Electronic Supplementary Material (ESI) for Molecular BioSystems

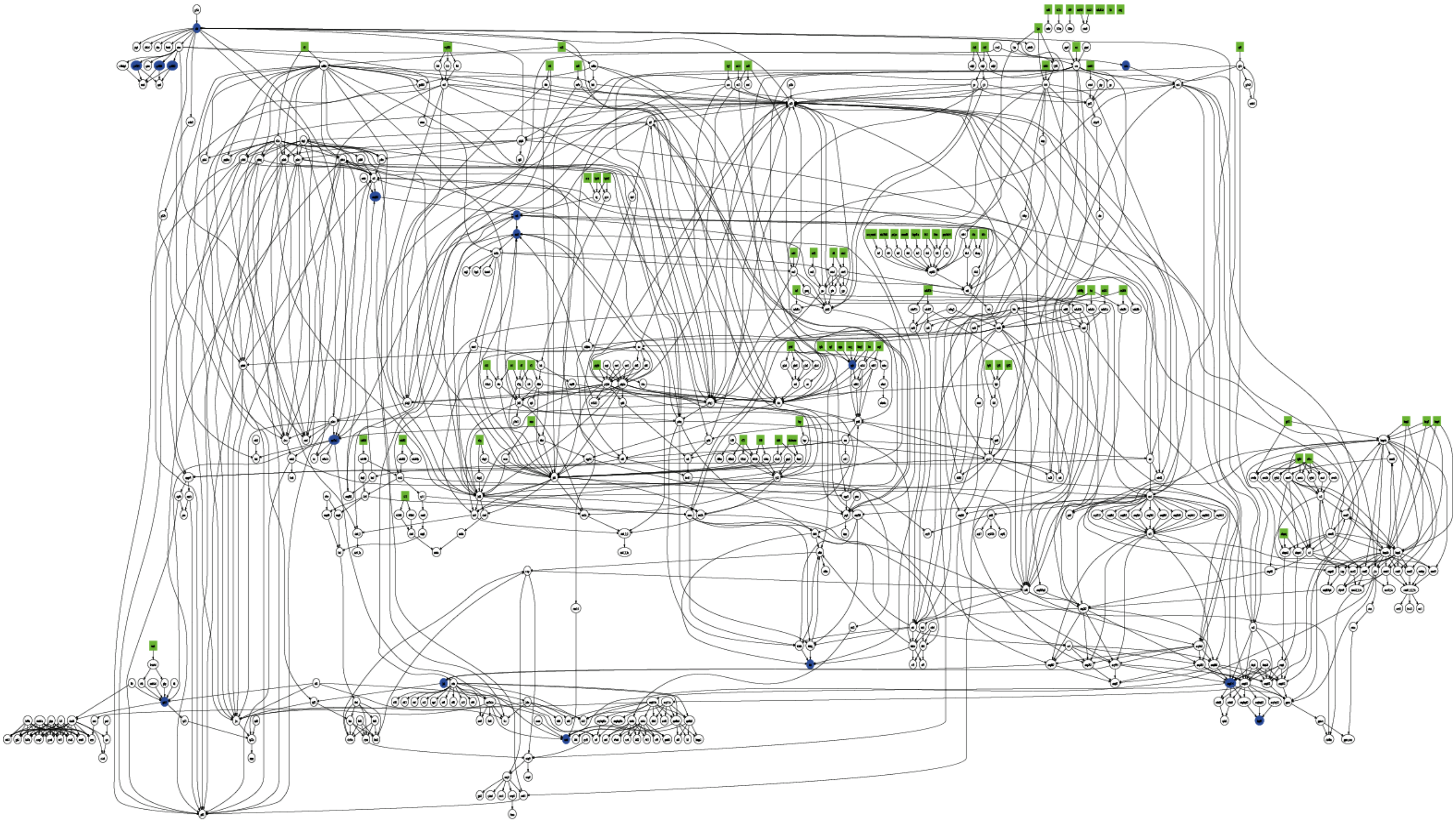
SUPPLEMENTARY MATERIAL FOR:

Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data

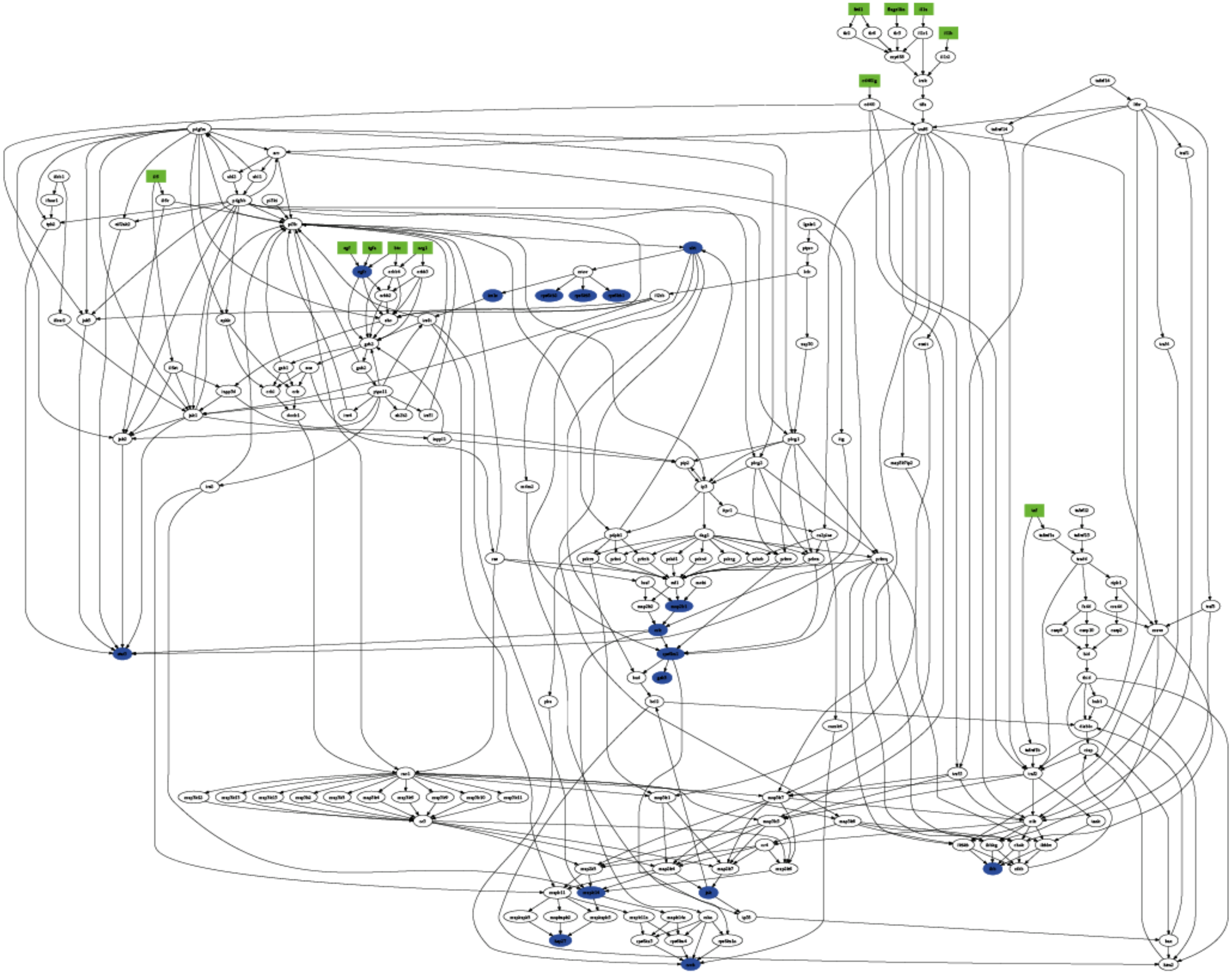
Ioannis N. Melas, Alexander Mitsos, Dimitris E. Messinis, Thomas S. Weiss, Julio-Saez Rodriguez, Leonidas G. Alexopoulos

Contents:

Supplementary Material 1 – Illustration of canonical and observable-controllable pathways



Supplementary Figure 1: Canonical pathway constructed from literature. Numbers 533 nodes, 1064 reactions and serves as starting point for the analysis described in this paper.

