_j, S_\eta) \Rightarrow \eta(V_i) = 1$, that is, V_i is a measured species or a low-confidence reaction.

Lemma 6 (branching)

The BLSR graph that results from removing a low-confidence reaction vertex R_η (and all edges from and to this vertex) from a simple insulating modularization and from setting the label of the corresponding measured species $S_\eta = \mu(R_\eta)$ to 0 is a simple insulating modularization.

In our example network (Fig. 1 *B*), we could, for example, completely remove the low-confidence reaction vertex R_β and obtain an insulating modularization by measuring S_2 because measuring S_2 and S_4 would result in an insulating modularization of the original network. In this case, the first module remains unchanged, while the second module disappears (S_3, S_4, and R_3 would be part of no module). On the other hand, if we would remove the low-confidence reaction R_α and only measure S_4, we would obtain a single module with all vertices.

Proof

Assume $\mu(R_\eta) = S_\eta$, $R_\eta \in V_{R\eta}$, and $S_\eta \in V_{S\eta}$. Furthermore, assume $M(S'_\eta)$ being a different module ($S'_\eta \neq S_\eta$) with $S_\eta \in I(S'_\eta)$. Then, removing the low-confidence reaction R_η and the measured species label from S_η, there will be a new module $\tilde{M}(S'_\eta) = M(S_\eta) \cup M(S'_\eta)$ with interface $\tilde{I}(S'_\eta) = (I(S'_\eta) \setminus \{S_\eta\}) \cup (I(S_\eta) \setminus \{R_\eta\})$. If $S_\eta \notin I(S''_\eta)$, $S''_\eta \neq S_\eta$: $\tilde{M}(S''_\eta) = M(S''_\eta)$, and $\tilde{I}(S''_\eta) = I(S''_\eta)$.

Lemma 7 (strictness)

The BLSR graph resulting from simultaneously setting the label of a low-confidence reaction vertex R_η and its corresponding measured species vertex $S_\eta = \mu(R_\eta)$ of a strict insulating modularization to 0 is a strict insulating modularization.

Note that the difference between Lemmas 6 and 7 is that for a strict insulating modularization the respective vertex of a low-confidence reaction R_η is not removed from the BLSR-graph together with the label of the corresponding measured species, but only its low-confidence label. The proof is according to Lemma 6.

Because measuring species S_2 and S_4 in our example (Fig. 1 *B*) results in a strict insulating modularization, we could—instead of removing a low-confidence reaction vertex as before—simply change its label such that it becomes a high-confidence reaction. Removing the label of R_α and measuring only S_4, or removing the label of R_β and measuring only S_2, would then also result in a strict insulating modularization.

Definition 8

Let the bijective function $l_S : \{1,\ldots,n\} \rightarrow V_S$, and respectively, $l_R : \{1,\ldots,r\} \rightarrow V_R$, induce an order on the elements of V_S, respectively V_R, of a BLSR graph. Without loss of generality, let $\forall j \in \{1,\ldots,|V_{R,\eta}|\} : l_R(j) \in V_{R,\eta}$, and $\forall i \in \{1,\ldots,|V_{S,\eta}|\} : l_S(i) \in V_{S,\eta}$.

Then there exists a unique $(n + r) \times (n + r)$ 0-1 biadjacency matrix

$$\mathbf{A_E} = \begin{pmatrix} \mathbf{0} & \mathbf{A_{RS}} \\ \mathbf{A_{SR}} & \mathbf{0} \end{pmatrix}, \quad (5)$$

with the element $a_{ij} \in \{0,1\}$ of the $n \times r$ 0-1 matrix $A_{RS} = (A_{RS,\eta}, A_{RS,\not\eta})$ equal to 1 if $(R_j, S_i) \in E$, and 0 otherwise. Likewise, the element $b_{ji} \in \{0,1\}$ of the $r \times n$ 0-1 matrix $A_{SR} = (A^T_{SR,\eta}, A^T_{SR,\not\eta})^{\mathrm{T}}$ is equal to 1 if $(S_j, R_i) \in E$, and 0 otherwise.

Theorem 9 (main result)

Let G_{BLSR} be a binary-labeled species reaction graph, and $\mathcal{C}$ be the set of all 0-1 matrices $C = [c_{ij}]_{|V_{R,\eta}|\times|V_S|}$, with

$$\forall i \in \{1, \ldots, |V_{R,\eta}|\} : \sum\nolimits_{j=1}^{|V_S|} c_{ij} = 1.$$

For a given $C \in \mathcal{C}$, let $D_{\Sigma,0}$ be defined as

$$D_{\Sigma,0}(E) = Ee^{AF}A_{RS,\eta} \tag{6}$$

with $F = (\mathrm{I} - C^T C)$, and $A = A_{RS,\not\eta}\, A_{SR,\not\eta}$.

Then, G_{BLSR} is a simple insulating modularization, if $|V_{S,\eta}| = |V_{R,\eta}| \wedge \exists\, !C \in \mathcal{C},\ c_{ij} = 1 \Rightarrow S_j \in V_{S,\eta}$, such that $D_{\Sigma,0}(C) = \mathrm{diag}(d_i), d_i > 0$, with $\mathrm{diag}(d_i) = [d_{ij}]$ as a diagonal matrix with $d_{ij} = d_i$ if $i = j$ and $d_{ij} = 0$ otherwise.

The matrix C defines the mapping between the low-confidence reactions and their associated measured output species:

$$c_{ij} = 1 \Leftrightarrow \mu(l_R(i)) = l_S(j). \tag{7}$$

Furthermore, G_{BLSR} is a strict insulating modularization, if G_{BLSR} is a simple insulating modularization and $D_{\Sigma,0}\,(A_{SR,\eta}F) = \mathrm{diag}(\delta_i)$, $\delta_i \geq 0$.

Note that for a strict insulating modularization, $D_{\Sigma,0}(A_{SR,\eta}F)$ has to be a diagonal matrix with diagonal elements ≥ 0, whereas the diagonal elements of $D_{\Sigma,0}(C)$ have to be strictly >0.

For the proof of this theorem and alternative formulations, see the Supporting Material.

Lemma 10

Let G_{IM} be a simple insulating modularization (IM) with the matrices C, A, F, and $A_{RS,\not\eta}$ as described above. Then, species j belongs to module i if element $m_{s,ij}$ of $M_S = Ce^{AF}$ is unequal to 0, and reaction k belongs to the module if element $m_{r,ik}$ of $M_R = Ce^{AF}\, A_{RS,\not\eta}$ is unequal to 0.

The proof for this lemma follows the proof for Theorem 9.

Finding insulating modularizations

An algorithm for computing all insulating modularizations has to solve the modularization problem defined by the following.

Problem 11 (modularization problem)

Input. An SR-graph $G_{SR} = (V_S, V_R, E)$ and a low-confidence reaction label function $\eta_R{:}V_R \rightarrow \{0,1\}$.

Output. The unique set $L_\Sigma = \{\eta_{S,1} \cdots \eta_{S,k}\}$ of all output label functions $\eta_{S,i}$, such that $\forall\ i \in 1,\ldots,\ k{:}G_{IM} = (V_S, V_R, E, \eta_{s,i}, \eta_R)$ is a simple insulating modularization.

In the example discussed in the previous section (Fig. 1 *B*), there are, in principle, $\binom{4}{2} = 6$ possibilities to choose two measured species out of a total of four species. It is easy to check that an algorithm to solve this modularization problem should have as an output that only two of these possibilities, namely either to measure S_2 and S_3, or S_2 and S_4, correspond to a simple insulating modularization. While in this small example it is feasible to check all possibilities by hand, it is desirable to have an efficient algorithm to solve larger modularization problems.

An exhaustive search to find all insulating modularizations for a given modularization problem with $|\eta_R| \ll |V_S|$ has computational cost that is exponential in $|\eta_R|$ (see the Supporting Material for details). Thus, the problem becomes computationally intractable for larger values of $|\eta_R|$. However, for a fixed $|\eta_R|$, we have polynomial complexity in $|V_S|$. Hence, networks with many vertices but few low-confidence reactions can be modularized by exhaustive search.

