*Bioinformatics*, 32, 2016, i297–i305
doi: 10.1093/bioinformatics/btw268
ISMB 2016

OXFORD

# Linear effects models of signaling pathways from combinatorial perturbation data

**Ewa Szczurek[1,*] and Niko Beerenwinkel[2,3,*]**

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland, [2]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland and [3]SIB Swiss Institute of Bioinformatics

*To whom correspondence should be addressed.

## Abstract

**Motivation:** Perturbations constitute the central means to study signaling pathways. Interrupting components of the pathway and analyzing observed effects of those interruptions can give insight into unknown connections within the signaling pathway itself, as well as the link from the pathway to the effects. Different pathway components may have different individual contributions to the measured perturbation effects, such as gene expression changes. Those effects will be observed in combination when the pathway components are perturbed. Extant approaches focus either on the reconstruction of pathway structure or on resolving how the pathway components control the downstream effects.

**Results:** Here, we propose a linear effects model, which can be applied to solve both these problems from combinatorial perturbation data. We use simulated data to demonstrate the accuracy of learning the pathway structure as well as estimation of the individual contributions of pathway components to the perturbation effects. The practical utility of our approach is illustrated by an application to perturbations of the mitogen-activated protein kinase pathway in *Saccharomyces cerevisiae*.

**Availability and Implementation:** lem is available as a R package at http://www.mimuw.edu.pl/~szczurek/lem.

**Contact:** szczurek@mimuw.edu.pl; niko.beerenwinkel@bsse.ethz.ch

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Signaling pathways convey stimuli from the outside or inside of the cell to generate required cellular response. For example, under osmotic stress, the high osmolarity glycerol (HOG) mitogen-activated protein kinase (MAPK) pathway in yeast is activated, and the signal is transported from the receptors down to the MAPKK Pbs2, which in turn phosphorylates the MAPK Hog1. Finally, Hog1 regulates several transcription factors, which activate the hyper-osmotic stress response genes (Hohmann, 2002; O'Rourke and Herskowitz, 2004).

The central means to study signaling pathways is by cellular perturbations. Hence, computational analysis, modeling and interpretation of perturbation data constitute the crucial tools in the field (Markowetz, 2010). Examples of experimental perturbations include CRISPR-Cas genome editing, genetic knock-outs (gene deletions) or transcriptional knock-downs through RNA interference (RNAi). For the HOG pathway in *Saccharomyces cerevisiae*, its components were perturbed by deletion and the effect of these perturbations was assessed in osmotic stress conditions by measuring global expression changes between perturbed and wild-type cells (O'Rourke and Herskowitz, 2004). The rationale behind the perturbation studies is that interrupting the signal flow in the pathways gives insight into both their structure and their downstream targets. First, with the interruption at a certain node in the pathway, the signal cannot be transmitted further. For example, when the MAPKK is deleted, the MAPK downstream cannot be phosphorylated. In this way, perturbations propagate in the pathway along its edges in the same way as the signal. Second, each node in the pathway may have its own (direct or indirect) contribution to the perturbation effects, such as gene expression changes. Those effects will be observed in combination when the pathway components are perturbed. The focus of the present work is computational analysis of cellular signaling pathways from perturbation data. The aim is to model the structure of interactions within the pathway as well as the contributions of its components to the observed perturbation effects.

The proposed approach is motivated by two problems that are inherent to the analysis of perturbation data. The first problem concerns the perturbation-effect gap problem, a common discrepancy between the perturbed and the observed variables. In most experimental studies, the perturbed signaling pathway constitutes one layer of the system (layer 1 in Fig. 1A), only indirectly connected to