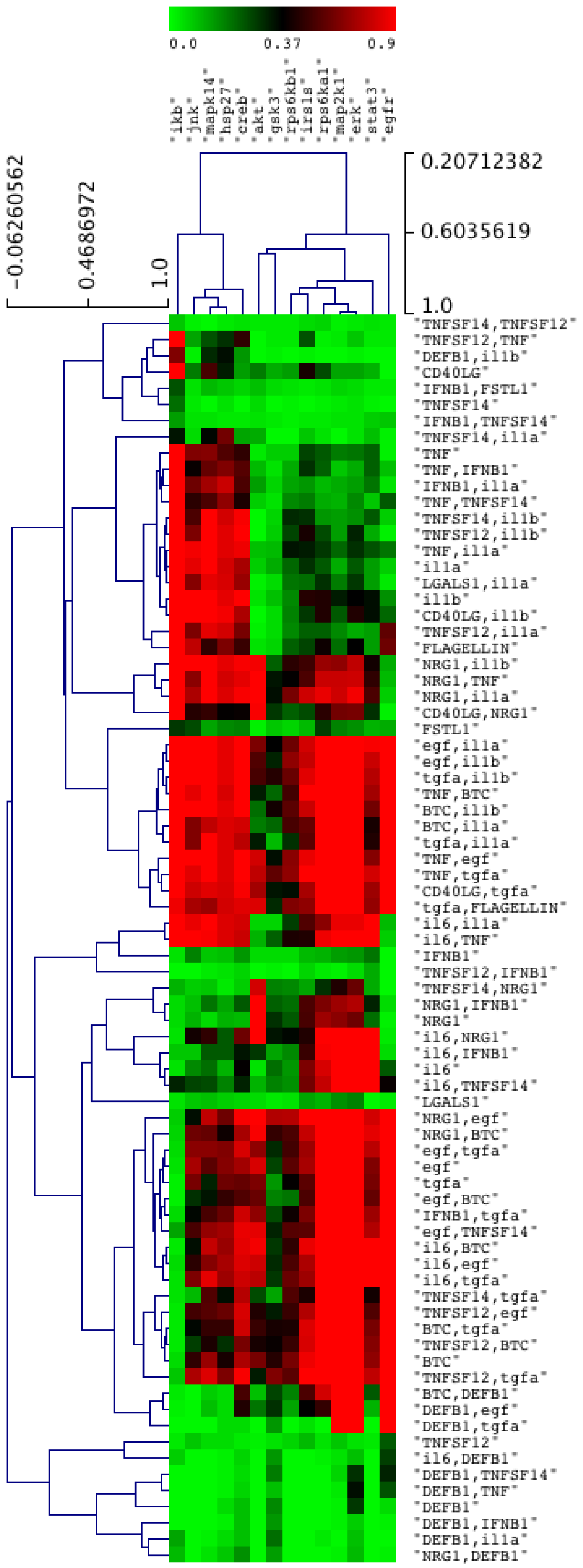


Supplementary Figure 2: Observable-controllable part of the canonical pathway. Numbers 177 species, 365 reactions and is the pathway to be optimized by the ILP algorithm.