In fact, the following theorem shows that it is unlikely that a polynomial-time algorithm to solve Problem 11 exists:

Theorem 12 (NP-hardness)

The modularization problem is nondeterministic polynomial-time (NP)-hard. The proof of this theorem, which is based on showing that the problem of deciding if at least one simple insulating modularization exists is NP-complete, is given in the Supporting Material.

However, superpolynomial runtime in the number of low-confidence reactions as indicated by Theorem 12 is a worst-case scenario. The structure of many practically relevant modularization problems allows one to determine all possible simple insulating modularizations in reasonable time using a recursive branch-and-bound algorithm (see the Supporting Material for details) based on Lemma 6. It exploits that the first k, $k < |\eta_R|$, measured outputs of a final modularization also define an insulating modularization for a reduced modularization problem with only the first k low-confidence reactions. Thus, if the solution for the respective modularization problem defined by the first $|\eta_R|-1$ low-confidence reactions is already known, finding the solution to the original problem requires only testing all concatenations of the subsolutions to the remaining, not yet-labeled, species. The complexity of each of the $|\eta_R|$ recursion steps depends on the number of solutions found in the previous step and $|V_S|$, but not on $|\eta_R|$. In practice, we experienced significantly higher efficiency compared to an exhaustive search, making it possible to modularize networks with many species and several low-confidence reactions in reasonable time (see the Supporting Material for details).

Example network

To demonstrate our modularization approach, we considered the small species-reaction network depicted in Fig. 2 consisting of nine species, ten high-confidence, and two low-confidence reactions. Our branch-and-bound algorithm identified the six distinct insulating modularizations shown in the network representations in Fig. 3.

For more complex problems, one may obtain combinatorially many possible insulating modularizations, typically when the measured-output labels for several modules can be chosen (partly) independently. Albeit being able to choose between different sets of outputs is an advantage when it comes to (feasible) experimental designs, a more compact and intuitive representation of the different possibilities than an unsorted list seems favorable. Each element of L_Σ can be interpreted as $|\eta_R|$ subsequent (ordered) choices of outputs leading to an insulating modularization, and each sequence of choices not represented by an element in L_Σ as not leading to a modularization. Similar to binary functions represented by a truth table, our multivariate function defining whether a given output-labeling function leads to an insulating modularization can be compactly represented by a so-called reduced ordered multiple-valued decision diagram (ROMDD, see Srinivasan et al. (38) and Miller (39)), a rooted directed acyclic graph with a minimal number of vertices for a given order of variables (see Fig. 2). The diagram is traveled from the root to the leaves. In our case, the outgoing edges represent all possible choices that a given output label, represented by the current non-leaf vertex, can be assigned to. Finally, the leaves of an ROMDD depict whether a given labeling function, represented by the directed path from the root to the respective leaf, leads to a case of strict, simple, or no modularization. For our purposes, we adjusted the algorithm for the construction and modification of ROMDD presented in Miller and Drechsler (40) to handle decisions with more than four-valued variables.

The ROMD diagram of the example network (Fig. 2) and the six corresponding modularizations (Fig. 3) illustrate several important properties of our modularization approach:

1. The choice of an output for one module depends on the choice of all others, e.g., measuring species S_6 requires measuring species S_3, too.
2. Not all species and reactions belong to a module, such as species S_7 and reaction R_{11}.
3. Modules might overlap; species S_4 and reaction R_6, for example, can belong to more than one module.
4. The same species might be measured in different modularizations to identify the existence of distinct low-confidence reaction such as, for example, species S_9.
5. Our modularization approach can work for highly connected networks with a substantial number of bidirectional interactions.

JAK2/STAT5 signaling

In a recent study, Bachmann et al. (7) analyzed the possibility for the existence of a dual feedback mechanism in JAK2/STAT5 signaling, and its implications for the dynamic response to external *Epo* stimulation. Based on their results, they proposed an ordinary differential equation model composed of 26 states and 36 reactions. With several well-designed experiments, Bachmann et al. (7) could show the existence of the feedbacks, and identify most of their model's parameters. Here, we put ourselves in the position of a researcher having developed a hypothetical network structure represented by the model (7), but who is unaware of the experiments that were performed to validate this structure. In this thought-experiment, we use our modularization approach to propose experimental designs to prove or disprove the existence of the (assumed) low-confidence dual feedback mechanism.

We imported the Systems Biology Markup Language (SBML) description of the JAK2/STAT5 signaling model (7) from the BioModels Database (41) and automatically generated its respective species-reaction graph (Fig. 4). Subsequently, we labeled the reactions corresponding to

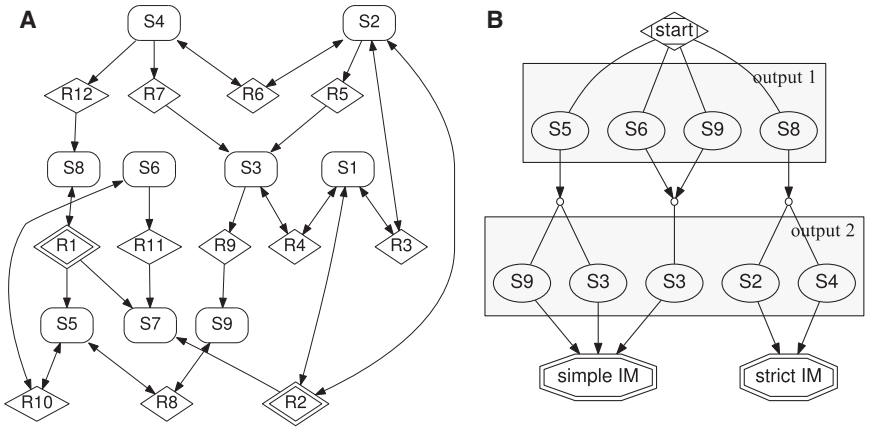

FIGURE 2 Example of modularization problem. (*A*) Network representation. (*Box-shaped vertices*) Species; (*diamond-shaped vertices*) reactions. Low-confidence reaction vertices are marked by two borders. (*B*) Reduced ordered multiple-valued decision diagram depicting the choices of species to measure in the example network to obtain an insulating modularization (IM). Conveniently, only decisions leading either to a strict or a simple modularization are shown. All graphs were drawn with GRAPHVIZ (43).

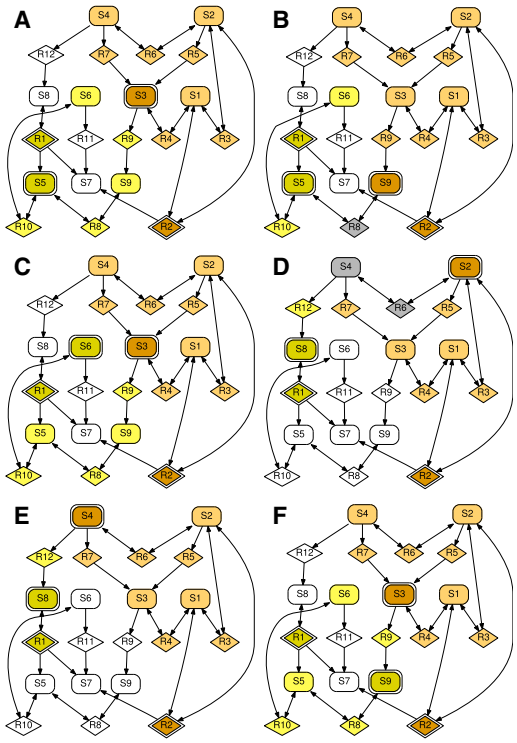

FIGURE 3 Distinct insulating modularizations of the example network. (*Box-shaped vertices*) Species; (*diamond-shaped vertices*) reactions. (*Yellow vertices*) Part of Module 1; (*brown vertices*) part of Module 2; (*gray vertices*) parts of both modules; and (*white vertices*) no module. Vertices representing the measured species of a module and the corresponding low-confidence reactions are filled differently (*slightly darker color* than their corresponding modules) and they have two borders (note that, by definition, low-confidence reactions do not belong to the respective module, but to its interface). All graphs were drawn with GRAPHVIZ (43). To see this figure in color, go online.

the dual feedback loop as being of low confidence. Specifically, this labeling includes the following interactions:

1. CIS and SOCS3 production rates are under transcriptional control of activated and nuclear localized STAT5 (reactions R_{17} and R_{27});
2. SOCS3 and CIS inhibit STAT5 activation by the *Epo* receptor complex (R_{14}); and
3. Phosphatase activity of SHP-1 requires prior binding to an *Epo* receptor with phosphorylated residue Tyr^{429} (R_{11}).