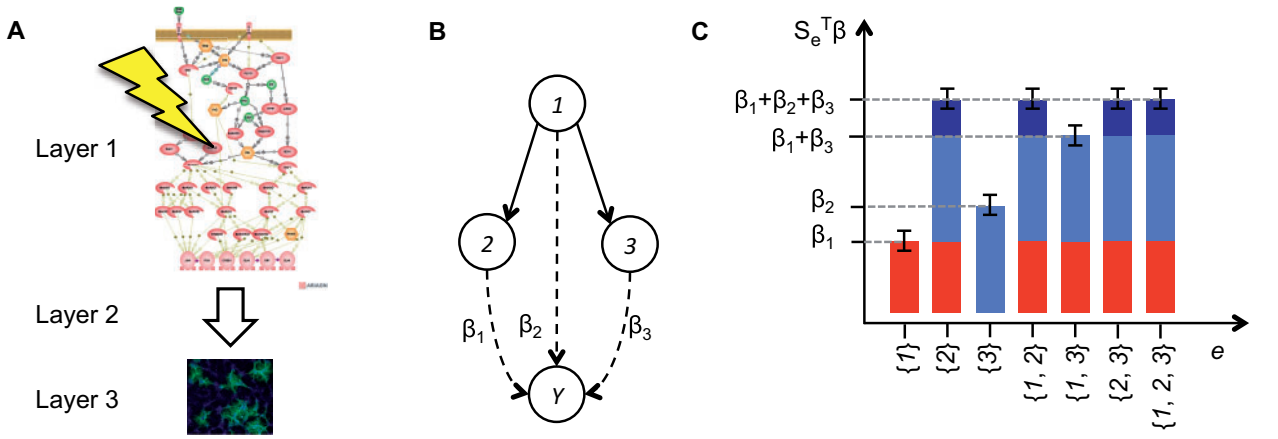


**Fig. 1.** Linear effects model. (**A**) Three layers of the system: 1 perturbed signaling pathway, 2 intermediate and 3 observed effects. (**B**) Genes (circles, here *1, 2, 3*) are directly or indirectly (via propagation in the pathway) perturbed in experiments. Bold arrows indicate how perturbations propagate within the pathway. Dashed arrows show the individual contributions of the genes to the observed perturbation effects *Y*. LEM assumes that *Y* is normally distributed around the mean equal to the weighted sum of individual gene effects (here $\beta_1$, $\beta_2$, $\beta_3$), with weights set to perturbation states. (**C**) Example means (*y*-axis) for all possible perturbation experiments (*x*-axis), as expected in the LEM with pathway structure as in B, and with $\beta_1, \beta_2$ and $\beta_3$ values indicated by red, light blue and dark blue bars, respectively. Whiskers indicate example error.

a distinct layer of measured effects (layer 3). Usually, the states of signaling genes are hidden, and their interconnections, i.e. the structure of the pathway is unknown. Layer 2 in Figure 1A corresponds for example to gene regulation, where pathway components control gene expression downstream, and changes of this expression due to perturbations are observed. The second problem is that gene expression is rarely controlled by a single gene in the pathway. Instead, the measured effects arise from a concerted contribution of several pathway components. In addition, gene regulators consist not only of transcription factors, placed in the 'bottom' of the pathway, but also can include their upstream kinases, which via regulation of other factors have their independent contribution to the effects.

Probabilistic graphical models were successfully applied to perturbation data previously (Pe'er et al., 2001; Friedman, 2004; Rogers and Girolami, 2005; Sachs et al., 2005; Gat-Viks *et al.*, 2006; Markowetz *et al.*, 2007; Ellis and Wong, 2008; Fröhlich et al., 2009a; Bender *et al.*, 2010). In these studies, however, the perturbation-effect gap problem was not addressed, since the signaling pathway variables are considered observed and not hidden. Either the data explicitly reported signaling variables as protein concentrations (Sachs *et al.*, 2005; Fröhlich *et al.*, 2009; Bender *et al.*, 2010), or expression levels were used as a proxy for the states of signaling genes. Nested effects models (NEMs) (Markowetz *et al.*, 2005; Markowetz *et al.*, 2007; Tresch and Markowetz, 2008; Fröhlich *et al.*, 2008, 2009b) and their extensions (Anchang *et al.*, 2009; Fröhlich *et al.*, 2011; Siebourg-Polster *et al.*, 2015) specifically address the perturbation-effect gap problem. NEMs are represented by directed graphs, and have distinct nodes for the variables representing signaling genes and for the downstream effects. The crucial assumption behind NEMs is that perturbation effects show a nested subset hierarchy which reflects the hierarchy of nodes in the signaling network. With a simple, deterministic model of signaling pathways, and with a probabilistic take on the observed effects, NEMs constitute an attractive approach for learning their structure. One disadvantage of this model, however, is an assumption that each effect gene is regulated only by a single gene in the pathway. Moreover, the discrepancy between model predictions and observed effects is evaluated assuming fixed noise levels as parameters, which cannot be estimated from the data together with the model, but as a preprocessing step. Finally, other previous studies concentrated solely on elucidating the link between the pathway and the observed effects (layer 2 in Fig. 1A; Gat-Viks and Shamir, 2007; Szczurek *et al.*, 2009). These approaches worked with an a priori known and given pathway graph, and aimed at either small refinements to the known graph, or resolving the detailed mechanisms governing the regulation of the downstream targets by the pathway components, representing them as logic functions or discrete probability distributions.