Supplementary Material 2 – Clustering analysis of the full combinatorial dataset

Additionally to the clustering analysis presented in the main text (see Results section, Clustering) where the hierarchical clustering of the single treatments part of the combinatorial dataset was illustrated, here we provide the clustering of the full combinatorial dataset (Supplementary Figure 3). Although manual inspection of the results is difficult, a few basic trends can be identified: The signals' hierarchical tree remains the same as before, with signals in the pro-inflammatory pathways (IKB, JNK, MAPK14, HSP27 and CREB) clustered together and in similar fashion, signals in the pro-growth pathways (AKT, GSK3, RPS6KB1, IRS1S, RPS6KA1, MAP2K1, ERK, STAT3 and EGFR) clustered together. Moreover, a few major clusters can be identified in the ligands tree: Starting from the top, (i) the combinations of TNFSFS12, TNFSF14, DEFB1, FSTL1, CD40LG and IFNB1 are clustered together, since all of them slightly activate the pro-inflammatory signals cluster. (ii) Combinations of TNF, IL1A, IL1B and FLAGELLIN are clustered together for activating at saturation level the pro-inflammatory cluster; (iii) combinations of strong pro-inflammatory with pro-growth activators are clustered together for activating most of the measured signals at saturation level. (iv) Combinations of pro-growth activators with themselves or with weak pro-inflammatory activators are also clustered together for activating mostly the pro-growth signals.

The clustering of the full combinatorial dataset confirms the analysis presented in the main text of this paper, while no novel synergistic effects could be identified.



Supplementary Figure 3: Hierarchical clustering of the full combinatorial dataset.

Supplementary Material 3 – Removal of feedback loops

Presence of feedback loops presents an obstacle for optimizing Boolean topologies, since it leads to a loop that requires no stimulus for activation. Thus, an important step before optimization is the identification and removal of feedback loops. Even though removal of reactions from the generic topology should have been avoided, the feedback loops usually are considered "late events" that take place after activation. Since our pathways are trained with "early response" experimental data, reactions that take place at later stages can be removed.

Without loss of generality we assume a positive feedback loop numbering n species.

A_1 -> A_2 -> ... -> A_n -> A_1

Equations (3) to (6) in Materials and Methods, here repeated for consistency, are:

$z_i^k \text{ £ } x_j^k,\ i=1,\ldots,n_r,\ k=1,\ldots,n_e,\ j \text{ Î } R_i$ (3) , R is the set of reactants in reaction i

$z_i^k \text{ £ } 1-x_j^k,\ i=1,\ldots,n_r,\ k=1,\ldots,n_e,\ j \text{ Î } I_i$ (4) , I is the set of inhibitors in reaction i

$z_i^k \text{ ³ } y_i + \text{å}(x_j^k - 1) - \text{å}\, x_j^k,\ i=1,\ldots,n_r,\ k=1,\ldots,n_e$ (5)

$x_j^k \text{ ³ } z_i^k\ \ i=1,\ldots,n_r,\ k=1,\ldots,n_e,\ j \text{ Î } P_i$ (6), P is the set of products in reaction i

z_i^k , expresses the activation of reaction i in experiment k (assumes only Boolean values)

(3), (4) and (5) imply that reaction i will take place if at least a reactant is present and no inhibitors are present, (6) implies that if reaction i takes place all products are formed. If a loop occurs in the initial topology A_1 -> A_2 -> ... -> A_n -> A_1, then starting from an arbitrary species in the loop A_1 and setting, $x_{A1}=1$, (3),(4),(5)=>$z_{An\text{->}A1}=1$, (6)=>$x_{An}=1$, $x_{An-1}=1$, ..., $x_{A2}=1$, $x_{A1}=1$.
In other words, species A_1 activates itself which is not desirable.

To prevent that, all feedback loops were identified beforehand using a custom depth-first search (DFS) algorithm. Each loop is characterized by n_L reactions with indices in the index set $L=\{1,\ldots,n\}$. For each such loop the following constraint was added prohibiting the ILP from including the loop in the solution:

$$\sum_{i \in L} y_i < n_L \ . \qquad (7)$$

The optimizer is thus forced to delete at least one of the reactions in the loop.