The last hypothesis (see also Klingmüller et al. (42)) was included although it is not essential for the dual feedback mechanism. Reactions R_{11}, R_{14}, R_{17}, and R_{27} also represent other molecular interactions and dependencies than the stated ones. Consequently, nonexistence of a described interaction would only imply modification, but not a complete removal of the respective low-confidence reaction. A strict insulating modularization guarantees that the measured outputs are insulated from all other low-confidence reactions except their corresponding ones for

FIGURE 4 Example insulating modularization for the JAK2/STAT5 signaling network. The graphical notation corresponds to Fig. 3 (*open* for no module). Graph drawn with GRAPHVIZ (43). To see this figure in color, go online.

all models with arbitrarily reduced complexity of the low-confidence reactions (see Lemma 7). However, to preserve integrity, we also include simple modularizations in the following analysis.

Using a nonoptimized MATLAB (The MathWorks, Natick, MA) implementation of our branch-and-bound algorithm (see the Supporting Material) to solve this modularization problem (26 states and 36 reactions, of which 4 were marked to be low-confidence) resulted in a total of 147 possible insulating modularizations, of which 98 are strict, in reasonable computational time (<4 s on an Intel Core 2 Duo, 3.16 GHz, 4 GB RAM). In contrast to the example network above, the ROMDD (Fig. 5) shows that all measured outputs can be chosen independently. Only measuring phosphorylated nuclear or cytosolic STAT5 concentrations leads to strict modularizations; measuring nonphosphorylated STAT5 concentrations results in simple modularizations. This is because conservation of the total STAT5 concentration is only implicitly modeled (7) and, thus, not considered by our algorithm. Furthermore, to simplify analysis, Bachmann et al. (7) used artificial intermediate mRNA species (SOCS3nRNA1–5 and CISnRNA1–5) to emulate transcriptional delays. Again, our algorithm is agnostic of the specific modeling approach and it consequently proposes to measure the artificial species. However, it is straightforward to exclude these artifacts before or after applying our modularization algorithm, and several other insulating modularizations remain.

Remarkably, the strict insulating modularization shown in Fig. 4 requires us to measure SOCS3 and CIS mRNA

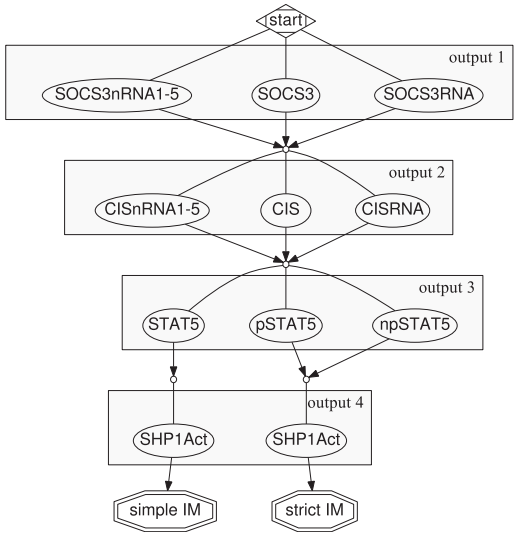

FIGURE 5 Reduced ordered multiple-valued decision diagram representing possible choices of tuples of suitable outputs to identify the low-confidence reactions in the JAK2/STAT5 pathway example. Conveniently, decisions representing the intermediate artificial mRNA species (SOCS3nRNA1–5 and CISnRNA1–5, respectively) were combined. Graph drawn with GRAPHVIZ (43).

concentrations, phosphorylated nuclear STAT5, and activated SHP-1. In Bachmann et al. (7), the authors measured CIS and SOCS3 expression profiles as well as the phosphorylation state of STAT5 over time, noting that several parameters in their model were not identifiable due to missing quantifications of (relative) SHP-1 activation levels. Thus, except for the SHP-1 module, their experiments were sufficient to modularize the JAK2/STAT5 signaling network using our approach, and to validate the different reactions comprising the proposed dual feedback mechanism separately. We believe that this shows that the freedom of choice our modularization approach offers by not only returning one, but several solutions in an easily comprehensible visual way, is essential to make an otherwise purely theoretical approach experimentally feasible.

DISCUSSION

Our modularization approach to split a biomolecular network into several, smaller modules differs in four main aspects from previous approaches:

1. Our modules have a practical meaning: a modularization directly instructs on which species should be experimentally measured to confirm or reject hypothetical reactions in a cellular network. Our approach thus shows a close relationship between modularization and experimental design, whereas most other modularization methods serve mainly to sort and visualize already available information.
2. Our modules can overlap each other and the union of all modules is not required to reconstitute the original network. The opposite requirements of other approaches seem intuitively reasonable, but they often lead to rather complex interfaces and so-called leftover or "scraggy" modules. Reactions and species unassigned to modules by our algorithm do not help in experimental design; requiring modules to contain unnecessary vertices would pose additional constraints on how the necessary vertices can be distributed.
3. Our approach can fail (return an empty set of possible modularizations) or give more than one solution. We believe that being able to fail is advantageous. Other methods returning modules for any given network structure leave the experimenter with the question of whether to trust the result, and relative quality indicators do not help in this regard. In contrast, representing all possible modularizations in a condensed and intuitive way as by the ROMDDs leaves the experimenter the freedom to select a set of outputs to implement.
4. Our approach is compatible with any modeling technique that describes which species influence each other (such that an adjacency matrix can be defined for a species-reaction graph). This applies to practically all contemporary modeling techniques that aim to represent real-world biomolecular networks, such as ordinary differential equations, stochastic representations, Boolean networks, and graphical models.

Note that we assume that outputs can be measured with sufficient precision and temporal resolution. In reality, all measurement methods are noisy and have limited temporal resolution. Thus, when using measured outputs as virtual inputs for our modules, the respective models will necessarily experience different input signals than the real-world networks. We expect that methods that quantify the agreement between a model and a real-world system, and which are able to deal with reasonable amounts of input noise, will also cope with measurement noise in our virtual inputs.

Future research might extend and adjust our approach, for instance, by increasing the amount of information used during the modularization—at the cost of restrictions to certain modeling techniques. One could use stoichiometric information, or tightly combine our approach with Bayesian analysis to further specify the experimental design in terms of type and the parameters of hypothetical reactions. We intentionally did not pursue these avenues here, to preserve the general applicability of our method. Furthermore, one could relax the requirement that each low-confidence reaction should be identifiable completely independently. In this case, Eq. 6 should result in lower triangular matrices (for a given order of low-confidence reactions) instead of diagonal matrices with nonzero diagonal elements. The existence of the first low-confidence reaction could then still be identified independently, and one could iterate over the remaining low-confidence reactions using this information. However, this modification would be prone to error propagation.

Finally, measuring a higher number of outputs than the number of low-confidence reactions in the network would help to ensure identifiability, verify the conclusions obtained by the outputs defining the modules, and—even more importantly—insulate the modules from each other. Thus, even when modularization as presented here fails, the principal idea of the approach could still be employed. Conversely, it would be possible to extend our approach to allow for more than one low-confidence reactions in each module. Thus, by accepting a certain small (exponential) increase in the number of candidate models per module, the number of outputs that are necessary to be implemented experimentally could be significantly decreased.

SUPPORTING MATERIAL

Two figures, one algorithm, 12 equations, References (44,45) and supplemental information are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(13)04219-7.

We thank Mikolaj Rybinski for valuable discussions about reduced ordered multiple-valued decision diagrams.

We acknowledge financial support by the Swiss Initiative for Systems Biology SystemsX.ch evaluated by the Swiss National Science Foundation (project YeastX).