Here, we propose a linear effects model (LEM) for modeling perturbation data that addresses both problems and can be applied to learn the structure of signaling networks together with individual contributions of their genes to the perturbation effects. The model contains a deterministic graph component representing how perturbations propagate within the signaling pathway. It assumes that the observed perturbation effects amount to a linear combination of the individual contributions of the perturbed pathway genes. Model inference does not require parameter estimation as a preprocessing step. Instead, inference of the pathway graph is performed within a Bayesian framework, together with the inference of hyper parameters defining distributions of the parameters of the model, including the noise distribution. We prove that for identifiability LEM requires perturbations of all single and all pairs of nodes in the pathway. Tests on simulated data demonstrated high accuracy of parameter estimation and excellent recovery of pathway structures already with small numbers of repeated experiments, and within a wide range of noise levels. In application to Δ*pbs*2 and Δ*hog*1 mutant data, LEM correctly identified the signal flow between Pbs2 and Hog1. In addition, LEM assigned high contributions to observed effects almost solely to Hog1, which is in accordance with the known roles of Pbs2 as the upstream kinase activator of Hog1, and Hog1 as a regulator of downstream transcription factors as well as a promoter-binding and gene-regulating protein.

## 2 Linear effects models

The LEM is defined for a set of genes $G = \{1, \ldots, n\}$, a set of $m$ perturbation experiments $E$ targeting the genes, and their effect measurements $Y$. The genes constitute nodes of a deterministic pathway

graph $\mathcal{G} = (G, W)$ and are connected with a set of directed edges $W$. Edges in $W$ represent propagation of the perturbations within the pathway, and we assume that $\mathcal{G}$ is transitively closed. For each edge $(i, j)$ we call gene $i$ a parent and of $j$, and $j$ a child of $i$. Two nodes $i$ and $j$, which are not connected via a direct edge are called cousins, including nodes with no incoming edges. Since $\mathcal{G}$ is transitively closed, each pair of nodes in $\mathcal{G}$ can only either be in a parent–child relation, or be cousins, which defines their family relation in $\mathcal{G}$. Each perturbation experiment $e$ targets one or more gene in the pathway, and will be represented as the set of targeted genes, for example {1, 2, 3}. The set of experiments $E$ and the pathway graph $\mathcal{G}$ together define a binary matrix $S$, referred to as perturbation states matrix. For given experiment $e$ and gene $g$, entry $S_{e,g} = 1$ if gene $g$ is (directly or via propagation in the pathway) perturbed when experiment $e$ is performed, and 0 otherwise. Denote as $\mathrm{Pa}(g)$ the set of parents nodes of gene $g$ in $\mathcal{G}$. Perturbations in the pathway propagate via logical disjunction, meaning that $S_{e,g} = 1$ if experiment $e$ directly targets $g$, or if for any of its parents $p \in \mathrm{Pa}(g)$ the perturbation status $S_{e,p}$ is 1:

$$S_{e,g} = \begin{cases} 1 & \text{if } e \text{ targets } g, \\ \vee\{S_{e,p} | p \in \mathrm{Pa}(g)\} & \text{otherwise.} \end{cases} \quad (1)$$

Here, we assume that a disjunction over an empty set is 0, i.e. perturbation state $S_{e,g}$ of a gene $g$ without parents can attain value 1 only when $e$ directly targets $g$. The data $Y$ quantify the magnitude of perturbation effects, with $Y_e$ recording the effect of experiment $e$. For example, the values of $Y$ may correspond to absolute values of log gene expression change. Thus, LEM is not concerned with the direction of the effects, i.e. whether it is gene repression or activation, but how large these effects are on absolute scale. For simplicity we now assume that $Y$ is a one-dimensional vector, for example corresponding to expression changes of a single gene. Extension to multidimensional vectors (for many genes) is straightforward and explained below. Finally, a vector of parameters $\beta = [\beta_1, \ldots, \beta_n]^T$, with each $\beta_i > 0$, represents the individual contributions of the genes to the observed perturbation effects. Formally, the model assumes that $Y$ is a random variable, normally distributed around a linear combination of the individual gene contributions, with weights set to their perturbation states (Fig. 1B)