Supplementary Material 4 – Construction of toy model and performance assessment

To validate the performance of the optimization algorithm a toy model consisting of 29 nodes and 35 reactions is constructed as an oversimplification of the canonical pathway. The toy model includes only 5 stimuli (TGFA, BTC, NRG1, IL1B, IL1A) and 3 signals (MAP2K1, AKT, IKB) (see Supplementary Material 3, Figure 3A) and serves to better illustrate the difference between positive and negative size weights, as well as the execution of cross-validation studies. The main topological features include the activation of MAP2K1 and AKT from TGFA, BTC and NRG1 via a number of alternative pathways, the activation of AKT from IL1A and IL1B via SRC, and the activation of IKB from TGFA, BTC, NRG1 via MAP3K8 and from IL1A, IL1B via TRAF6. An accompanying dataset is also constructed, consisting of single treatments of the above mentioned stimuli in a total of 6 experimental conditions (including the no-inhibitor experiment). The ILP prunes the initial topology to best fit the dataset at hand by minimizing the objective function:

$\sum a_j^k \,|\, x_j^k - x_j^{k,m} \,|+\sum b_i y_i$ where,

$j=\{1,\ldots,n_s\}$ are the species (nodes) included in the pathway,

$i=\{1,\ldots,n_r\}$ are the reactions,

$k=\{1,\ldots,n_e\}$ are the experiments,

x_j^k is the simulated value of species *j* in the experiment *k*,

$x_j^{k,m}$ is the measured value of species *j* in experiment *k*,

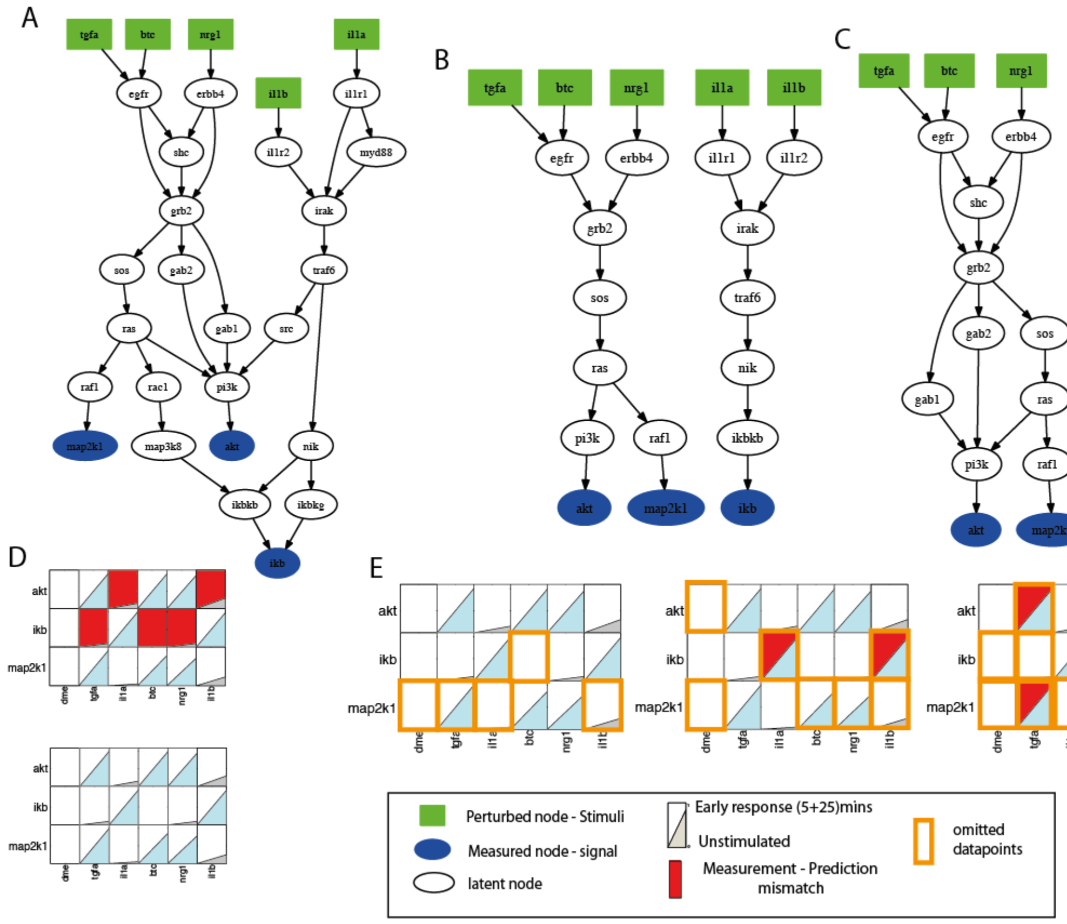
a_j^k are weights (for species *j* in experiment *k*)

y_i is the presence (or absence) of reaction *i* (takes Boolean values only)

b_i are weights (for reaction *i*)