REFERENCES

1. Kirk, P., T. Thorne, and M. P. Stumpf. 2013. Model selection in systems and synthetic biology. *Curr. Opin. Biotechnol.* 24:767–774.
2. Kuepfer, L., M. Peter, …, J. Stelling. 2007. Ensemble modeling for analysis of cell signaling dynamics. *Nat. Biotechnol.* 25:1001–1006.
3. Szederkényi, G., J. R. Banga, and A. A. Alonso. 2011. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Syst. Biol.* 5:177.
4. Sunnåker, M., E. Zamora-Sillero, …, J. Stelling. 2013. Automatic generation of predictive dynamic models reveals nuclear phosphorylation as the key MSN2 control mechanism. *Sci. Signal.* 6:ra41.
5. Xu, T.-R., V. Vyshemirsky, …, W. Koch. 2010. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.* 3:ra20.
6. Turkheimer, F. E., R. Hinz, and V. J. Cunningham. 2003. On the undecidability among kinetic models: from model selection to model averaging. *J. Cereb. Blood Flow Metab.* 23:490–498.
7. Bachmann, J., A. Raue, …, U. Klingmüller. 2011. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.* 7:516.
8. Hartwell, L. H., J. J. Hopfield, …, A. W. Murray. 1999. From molecular to modular cell biology. *Nature.* 402:C47–C52.
9. Alexander, R. P., P. M. Kim, …, M. B. Gerstein. 2009. Understanding modularity in molecular networks requires dynamics. *Sci. Signal.* 2:pe44.
10. Kaltenbach, H.-M., and J. Stelling. 2012. Modular analysis of biological networks. *Adv. Exp. Med. Biol.* 736:3–17.
11. Ravasz, E., A. L. Somera, …, A.-L. Barabási. 2002. Hierarchical organization of modularity in metabolic networks. *Science.* 297:1551–1555.
12. Bowsher, C. G. 2011. Information processing by biochemical networks: a dynamic approach. *J. R. Soc. Interface.* 8:186–200.
13. Newman, M. E. 2006. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA.* 103:8577–8582.
14. Girvan, M., and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA.* 99:7821–7826.
15. DasGupta, B., and D. Desai. 2013. On the complexity of Newman's community finding approach for biological and social networks. *J. Comput. Syst. Sci.* 79:50–67.
16. Eisen, M. B., P. T. Spellman, …, D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.* 95:14863–14868.
17. Clauset, A., M. E. J. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 70:066111–066116.
18. Saez-Rodriguez, J., A. Kremling, and E. D. Gilles. 2005. Dissecting the puzzle of life: modularization of signal transduction networks. *Comput. Chem. Eng.* 29:619–629.
19. Saez-Rodriguez, J., S. Gayer, …, E. D. Gilles. 2008. Automatic decomposition of kinetic models of signaling networks minimizing the retroactivity among modules. *Bioinformatics.* 24:i213–i219.
20. Ederer, M., T. Sauter, …, F. Allgöwer. 2003. An approach for dividing models of biological reaction networks into functional units. *Simulation.* 79:703–716.
21. DasGupta, B., G. A. Enciso, …, Y. Zhang. 2007. Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. *Biosystems.* 90:161–178.
22. Kaltenbach, H.-M., S. Constantinescu, …, J. Stelling. 2011. Graph-based decomposition of biochemical reaction networks into monotone subsystems. *In* Algorithms in Bioinformatics Springer, New York, pp. 139–150.
23. Stelling, J., S. Klamt, …, E. D. Gilles. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature.* 420:190–193.
24. Klamt, S. 2005. Generalized concept of minimal cut sets in biochemical networks. *Biosystems.* 83:233–247.
25. Papin, J. A., J. L. Reed, and B. O. Palsson. 2004. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.* 29:641–647.
26. Kashtan, N., and U. Alon. 2005. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. USA.* 102:13773–13778.
27. Kholodenko, B. N., A. Kiyatkin, …, J. B. Hoek. 2002. Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA.* 99:12841–12846.
28. Sontag, E., A. Kiyatkin, and B. N. Kholodenko. 2004. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics.* 20:1877–1886.
29. Banga, J. R., K. J. Versyck, and J. F. van Impe. 2002. Computation of optimal identification experiments for nonlinear dynamic process models: a stochastic global optimization approach. *Ind. Eng. Chem. Res.* 41:2425–2430.
30. Faller, D., U. Klingmüller, and J. Timmer. 2003. Simulation methods for optimal experimental design in systems biology. *Simulation.* 79:717–725.
31. Chen, B. H., and S. P. Asprey. 2003. On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. *Ind. Eng. Chem. Res.* 42:1379–1390.
32. Kremling, A., S. Fischer, …, E. D. Gilles. 2004. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res.* 14:1773–1785.
33. Liepe, J., S. Filippi, …, M. P. Stumpf. 2013. Maximizing the information content of experiments in systems biology. *PLOS Comput. Biol.* 9:e1002888.
34. Reference deleted in proof.
35. Jefferys, W. H., and J. O. Berger. 1992. Ockham's Razor and Bayesian analysis. *Am. Sci.* 80:64–72.
36. Sedoglavic, A. 2001. A probabilistic algorithm to test local algebraic observability in polynomial time. *In* Proceedings of the 2001

International Symposium on Symbolic and Algebraic Computation Association for Computing Machinery, New York, pp. 309–317.

37. Cormen, T. H., C. E. Leiserson, …, C. Stein. 2001. Introduction to Algorithms, 2nd Ed. MIT Press, Cambridge, MA.
38. Srinivasan, A., T. Ham, …, R. K. Brayton. 1990. Algorithms for discrete function manipulation. *In* 1990 IEEE International Conference on Computer-Aided Design, ICCAD-90. Digest of Technical Papers Institute of Electrical and Electronics Engineers, New York, pp. 92–95.
39. Miller, D. 1993. Multiple-valued logic design tools. *In* Proceedings of The 1993 23rd IEEE International Symposium on Multiple-Valued Logic Institute of Electrical and Electronics Engineers, New York, pp. 2–11.
40. Miller, D., and R. Drechsler. 1998. Implementing a multiple-valued decision diagram package. *In* Proceedings of the 1998 28th IEEE International Symposium on Multiple-Valued Logic Institute of Electrical and Electronics Engineers, New York, pp. 52–57.
41. Li, C., M. Donizelli, …, C. Laibe. 2010. BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* 4:92.
42. Klingmüller, U., U. Lorenz, …, H. F. Lodish. 1995. Specific recruitment of SH-PTP1 to the erythropoietin receptor causes inactivation of JAK2 and termination of proliferative signals. *Cell.* 80: 729–738.
43. Ellson, J., E. R. Gansner, …, G. Woodhull. 2001. GRAPHVIZ—open source graph drawing tools. *In* Lecture Notes in Computer Science Springer-Verlag, Dordrecht, The Netherlands, pp. 483–484.
44. Strassen, V. 1969. Gaussian elimination is not optimal. *Numer. Math.* 13:354–356.
45. Higham, N. J. 2005. The scaling and squaring method for the matrix exponential revisited. *SIAM J. Matrix Anal. Appl.* 26:1179–1193.

Cutting the Wires: Modularization of Cellular Networks for Experimental Design

Moritz Lang,[†]* Sean Summers,[‡] and Jörg Stelling[†]

[†]Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich, and Swiss Institute of Bioinformatics, Basel, Switzerland; and [‡]Automatic Control Laboratory, Eidgenössische Technische Hochschule Zürich, Zurich, Switzerland

Supporting Material

Proof of Theorem 9

For *simple insulating modularization*s, there (i) exists a path of length 2 from the low-confidence reaction $R_j = l_R(j) \in V_{R,\eta}$, $j \in \{1, \ldots, |V_{R,\eta}|\}$, to species $S_i = l_S(i) \in V_S$, $i \in \{1, \ldots, n\}$ if the element a_{ij} of $A_{RS,\eta}$ is equal to one. There (ii) exists a path of length 3 from species $S_j = l_S(j)$ which is not measured ($S_j \in V_S \setminus V_{S,\eta}$) to species $S_i = l_S(i) \in V_S$ including only high-confidence reactions if element a_{ij} of $A_{RS,\not\eta} A_{SR,\not\eta} F$ is unequal zero. (iii) Species $S_j = l_S(j) \in V_S$ is measured if $\exists i : c_{ij} = 1$, $C = [c_{ij}]$.