$$Y_e = \sum_g S_{e,g}\beta_g + \epsilon_e = S_e^T\beta + \epsilon_e, \quad (2)$$

where $\epsilon_e$ stands for measurement error, $\epsilon \sim N(0, c^{-1}I)$, with $c$ denoting the precision parameter (inverse variance), and where $S_e^T$ denotes the $e$-th row of matrix $S$. Note that from the assumptions that $S$ is binary, that each experiment $e$ targets at least one gene, and that the contributions $\beta$ are positive, it follows that the means $S_e^T\beta$ of the normal distributions for $Y_e$ are also positive. This is with accordance with the fact that $Y$ records the effect magnitudes and not their direction. Equation (2) can be read as the linear regression equation with design matrix $S$ and coefficients $\beta$. With these assumptions, the log likelihood function for the LEM takes the form

$$\ln(p(Y|S, \beta, c)) = \sum_{e=1}^{m} \ln\left(\mathcal{N}(Y_e|S_e^T\beta, c^{-1})\right) \quad (3)$$

$$= \frac{m}{2}\ln(c) - \frac{m}{2}\ln(2\pi) - \frac{c}{2}\sum_{e=1}^{m}\left[Y_e - S_e^T\beta\right]^2. \quad (4)$$

As a direct implication from Equation (2) and the fact that $S$ is binary, we have

FACT 1. *Denote D(g) the descendants of gene g in graph $\mathcal{G}$, i.e. the set of nodes reachable along directed edges from g. In the LEM, the effect observed under perturbation targeting g, denoted $Y_{\{g\}}$, is given by*

$$Y_{\{g\}} = \beta_g + \sum_{h \in D(g)} \beta_h + \epsilon_{\{g\}}.$$

Thus, the model makes explicit how the total effect of perturbing a single node in the pathway graph distributes across its sub-graph into the individual contributions of the descendant genes.

## 2.1 Learning the pathway structure

In most applications, the goal is to infer the pathway structure $\mathcal{G}$ from observed data. To learn a LEM, the model space needs to be searched evaluating candidate pathway graphs in terms of model fit to the data. For LEM, we implemented exhaustive search for small structures (up to five nodes) and greedy hill climbing in model space for larger ones (Russell and Norvig, 2003). Examples of models with developed greedy model search include Bayesian Networks (Chickering, 2003) or NEMs (Fröhlich *et al.*, 2009b). Here, we relied on corresponding procedures developed for the NEMs due to similarity of the deterministic pathway graph in our and the NEM model. Exhaustive LEM search enumerates and evaluates all possible distinct models of the size equal to the number of genes. More specifically, it enumerates one model per equivalence class by considering only transitive models with collapsed cycles (see below). Greedy search of model space traverses from one model to another in small steps corresponding to adding edges, greedily choosing the next graph as the one with the largest evaluation score.

Recall that the given set of experiments $E$ and the candidate pathway graph $\mathcal{G}$ define the perturbation states matrix $S$. In both exhaustive and greedy LEM search procedures, each considered candidate graph $\mathcal{G}$ is evaluated in a Bayesian manner, using marginal likelihood for Bayesian linear regression (Bishop, 2006). To this end, we employ a flat prior on all possible graphs, and assume that the precision parameter $c$ is a constant, while the prior distribution of the $\beta$ parameters, denoted $p(\beta|b)$, is a zero mean isotropic Gaussian with precision $b$, $\beta \sim N(0, b^{-1}I)$. Thus, we have two hyper-parameters, $b$ and $c$. In a fully Bayesian setup, we would consider priors on the hyper-parameters, and to compute the marginal likelihood $p(Y|S)$, we would marginalize over both the hyper-parameters $b$ and $c$, and the parameters $\beta$. For the sake of the efficiency of computations, we use an (empirical Bayes) approximation instead, taking point estimates $\widehat{b}, \widehat{c}$ of the hyper-parameters, and computing the marginal likelihood function $p(Y|S, \widehat{b}, \widehat{c})$, which involves integrating over only the parameters $\beta$. The point estimates are obtained by maximizing the marginal likelihood. For given $b$ and $c$ values, the log marginal likelihood function takes the form

$$l(Y|S, b, c) = \ln\left(\int p(Y|S, \beta, c)p(\beta|b)d\beta\right), \quad (5)$$