The toy model is optimized using two different settings: (i) small positive reaction weights (b_i) enforcing the minimization of the pathway size together with the experiments-topology mismatch (see Supplementary Material 3, figure 3B). (ii) small negative reaction weights (b_i) enforcing the maximization of the pathway size while minimizing the experiments-topology mismatch. This setting results in the superset of possible solutions (see Supplementary Material 3, figure 3C). It becomes apparent that both solutions share the same basic connectivity patterns, TGFA, BTC and NRG1 activate MAP2K1 and AKT, IL1A and IL1B activate IKB. Their difference lies in the fact that the superset of possible solutions includes all paths fitting these patterns while pathway in Supplementary Material 3, figure 3B includes only the shortest. No unresolved fitness error is observed (for both solutions).

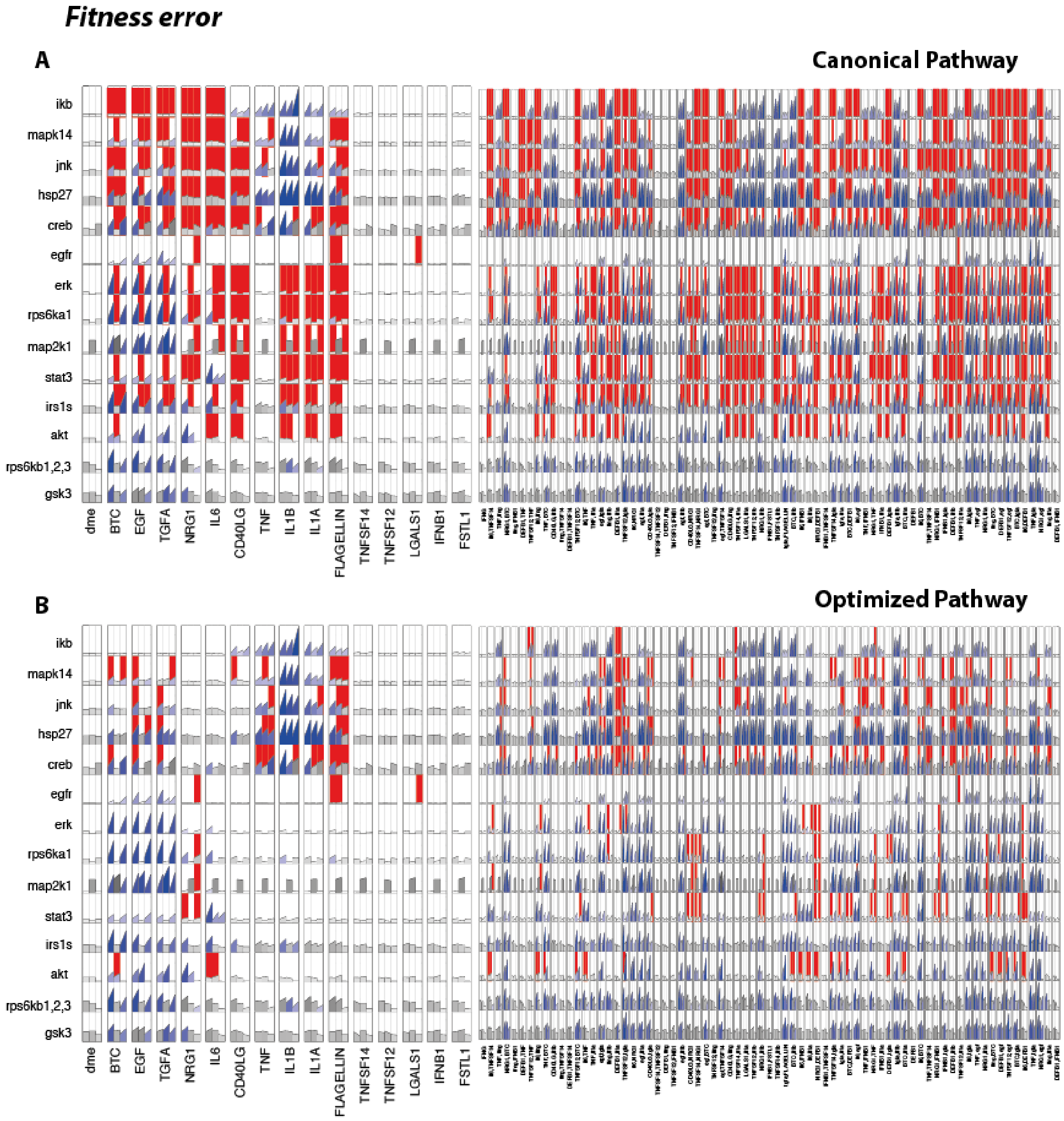
A better assessment of how the ILP formulation performs, is obtained by omitting datapoints from the initial dataset and monitoring the remaining fitness error. 3 subsets of the initial dataset were constructed by omitting measurements in a random manner. The toy model was then trained and the unresolved fitness error was plotted (red background in Supplementary Material 3, Figure 3E). In the first of these plots, although 5 of the original 18 datapoints were left out of the dataset no experiments-topology mismatch is observed in the solution. The reason for such a solid performance under missing measurements can be traced to the high degree of overlap between the pathways. In the toy model TGFA, BTC and NRG1 signal via almost identical pathways (same for IL1A and IL1B) creating many internal replicates, allowing the removal of a substantial part of the dataset without affecting the goodness of fit of the solution. The same applies for the large pathway, as it was mentioned in the Discussion part, resulting in a procedure very robust against missing measurements and incomplete datasets. As we gradually omit more measurements i) internal replicates are no longer present, ii) subsets of the topology become non-observable and are removed altogether, iii) and the solution fails to fit the latent signals.