A path from a low-confidence reaction $R_j \in V_{R,\eta}$ to a measured output $S_i \in V_{S,\eta}$ that does not contain any other low-confidence reactions or outputs is a combination of (i), an arbitrary amount of (ii), and (iii). Thus, such a path exists if and only if at least one element of the j^{th} column of

$$M_i = \begin{pmatrix} c_i^T \left[AF\right]^0 A_{RS,\eta} \\ c_i^T \left[AF\right]^1 A_{RS,\eta} \\ \vdots \\ c_i^T \left[AF\right]^{n-1} A_{RS,\eta} \end{pmatrix} \tag{1}$$

is unequal zero, with c_i^T the i^{th} row of C, and $A = A_{RS,\not\eta} A_{SR,\not\eta}$. Note that the longest possible simple path in a network with n species vertices is of length smaller or equal to $2n+1$, thus allowing M_i to be finite. The graph is a *simple insulating modularization* iff $|V_{S,\eta}| = |V_{R,\eta}| =: |\eta|$, and all matrices M_i, $i \in 1 \ldots |\eta|$, have at least one nonzero entry in the i^{th} column, and only zero entries in all other columns.

When discarding the explicit information about the length of the path, the requirement for *insulating modularization*s (Eq. 1) can be written more compactly as

$$D_\Sigma(C) = C \left(\sum_{k=0}^{m \geq n-1} a_k (AF)^k \right) A_{RS,\eta}, \tag{2}$$

with $a_k > 0$ arbitrary, positive constants. In this formulation, the graph is a *simple insulating modularization*, iff $|V_{S,\eta}| = |V_{R,\eta}| =: |\eta|$ and if the matrix $D_\Sigma(C)$ is a diagonal matrix with nonzero diagonal elements.

By choosing $a_k = \frac{1}{k!}$ and $m \to \infty$, we obtain

$$\boxed{D_{\Sigma,0}(C) = C e^{AF} A_{RS,\eta}} \tag{3}$$

where e is the matrix exponential defined by $e^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$.

For *strict insulating modularization*s, there (i) exists a path of length 2 from the low-confidence reaction $R_j = l_R(j) \in V_{R,\eta}$, $j \in \{1, \ldots, |V_{R,\eta}|\}$, to species $S_i = l_S(i) \in V_S$, $i \in \{1, \ldots, n\}$ if the element a_{ij} of $A_{RS,\eta}$ is equal to one. There (ii) exists a path of length 3 from species $S_j = l_S(j)$ which is not measured ($S_j \in V_S \setminus V_{S,\eta}$) to species $S_i = l_S(i) \in V_S$ including only high-confidence reactions if element a_{ij} of $A_{RS,\eta}A_{SR,\eta}F$ is unequal zero. (iii) There exists a path of length 2 from species $S_j = l_S(j)$ which is not measured ($S_j \in V_S \setminus V_{S,\eta}$) to the low-confidence reaction $R_i = l_R(i) \in V_{R,\eta}$, $i \in \{1, \ldots, |V_{R,\eta}|\}$, iff element a_{ij} of $A_{SR,\eta}F$ is unequal zero.

A path from a first low-confidence reaction $R_j \in V_{R,\eta}$ to a second low-confidence reaction $R_i \in V_{R,\eta}$ that does not contain any other low-confidence reactions or a measured output is a combination of (i), an arbitrary amount of (ii), and (iii). According to *simple insulating modularization*s, one can derive that such a path exists iff the element δ_{ij} of

$$\boxed{D_{\Sigma,0}(A_{SR,\eta}F) = A_{SR,\eta}Fe^{AF}A_{RS,\eta}} \tag{4}$$

is greater than zero. Since for a *strict insulating modularization* such a path must not exist for two distinct low confidence reactions ($\delta_{ij} = 0 \ \forall i \neq j$), $D_{\Sigma,0}(A_{SR,\eta}F)$ must be a diagonal matrix.

Alternative Representation of Theorem 9

Eq. 2 can be extended to

$$D_\Sigma(C) = a_0 CA_{RS,\eta} + CAF\left(\sum_{k=0}^{\infty} a_{k+1}(AF)^k\right)A_{RS,\eta} \tag{5}$$

By using the identity that $F = (I - C^TC) = (I - C^TC) \cdot (I - C^TC)$, one obtains

$$D_\Sigma(C) = a_0 CA_{RS,\eta} + CA\left(\sum_{k=0}^{\infty} a_{k+1}\left[FAF\right]^k\right)FA_{RS,\eta}. \tag{6}$$

Finally, by choosing $a_0 = 1,\ a_k = \frac{1}{(k-1)!}\ \forall k = 1, 2, \ldots$, we obtain a more "symmetric" (in terms of the exponent) version of the formula:

$$\boxed{D_{\Sigma,1}(C) = \underbrace{CA_{RS,\eta}}_{\substack{\text{feed-}\\\text{through}}} + \underbrace{CA}_{\substack{\text{observed}\\\text{inner}\\\text{dynamics}}} \underbrace{e^{FAF}}_{\substack{\text{inner}\\\text{dynamics}}} \underbrace{FA_{RS,\eta}}_{\substack{\text{non-feed-}\\\text{through}\\\text{inputs}}}}. \tag{7}$$

Note that–in general–$D_{\Sigma,0} \neq D_{\Sigma,1}$ due to the different choice for the values of a_k.

The symmetry of the exponent in Eq. 7 allows us to rewrite the formula in a computationally more efficient way. We consider the $(n-r \times n)$ $0-1$ matrix C_0 such that $(\tilde{C} = (C^T, C_0^T)^T)$ is orthonormal and has full rank, with $n = |V_S|$ and $r = |V_{R,\eta}|$. Additionally, we define $I_{r,n} = (I_{r,r}, 0_{n-r,r})^T$, with $I_{r,r}$ the $(r \times r)$ identity matrix and $0_{n-r,r}$ the $(n-r, r)$ matrix of zeros. With these two definitions, Eq. 7 can be rewritten into

$$
\begin{aligned}
D_{\Sigma,1}(C) =& CA_{RS,\eta} + CAe^{(I-C^TC)A(I-C^TC)}(I - C^TC)A_{RS,\eta} && \text{(8a)}\\
=& CA_{RS,\eta} + CA\left(e^{(I-C^TC)A(I-C^TC)} - C^TC\right)A_{RS,\eta} && \text{(8b)}\\
=& CA_{RS,\eta} + && \text{(8c)}\\
& CA\left(\tilde{C}^T\tilde{C}e^{(I-C^TC)A(I-C^TC)}\tilde{C}^T\tilde{C} - \tilde{C}^TI_{n,r}I_{r,n}\tilde{C}\right)A_{RS,\eta}\\
=& CA_{RS,\eta} + && \text{(8d)}\\
& CA\tilde{C}^T\left(e^{\tilde{C}(I-\tilde{C}^TI_{n,r}I_{r,n}\tilde{C})A(I-\tilde{C}^TI_{n,r}I_{r,n}\tilde{C})\tilde{C}^T} - I_{n,r}I_{r,n}\right)\tilde{C}A_{RS,\eta}
\end{aligned}
$$

Since $\tilde{C}$ is orthonormal $(C_0 \cdot C^T = 0)$, and the matrix in the exponential may have only non-zero elements in its lower-right $n-r \times n-r$ sub-matrix, this leads to:

$$
\boxed{D_{\Sigma,1}(C) = CA_{RS,\eta} + CAC_0^Te^{C_0AC_0^T}C_0A_{RS,\eta}} \tag{9}
$$

Eq. 9 is computationally advantageous over Eq. 7 because the matrix in the exponential is $(n-r \times n-r)$ instead of $(n \times n)$, thus reducing the computational costs for taking the matrix exponential.