where $p(Y|S, \beta, c)$ is the likelihood introduced in Equation (3) and $p(\beta|b)$ is the prior distribution of $\beta$. Using the evaluation of the marginal likelihood function by Bishop, 2006 (Chapter 3.5), we have

$$l(Y|S, b, c) = \frac{n}{2}\ln(b) + \frac{m}{2}\ln(c) - E(\mu) + \frac{1}{2}\ln(|V|) - \frac{m}{2}\ln(2\pi), \quad (6)$$

where $|V|$ denotes the determinant of $V$ and where $V$ and $\mu$ are given by

$$V^{-1} = bI + cS^T S \tag{7}$$

$$\mu = cVS^T Y, \tag{8}$$

and $E(\mu)$ is computed as

$$E(\mu) = \frac{c}{2}||Y - S\mu||^2 + \frac{b}{2}\mu^T\mu.$$

With the requirement of combinatorial perturbations for identifiability (see below), the number of experiments $m$ exceeds the number of genes $n$, and we follow an iterative procedure to identify the estimates of hyper-parameters which maximize the marginal likelihood (Bishop, 2006). We start with initial values of $b$ and $c$. In each iteration, we first fix those values to compute $V$ and $\mu$ using the above equations, and second we recompute $b$ and $c$ using

$$b = \frac{n}{\mu^T\mu} \tag{9}$$

$$c = \frac{m}{||Y - S\mu||^2}. \tag{10}$$

We iterate until convergence.

### 2.2 Parameter inference

The parameters $\beta$ which describe the individual contributions of the genes to the effects are estimated as the mean of the posterior distribution inferred using the above Bayesian procedure

$$\widehat{\beta} = \mu.$$

### 2.3 Application to multidimensional data *Y*

In the case when more than one effect is measured in the experiment (e.g. expression changes of many genes), we deal with multidimensional data $Y$, which can be represented as a set of $k$ random variables $Y^j$, $j \in 1, \ldots, k$. We assume the contributions of the pathway components to each effect are different. Formally, this means that for each effect $j$ there is a different contribution vector $\beta^j$, and precision parameters $b^j$ and $c^j$. Thus, to compute the log marginal likelihood for a given pathway structure $\mathcal{G}$ and multidimensional data $Y$, we first estimate the hyper-parameters $b^j$ and $c^j$ for each effect $j$, and compute the marginal likelihood $l(Y^j|S, b^j, c^j)$ as explained above, and next we sum over the marginal likelihoods

$$l(Y|S, b^1, \ldots, b^k, c^1, \ldots, c^k) = \sum_{j=1}^{k} l(Y^j|S, b^j, c^j).$$

This procedure results also in $k$ estimators of the contribution vectors $\widehat{\beta}^j$, $j \in 1, \ldots, k$.

### 2.4 Integration of network prior

In the above considerations we assumed a flat prior over network structures. If available, existing knowledge of plausible networks should be formalized as a prior $P(\mathcal{G})$ over all possible networks $\mathcal{G}$. If the prior knowledge specifies that the network contains a specific set of edges, a simple prior reflecting that knowledge could for example be flat over the set of all graphs that include these edges, and fixed to 0 for all networks which miss them. Alternatively, to penalize network graph complexity, the prior could be a function inversely proportional to the number of its edges. Such a prior can be incorporated into model search procedures, where the set of experiments $E$ is given and candidate pathway graphs $\mathcal{G}$ are searched for the best fit with the data $Y$. To include the model prior, we set the prior over perturbation states matrix $S$, derived from the set of experiments and the candidate graph, as equal to the prior on $\mathcal{G}$, i.e. $P(S) = P(\mathcal{G})$, and in the search procedure we maximize the log posterior

$$l(Y|S, b, c) + \ln(P(S)),$$

instead of the log marginal likelihood (Equation (6)).

### 2.5 Model identifiability

Models that are undistinguishable in terms of their likelihood based on the observations belong to the same equivalence class and are not identifiable from data. For example, several Bayesian Networks with different directions of edges may equally well describe the same joint probability distribution over a set of random variables. Likewise, an equivalence class of NEMs with equal likelihood is defined by the set of all directed graphs with the same transitive closure. With the log likelihood function for the LEM defined by Equation (3), two different models $A$ and $B$ will belong to the same equivalence class if $S^A\beta^A = S^B\beta^B$ and $c^A = c^B$, where the upper indexes $A$ and $B$ indicate the states matrix and the parameters for the two models, respectively.

