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

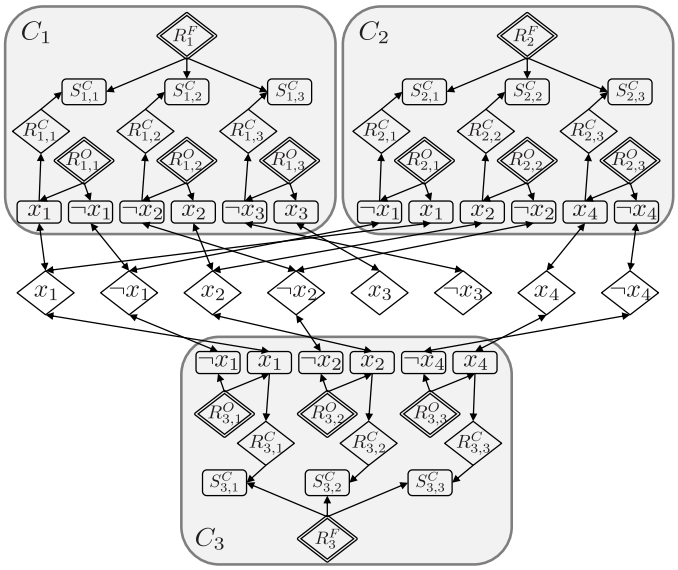

Figure S1: Reduction of an instance of the 3-CNF-SATISFIABILITY problem (Eq. 11) to an instance of the MODULARIZABLE problem. Box shaped vertices represent species, diamond shaped ones reactions. Low-confidence reaction vertices are marked by two borders. The light-gray boxes demarcate the widgets corresponding 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

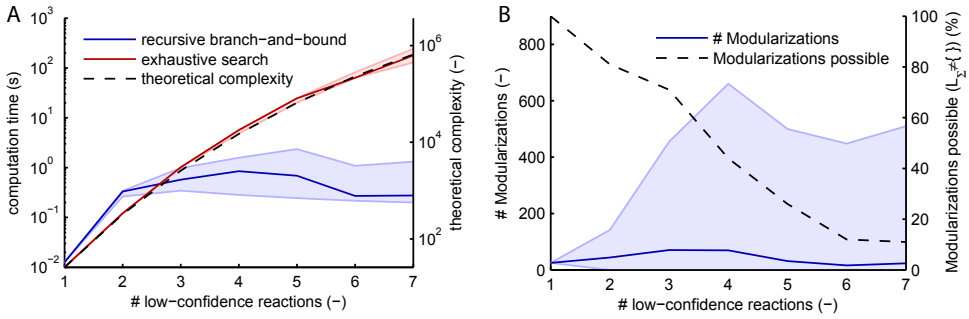

Figure S2: Evaluation of the branch-and-bound algorithm. (A) Computational time of an exhaustive search (red; median, 25% and 75% quantiles) and of our branch-and-bound algorithm (blue) to solve a modularization problem with $|\eta_R|$ low-confidence reactions generated as described in the text, compared 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.