NP-Hardness of Modularization Problems

To show that the modularization problem (Problem 11 in the main text) is NP-hard (see (1)), we define for the corresponding decision problem the formal language

Definition S1 (Modularizable)

$$
\begin{aligned}
MODULARIZABLE = \{&\langle G_{SR} = (V_S, V_R, E), \eta_R\rangle :\\
&\exists \eta_S, \textit{ such that } G_{IM} = (V_S, V_R, E, \eta_S, \eta_R)\\
&\qquad\qquad \textit{is a simple insulating modularization}\}\,.
\end{aligned}
$$

Definition S1 corresponds to deciding if at least one simple modularization exist for the corresponding modularization problem (Problem 11 in the main text). Clearly, a polynomial-time algorithm solving the modularization problem could be used to solve the decision problem in polynomial time, too, by simply checking if the set L_Σ of possible simple modularizations is empty or not:

$$\text{MODULARIZABLE} \leq_P \text{MODULARIZATION}. \tag{10}$$

However, the following theorem shows that a polynomial-time algorithm is unlikely to exist.

Theorem S2 *The modularizable problem is NP-complete.*

Theorem 12 in the main text follows because the modularizable problem is polynomial-time reducible to the modularization problem (Eq. 10).

Our proof for Theorem S2 is conceptually related to the proofs that the clique problem (1, page 1003ff), respectively the Hamiltonian-cycle problem (1, page 1008ff), are NP-complete. Furthermore, we utilize that the 3-conjunctive normal form (3-CNF) satisfiability problem is NP-complete (1, page 998ff). In the remainder of this section, we (i) shortly summarize the definition of the 3-CNF satisfiability problem, and (ii) utilize this satisfiability problem to proof Theorem S2.

3-CNF-SATISFIABILITY

The problem 3-CNF-SATISFIABILITY considers the decision problem if a Boolean formula $\phi(x_1, \ldots, x_n)$ in conjunctive normal form (CNF) with exactly three distinct literals l_1^r, l_2^r, and l_3^r in each of the k clauses C_r, $r \in 1, \ldots, k$, is satisfiable, that is, if at least one assignment (TRUE or FALSE) for the variables $x_1, \ldots, x_n$ exists such that ϕ evaluates to TRUE (1, page 998ff). In this definition, a literal is an occurrence of a variable x_j, $j \in 1 \ldots n$, or its negation $\neg x_j$. A clause C_r, $r \in 1, \ldots, k$, is the OR of one or more literals, and a Boolean formula in CNF is the AND of one or more clauses. For example,

$$\phi = \underbrace{(x_1 \vee \neg x_2 \vee \neg x_3)}_{C_1} \wedge \underbrace{(\neg x_1 \vee x_2 \vee x_4)}_{C_2} \wedge \underbrace{(x_1 \vee x_2 \vee x_4)}_{C_3} \tag{11}$$

is a 3-CNF Boolen formula with three clauses (C_1, C_2, and C_3) and six distinct literals (x_1, $\neg x_1$, x_2, $\neg x_2$, $\neg x_3$, and x_4).

Proof of Theorem S2

To prove Theorem S2, we have to show that MODULARIZABLE belongs to NP, and that deciding it is NP-hard. To show that MODULARIZABLE$\in$NP, for a given species reaction graph $G_{SR} = (V_S, V_R, E)$ and a low-confidence reaction label function η_R, we use the output label function η_S as a certificate. The verifying algorithm checks if $|\eta_R| = |\eta_S|$, and if the binary labeled species reaction graph $G_{BLSR} = (V_S, V_R, E, \eta_S, \eta_R)$ is a simple insulating modularization by utilizing the formulas given in Theorem 9 in the main text.

To prove that the decision problem MODULARIZABLE is NP hard, we show that 3-CNF-SATISFIABILITY$\leq_P$MODULARIZABLE. For this, we construct a SR-graph G_{SR} and a low-confidence reaction label function η_R for a given Boolean formula $\phi = C_1 \wedge C_2 \wedge \ldots \wedge C_k$ in 3-CNF and show that ϕ is satisfiable if and only if $\langle G_{SR}, \eta_R \rangle$ is modularizable.

Similar to (1, page 1008ff), we create a widget (a sub-graph enforcing certain properties; see Figure S1) for every clause C_r, $r \in 1, \ldots, k$, in ϕ. For each of the three literals l_1^r, l_2^r, and l_3^r in C_r, we create a low-confidence reaction node $R_{r,i}^O$, $i \in \{1, 2, 3\}$ as well as two species vertices $S_{r,i}^+$ and $S_{r,i}^-$ that correspond to the literal l_i^r, respectively its negation $\neg l_i^r$. Furthermore, we add a directed edge from $R_{r,i}^O$ to each species vertex $S_{r,i}^+$ and $S_{r,i}^-$. For each vertex $S_{r,i}^+$ (but not for $S_{r,i}^-$), we add a high-confidence reaction $R_{r,i}^C$, a species $S_{r,i}^C$, and the directed edges $(S_{r,i}^+, R_{r,i}^C)$ and $(R_{r,i}^C, S_{r,i}^C)$. Finally, we create one additional low-confidence reaction R_r^F per widget, and the three directed edges $(R_r^F, S_{r,i}^C)$, $i \in \{1, 2, 3\}$.

It is easy to validate that in a simple modularization, $\forall i \in \{1, 2, 3\}$ either $S_{r,i}^+$ or $S_{r,i}^-$ (but not both), as well as one of the nodes $S_{r,1}^C$, $S_{r,2}^C$, $S_{r,3}^C$ have to be assigned as a measured species: a module defined by the measured species $S_{r,i}^+$ or $S_{r,i}^-$ will always have $R_{r,i}^O$ in its interface, and a module defined by the measured species $S_{r,1}^C$, $S_{r,2}^C$, or $S_{r,3}^C$ will always have R_r^F in its interface. Note that such an assignment is only possible if at least one of the species $S_{r,i}^+$, $i \in \{1, 2, 3\}$, is measured: selecting $S_{r,i}^-$ and $S_{r,i}^C$ as measured species does not lead to a simple modularization because the module defined by $S_{r,i}^C$ contains at least two low confidence reactions ($R_{r,i}^O$ and R_r^F) in its interface (compare Lemma 5 in the main text).

Selecting species $S_{r,i}^+$ ($S_{r,i}^-$) as a measured output corresponds to the assignment that the corresponding literal l_i^r in the clause C_r evaluates to TRUE (FALSE). One has to select one of the species $S_{r,i}^C$, $i \in \{1, 2, 3\}$ as a measured species because at least one literal in every clause has to evaluate

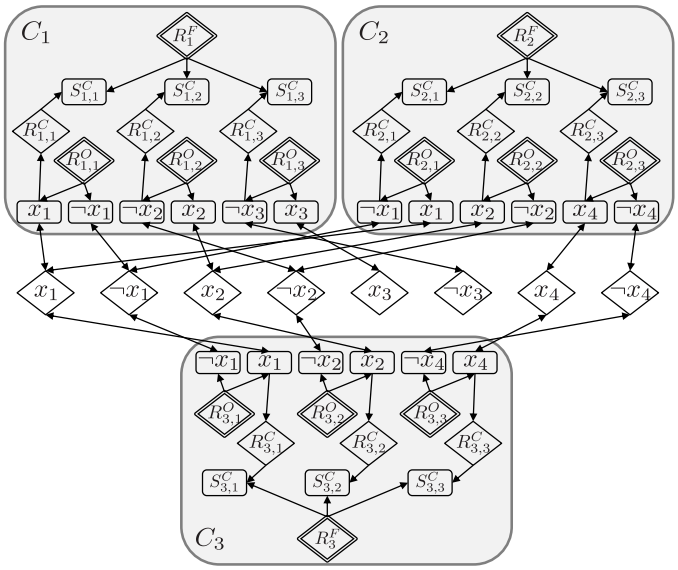

Figure S1: Reduction of an instance of the 3-CNF-SATISFIABILITY problem (Eq. 11) to an instance of the MODULARIZABLE problem. Box shaped vertices represent species, diamond shaped ones reactions. Low-confidence reaction vertices are marked by two borders. The light-gray boxes demarcate the widgets corresponding to the three clauses C_1, C_2 and C_3 in the Boolean formula ϕ. For convenience, the species vertices $S_{r,i}^+$ and $S_{r,i}^-$, $r \in 1, \ldots, k \wedge i \in \{1, 2, 3\}$ as well as the reaction vertices $R_{x,j}^+$ and $R_{x,j}^-$, $j \in 1, \ldots, n$ are labeled with their corresponding literals. For each widget, a simple modularization enforces for each literal that either the species corresponding to the literal or the species corresponding to its negation are measured, as well as that at least one species corresponding to one of the literals is measured. In a simple modularization, the high-confidence reactions connecting the widgets enforce that only sets of species are measured that correspond to a consistent TRUE assignment of the Boolean variables (respectively the literals) in and between the clauses.

to TRUE.