We now list the general constraints for the LEM to be identifiable. First, we put the requirement that the model graph $\mathcal{G}$ is transitively closed. Due to the assumption of perturbation propagation, for a given set of experiments $E$ two different model graphs with the same transitive closure would result in the same perturbation states matrix $S$, and with the same values of the parameters $\beta$ and $c$ would obtain equal likelihoods (compare Equation (3)). In addition, the LEM does not allow negative nor zero $\beta$ values. In the most degenerate example, consider a LEM with a given graph $\mathcal{G}$ and a zero $\beta$ vector. In such a case, any other model with a different graph, the same parameter $c$ and the same zero parameter $\beta$ vector would have equal likelihood. Likewise, the positive and negative $\beta$ values can cancel out, resulting in zero total expected effects, giving room for equally likely but different model graph structures. Finally, note that as in the NEMs, since the model graphs are transitive, cycles are cliques. Indeed, for model graphs with cycles, perturbation of any of the genes in the cycle propagates to all genes in the cycle, so that they are always either perturbed or not all at once. For those genes, their exact corresponding $\beta$ values in the LEM are not identifiable. Assume a given LEM with pathway graph $\mathcal{G}$ and a set of genes $C \subset G$ in a cycle and parameters $\beta^*$, $c^*$. The entries of the states matrix $S^*$ of this graph for those rows which perturb any gene in $C$ have all values in columns $C$ filled with value 1. There are infinitely many possible models, which have the same model graph $\mathcal{G}$ and the same parameter $c^*$, but different parameter vectors $\beta$ and equal likelihood. The $\beta$ vector of any such model has the same values as $\beta^*$ in entries not corresponding to genes in $C$, and such entries for the genes in $C$ which sum up to $\beta_C = \sum_{g \in C} \beta^*_g$. Nevertheless, all these models will have the correct pathway graph structure, with $C$ in a clique. Therefore, the model structure of such LEMs is identifiable. For models with cycles, we collapse the set of nodes in the cycle into a single node. Since we evaluate models using marginal likelihood where the $\beta$ parameters are integrated out, all models, including those with reduced number of nodes due to collapsing, can effectively be compared. If needed, each collapsed node can be expanded into its corresponding clique $C$, and the $\beta$ values for the genes in the clique can be estimated as $\frac{\beta_C}{|C|}$, where $|C|$ is the size of set $C$.

We will now derive that the set of perturbations of both all single nodes and all pairs of nodes in $S$ (together $\binom{n+1}{2}$ experiments for $n$

genes) is required to identify the LEM from the data $Y$. The following is proven in the Supplementary Material and illustrated in Figure 2.

LEMMA 1: *Each LEM graph $\mathcal{G} = (G, W)$ is uniquely defined by stating the family relation for each pair of its nodes $i, j \in G$.*

LEMMA 2: *Let $Y$ be a vector of effects for perturbations of single nodes in a given set $G$. There exist LEMs with model graphs on $G$ which are not identifiable from $Y$.*

LEMMA 3: *Let $Y$ be a vector of effects for perturbations of double nodes in a given set $G$. There exist LEMs with model graphs on $G$ which are not identifiable from $Y$.*

THEOREM 1: *Let $Y$ be a vector of effects for perturbations of all single nodes and all pairs of nodes in a given set $G$. All LEMs with model graphs on $G$ are identifiable from $Y$.*

## 3 Applicability to noisy data in various experimental setups

To assess the performance of our model, we applied it to simulated data with various levels of noise, small and large pathway sizes, violations to model assumptions and with different numbers of experimental repeats. First, we tested the ability of exhaustive LEM search to correctly identify all possible pathway structures $\mathcal{G}$ with three nodes (with the assumptions of transitive closure and excluding the model collapsing into a single cycle, there are 28 of them) and to correctly estimate their simulated contributions $\beta$. We simulated worst case, one-dimensional data vector $Y$, for all possible perturbations of single and pairs of nodes, and with five different experimental setups, where the number of times each experiment was repeated was equal $1, 2, 3, 4$ or $5$. Each setup was simulated with five different levels of noise ($\sigma = \sqrt{c^{-1}} \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$, where $\sigma$ denotes standard deviation of error terms in Equation (2)). For each simulated pathway graph, noise level and experimental repeat number, the model parameters $\beta$ were obtained as absolute values sampled 30 times from the standard normal distribution. Thus, the noise ranged from one hundredth to one half of the standard deviation of the beta values.