Supplementary figure 3: Toy model – performance assessment. (a) The toy model consists of 5 stimuli, 3 measured signals and a total of 29 nodes and 35 reactions. (b) The toy model is optimized by the ILP using two different settings, first using a small positive reactions weight, resulting in a minimum-size solution, and second (c) using a small negative reactions weight, resulting in a maximum-size solution. (d) Topology – prediction mismatch of the original and optimized pathway. (e) Topology - prediction mismatch after optimizing the toy model with subsets of the original dataset.

Supplementary Material 5 – Inspection of fitness error before and after the optimization procedure

The experiments-topology mismatch is illustrated in Suppl. Figure 4, before and after optimization procedure. The ILP formulation has decreased the fitness error from 31% to 7%. As illustrated in Suppl. Figure 4, there are 2 main areas where the error has been removed drastically : (i) measurement of pro-inflammatory signals (IKB, MAPK14, JNK, HSP27, IKB, CREB) under pro-growth treatments (BTC, EGF, TGFA, NRG1, IL6) and (ii) measurement of pro-growth signals (ERK, RPS6KA1, MAP2K1, AKT, IRS1S, STAT3) under pro-inflammatory treatments (TNF, IL1A, IL1B, CD40LG, FLAGELLIN). Unresolved fitness error is identified mostly under FLAGELLIN, because of partial activation of MAPK14, JNK, HSP27 and CREB. According to the generic topology FLAGELLIN pathway overlaps with the rest of the pro-inflammatory stimuli (IL1A, IL1B, TNF and CD40LG) that fully activate the above mentioned signals. Therefore, minimization of the objective function implies fitting IL1A, IL1B, TNF, CD40LG by including paths to MAPK14, JNK, HSP27 and CREB, and misfitting FLAGELLIN. The same can be observed for IL6 induced AKT activation.



Supplementary Figure 4: Fitness error before and after the optimization procedure. (a) Fitness error before the optimization procedure. (b) Fitness error after the optimization procedure