To enforce a consistent truth assignment of the Boolean variables (respectively the literals) in and between the clauses/widgets, we add two high-confidence reactions $R_{x,j}^+$, respectively $R_{x,j}^-$, for each Boolean variable x_j, $j \in 1, \ldots, n$, to the graph (see Figure S1), corresponding to the assignment $x_j = TRUE$, respectively $x_j = FALSE$. We add a bidirectional

edge (or two unidirectional edges in opposing directions) between a vertex in $\{S^+_{r,i}, S^-_{r,i} : r \in 1,\ldots,k \wedge i \in \{1,2,3\}\}$ and a vertex in $\{R^+_{x,j}, R^-_{x,j} : j \in 1,\ldots,n\}$ if the corresponding literal l^r_i is equivalent to the assignment of x_j. In a simple modularization, this enforces to either assign all species vertices adjacent to a reaction vertex $R^+_{x,j}$, respectively $R^-_{x,j}$, to be measured, or none: a partial assignment would lead to multiple low-confidence reactions in the interfaces of the corresponding modules (see Figure S1).

For a given 3-CNF-SATISFIABILITY problem with n Boolean variables and k clauses, our reduction algorithm described above creates a SR-graph with $2n + 7k$ reaction vertices ($4k$ of which are labeled low-confident), $9k$ species vertices, and $27k$ directed edges. Hence, the SR-graph G_{SR} and the low-confidence reaction label function η_R can be computed from a Boolean function ϕ in 3-CNF in polynomial time.

To show that the transformation of ϕ into (G_{SR}, η_R) is a reduction, we have to show that a satisfying assignment to the variables in ϕ corresponds to a simple modularization of (G_{SR}, η_R), and, conversely, that a simple modularization of (G_{SR}, η_R) corresponds to a satisfying assignment of the variables in ϕ. A satisfying assignment of ϕ directly corresponds to measuring either species $S^+_{r,i}$ or $S^-_{r,i}$, $r \in 1,\ldots,k$, $i \in \{1,2,3\}$, since for each literal in each clause either the literal or its negation is TRUE. In each clause at least one literal has to evaluate to TRUE, say l^r_j. Then, species $S^C_{r,j}$ can be assigned to be a measured. Finally, in each widget there is a consistent choice of measuring either $S^+_{r,i}$ or $S^-_{r,i}$, implying that all or none of the species adjacent to a reaction vertex $R^+_{x,j}$, $j \in 1,\ldots,n$, respectively $R^-_{x,j}$, are measured. Thus, the number of measured outputs is the same as the number of low-confidence reactions, and each module defined by a measured species has exactly one low-confidence reaction in its interface, corresponding to a simple modularization.

Conversely, if (G_{SR}, η_S, η_R) is a simple modularization, it is guaranteed that the truth assignments of the literals between the clauses is consistent; otherwise at least one module defined by a measured species $S^+_{r,i}$ or $S^-_{r,i}$ that has more than one low-confidence reaction in its interface would exist. Furthermore, in each widget either $S^C_{r,1}$, $S^C_{r,2}$, or $S^C_{r,3}$ is a measured species, say $S^C_{r,j}$, which implies that also $S^+_{r,j}$ is measured. Hence, in the respective clause at least the literal l^r_j evaluates to TRUE. Because the literals in the clauses are assigned consistently and at least one literal in each clause evaluates to TRUE, ϕ evaluates to TRUE corresponding to a satisfying assignment of the Boolean variables in ϕ.

Branch-and-Bound Algorithm for Modularization Problems

An exhaustive search to find all *insulating modularization*s for a given modularization problem would require to iterate over $\frac{|V_S|!}{(|V_S|-|\eta_R|)!}$ possible assignments of $|\eta_R|$ output labels to $|V_S|$ different species. However, $|\eta_R|!$ of these tests include the same set of outputs, albeit in different order. Computationally, the correct order of the measured outputs for an *insulating modularization* can be efficiently determined *a posteriori*, if the set of measured outputs is known. Thus, instead of directly searching tuples of outputs such that $D_{\Sigma,1}(C)$ is a diagonal matrix with non-zero diagonal entries (Eq. 9), we first search for sets of outputs such that $D_{\Sigma,1}(C)$ has exactly one non-zero element in each row and column, and afterwards we sort the outputs to fulfill the original condition. This reduces the number of necessary checks to $\binom{|V_S|}{|\eta_R|} = \frac{|V_S|!}{|\eta_R|!(|V_S|-|\eta_R|)!}$.

Checking if one output labeling function is part of the solution to a modularization problem—solving Eq. 9 for a given C—requires calculating two matrix multiplications and a matrix exponent (the costs for left or right multiplying a matrix X with C or C_0 are negligible). The two matrix multiplication require less than $O(|\eta_R|(|V_S| - |\eta_R|)^2)$ (2). The exponential of a matrix X can be precisely and efficiently calculated via Padé approximation with $\tau = 6 + \max\left(\left\lceil \log_2 \frac{\|X\|_\infty}{5.4} \right\rceil,\ 0\right)$ matrix multiplications (3). The value of τ depends on the maximal amount of inward connections of a vertex in the network, and, thus, scales with increasing connectivity of the network (usually $\tau < 10$). Each of these matrix multiplications has complexity $O((|V_S| - |\eta_R|)^3)$, such that an exhaustive search has complexity

$$O\left(\binom{|V_S|}{|\eta_R|} \cdot \left(2|\eta_R|(|V_S| - |\eta_R|)^2 + \tau(|V_S| - |\eta_R|)^3\right)\right). \tag{12}$$

In the following, we present our recursive branch-and-bound algorithmic solution for *simple insulating modularization*s (see main text for an intuitive description); if a given *simple insulating modularization* is *strict* can be easily checked with the formulas given in Theorem 9 in the main text, and the species and reactions belonging to a given module or interface can be obtained with the formulas given in Lemma 10 in the main text.

The complexity and, thus, the expected evaluation time of our recursive branch-and-bound algorithm highly depends on the specific modularization problem, and can only be upper bounded (see main text). However, to validate that our branch-and-bound algorithm performs significantly better than an exhaustive search for many modularization problems, we decided

Data: SR graph $G_{SR} = (V_S, V_R, E)$ defining the network, and the tuple $V_{R,\eta}$ of low-confidence reactions.
Result: Set L_Σ of all tuples of outputs leading to a *simple* modularization.

begin
  Create matrices $A_{SR,\not\eta}$, $A_{RS,\not\eta}$, $A_{SR,\eta}$, $A_{RS,\eta}$
  if *length* $V_{R,\eta}$ = *1* **then**
    $L_0 := ()$
  else
    $L_0 :=$ InsuMod($(V_S, V_R \setminus \{V_{R,\eta}(end)\}, E)$, $V_{R,\eta}$(1:end-1))
  end
  $L_\Sigma := \{\}$
  foreach $V_{S,\eta} \in L_0$ **do**
    foreach $S \in V_S \setminus V_{S,\eta}$ **do**
      $\tilde{V}_{S,\eta} := V_{S,\eta}$ concat (S)
      Construct matrix C from $\tilde{V}_{S,\eta}$
      Calculate $D_{\Sigma,0}$ (see Theorem 9)
      if $D_{\Sigma,0} = diag(\sigma_i)$, $\sigma_i > 0$ **then**
        $L_\Sigma := L_\Sigma \cup \{\tilde{V}_{S,\eta}\}$
      end
    end
  end
end

Function InsuMod(G_{SR}, $V_{R,\eta}$)

to compare the runtime of the two algorithms for automatically generated modularization problems of various complexity in $|\eta_R|$.