Figure 3 indicates that only extreme levels of noise ($\sigma \in \{0.25, 0.5\}$) for few experimental repeats are an issue for parameter estimation in LEM when such small networks are considered. Otherwise, the median correlation between estimated and true $\beta$ values is close to 1 (Fig. 3A). Similarly, root mean squared error (RMSE; Fig. 3B) of the estimations is only increased for large noise and few experimental repeats.

We refer to pathway structure as perfectly reconstructed when all the inferred edges of the pathway graph are exactly the same as the edges of the original used for simulation (Fig. 4A). Perfect reconstruction rate of small networks is sensitive to noise values, with the fraction of perfectly learned structures for $\sigma = 0.01$ equal almost 1, and for $\sigma = 0.5$ around 0.5. The fraction of perfectly learned structures increases with the number of experimental repeats. Faults in reconstruction may result from cases where the simulated $\beta$ are very small and due to noise cannot effectively be distinguished from 0, raising practical identifiability issues. These cases can be resolved by increasing the power of the estimation by adding more experimental repeats. Interestingly, sensitivity (fraction of true edges that are correctly identified as such) of exhaustive search is almost perfect already for one experimental repeat; with just a few exceptions observed for extreme noise levels (Fig. 4B). Specificity (fraction of missing edges that are correctly identified as such) is more affected by noise and repeat number, but still, its median is close to one in all cases but for a single repeat and the highest noise considered ($\sigma = 0.5$; Fig. 4C).