The structures of naturally evolved molecular signaling networks are constrained by their functionality. However, since these constraints are only poorly understood, it is not possible to automatically generate "typical" signaling networks for speed assessments of our algorithm. Therefore, we decided to take the network structure of the JAK2/STAT5 signaling model (4), and to generate in total 700 artificial modularization problems by randomly assigning low-confidence labels to the reactions in this model. We implemented our branch-and-bound algorithm and an exhaustive search in MATLAB (Release R2010a, The MathWorks, Natick, MA) and determined their computational times on an Intel Core 2 Duo, 3.16GHz, with 4GB RAM.

Fig. S2 shows that the computational time of the exhaustive search algorithm scales–as theoretical predicted (Eq. 12)–approximately exponentially with $|\eta_R|$. For less than two low-confidence reactions, the computation time of the exhaustive search is slightly lower than for our recursive branch-and-bound algorithm (both below 1 second). However, for more than two low-confidence reactions, the computational time required by the branch-and-bound algorithm seems to saturate, such that it significantly outperforms an exhaustive search for more complex modularization problems.

We also assessed the maximal, minimal, and mean number of possible distinct *insulating modularization*s for different numbers $|\eta_R|$ of low-confidence reactions, as well as the percentage of modularization problems for which at least one *insulating modularization* is possible (Fig. S2). As expected, for all modularization problems with $|\eta_R| = 1$ there exist 25 different modularizations, equal to the number of dynamic states (the concentration of *Epo* is not influenced by any reaction, and, thus, is constant in the model). This shows that the dual feedback mechanism in the model has as a consequence that the concentration of all species (except *Epo*) are–directly or indirectly–influenced by the turn-over of any reaction. For increasing numbers of low-confidence reactions in the network, the maximal number of possible modularizations increases due to combinatorial explosion, whereas the percentage of modularization problems having a non-empty solutions decreases. Note that for $|\eta_R| = 7$ already around a fifth of all reactions are marked as being low-confident, and that in a valid *insulating modularization* more than a quarter of all states have to be measured. As stated in the main text, our modularization approach was designed for relatively well-known networks. Thus, it is rather surprising that still more than 10% of all randomly generated modularization problems with $|\eta_R| = 7$ have a non-empty

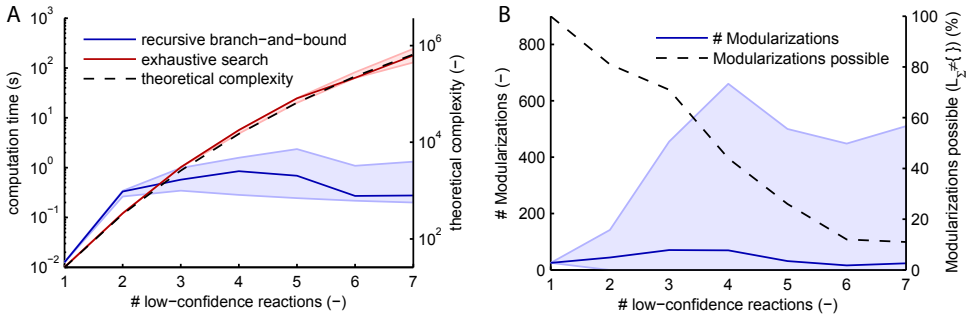

Figure S2: Evaluation of the branch-and-bound algorithm. (A) Computational time of an exhaustive search (red; median, 25% and 75% quantiles) and of our branch-and-bound algorithm (blue) to solve a modularization problem with $|\eta_R|$ low-confidence reactions generated as described in the text, compared to the theoretically predicted complexity (black, dashed; compare Eq. 12). Note the logarithmic scale of the y-axis. (B) Maximal, minimal, and mean number of modularizations found by either algorithm (blue), and percentage of modularization problems with a non-empty solution (black, dashed), i.e., for which at least one possible modularization exists. Both plots are based on 100 randomly generated modularization problems for each value of $|\eta_R|$.

solution. In this assessment, the majority of modularization problems with less than 10% of reactions marked as being low-confident has a non-empty solution. In reality, when encountering modularization problems with very high numbers of low-confidence reactions compared to the total number of reactions and species, one should consider merging several low-confidence reactions into one, especially if they are closely related, i.e. belong to a single hypothetical network extension.

It is important to note that our evaluation of the required computational time, as well as of the number of possible distinct *insulating modularization*s for different numbers $|\eta_R|$ of low-confidence reactions, highly depends on the specific way to generate the modularization problems. In general, we expect modularization problems in, for example, highly connected protein-protein interaction networks to have fewer possible solutions, and problems in networks including, for example, many non-reversible transcription and translation reactions to have higher probability that at least one possible modularization exists. The model of Bachmann et al. (4) can be seen as an intermediate between these two extremes since it includes protein-protein interactions at the *Epo* receptor complex as well as transcription and trans-

lation of *socs*3 and *cis*. Note, however, that the evaluations of computational time and number of possible modularizations for this specific problem is meant to provide intuition for our modularization approach, rather than to represent an exhaustive analysis.

Construction of Models

In this section, we shortly describe how to create the models with and without the low-confidence reaction of a module. These models can be used for the assessment of the existence of the respective low-confidence reaction using, for instance, Bayesian inference (5). Here, we assume that a model of the full network is given, as well as that an insulating modularization was already identified using our branch-and-bound algorithm. Furthermore, we assume that experimental time-series data $\{y_{it}\}_{t\in T_i}$ of each measured output $S_i \in V_{S,\eta}$ is available.

For simple modularizations, to construct the model of the i^{th} module without the low-confidence reaction, we utilize the formulas given in Lemma 10 in the main text to determine the species and reactions belonging to the module. All species (and their initial conditions) and the reactions with rate equations only depending on the species in the module are simply taken over from the model of the full network. For reactions with rates depending on the concentrations of species not in the module, the corresponding term in the rate equation is replaced by the respective measurement data $\{y_{it}\}_{t\in T_i}$, or by an appropriate spline approximation of the measurement data for continuous models. This is possible because all species on which the rate of a reaction in the module might depend are, by Definition 4 in the main text, either part of the module or of its interface, and all species in the interface of a module are measured outputs (Lemma 5 in the main text).

For modules of strict modularizations, also models can be constructed including the respective low-confidence reaction. To identify the species and reactions belonging to this model, we remove the low-confidence label of the respective reaction, that is, we append the column (row) of $A_{RS,\eta}$ ($A_{SR,\eta}$) corresponding to the low-confidence reaction to the matrix $A_{RS,\not\eta}$ ($A_{SR,\not\eta}$), and apply the formulas given in Lemma 10 in the main text (without recalculating the outputs). Given the species and reactions which belong to the model, we proceed as described above. Note that, by Definition 3b in the main text, the concentration of none of the species in this model is influenced by any other low-confidence reaction.

The models constructed as described above do not depend on any species

or reactions not in the module, but only on experimental measurement data that is used for the virtual inputs. Thus, it is possible to simulate the models and compare them to the experimental measurement data of the respective output separately, and in any order: the models of the modules are insulated from each other by using the concept of virtual inputs. For strict modularizations, if the models of different modules do not share common parameters, which is given if the modules do not overlap, the probability for the existence of one low-confidence reaction becomes conditionally independent of the existence of all other low-confidence reactions by applying our modularization approach, as stated in the main text.

Supporting References

1. Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein, 2001. Introduction to algorithms. The MIT press, Cambridge, Massachusetts, second edition.

2. Strassen, V., 1969. Gaussian elimination is not optimal. *Numer. Math.* 13:354–356.

3. Higham, N. J., 2005. The scaling and squaring method for the matrix exponential revisited. *SIAM J. Matrix Anal. Appl.* 26:1179–1193.

4. Bachmann, J., A. Raue, M. Schilling, M. Bohm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. Lehmann, J. Timmer, and U. Klingmuller, 2011. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol Syst Biol* 7.

5. Xu, T.-R., V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, G. S. Baillie, D. Ketley, A. J. Dunlop, G. Milligan, M. D. Houslay, and W. Kolch, 2010. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci Signal* 3:ra20.