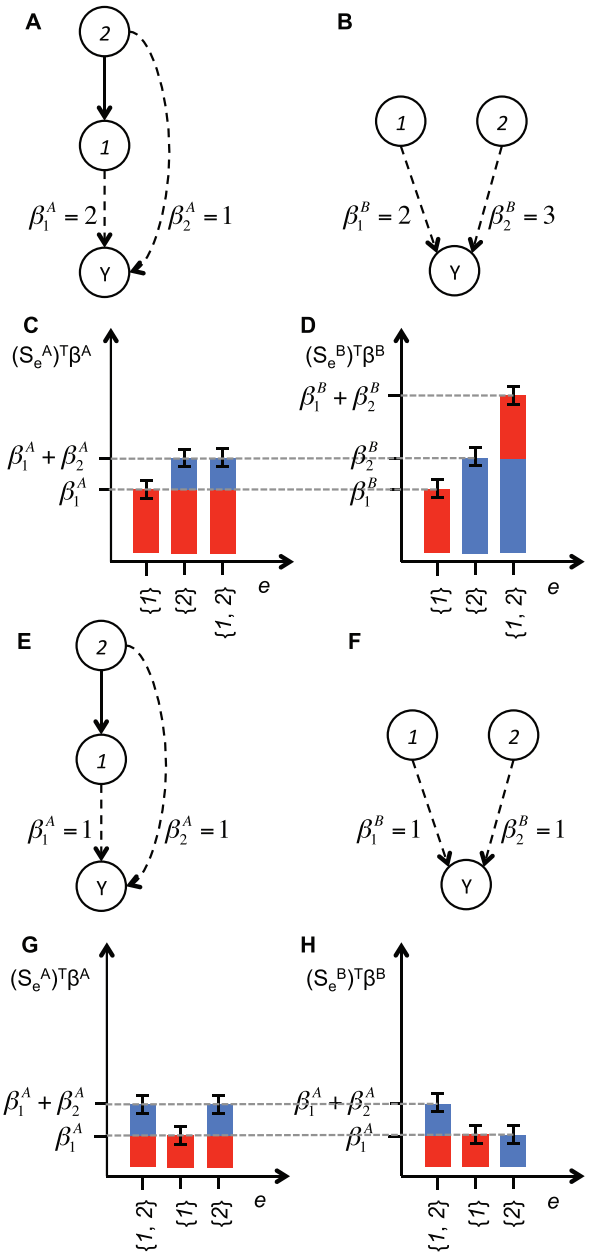


**Fig. 2.** Identifiability of the LEMs with only single and with only double-node perturbations. (**A**, **B**) Two distinct LEMs are considered: model *A* with parent gene *2* and child gene *1*, and contributions $\beta_1^A = 2$, $\beta_2^A = 1$ to the measured effects $Y$ (**A**), and model *B* with cousins *1* and *2*, and contributions $\beta_1^B = 2$, $\beta_2^B = 3$ (**B**). (**C**, **D**) For single node perturbation experiments *e*, models *A* and *B* assume equal means $(S_e^A)^T\beta^A = (S_e^B)^T\beta^B$ for the distributions of *Y*, and equal parameter *c* values would result in the same likelihood for both models. The model-predicted means would only be different if a double perturbation $\{1, 2\}$ was performed. (**E**, **F**) Two distinct LEM models: Model *A* with graph as in **A** and parameters $\beta_1^A = 1$, $\beta_2^A = 1$ (**E**) and model *B* as in **B** with parameters $\beta_1^B = 1$, $\beta_2^B = 1$ (**F**). (**G**, **H**) For the double perturbation $e = \{1, 2\}$ and with equal parameter *c*, the likelihood of these models given the data $Y_e$ is the same, since $(S_e^A)^T\beta^A = (S_e^B)^T\beta^B$. The likelihoods are only different for a single perturbation $\{2\}$.

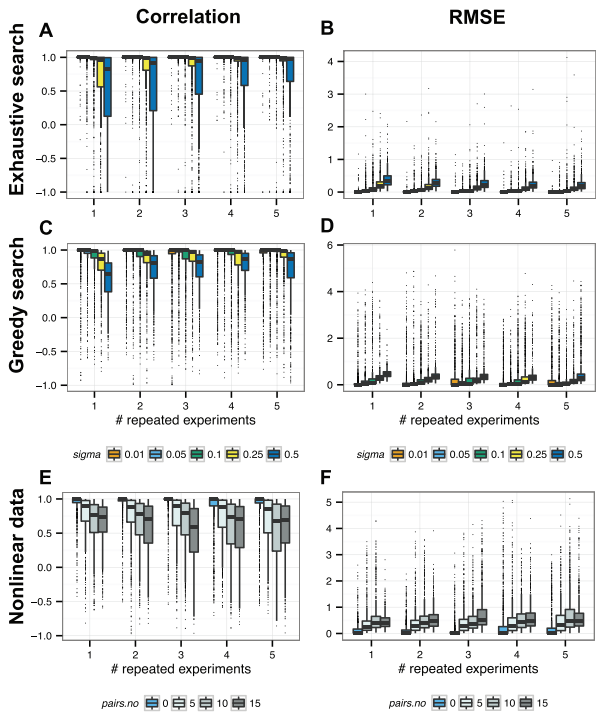


**Fig. 3.** LEMs allow accurate parameter estimation. (**A**, **C**) Distribution of the correlation between the true $\beta$ values used to simulate the data and the estimated values (*y*-axis) for increasing number of experimental repeats (*x*-axis) and for increasing noise (colours), for exhaustive search over 3-node networks (**A**) and greedy search over 10-node networks (**C**). (**B**, **D**) Distribution of RMSE in the same setup as in **A** and **C**. The performance of parameter learning is affected by extreme noise values only for few experimental repeats. (**E**) Distribution of the correlation between the true $\beta$ values and the estimated values (*y*-axis) for increasing number of experimental repeats (*x*-axis) and for increasing number of pairwise interaction terms (colours), which were used in linear combination with the true $\beta$s to generate the data (as violation to the model assumptions). The noise level is set to 0.05. (**F**) Distribution of RMSE in the same setup as in **E**.

To show that LEM can successfully be applied also to larger networks, we repeated the above described simulations, assuming the same noise levels and numbers of experimental repeats, each time generating 30 random networks of size 10, and for each sampling 30 absolute standard normal $\beta$ parameters. This time, to infer the network structures back from the simulated data, we applied a simple greedy hill climbing heuristic. This algorithm starts with an initial graph structure consisting of unconnected nodes, and iteratively adds one edge a time, choosing the additional edge that increases the marginal likelihood of the model the most. Compared with exhaustive search, the correlation of inferred to true $\beta$ parameters decreased, but to most extent for large noise values, where the median correlation dropped to 0.65 for the worst case of one experimental repeat (Fig. 3C). For noise levels $\leq 0.1$, median correlation remained around 1 for all numbers of experimental repeats. Median RMSE stayed below 0.5 for all tested cases (Fig. 3D). Fraction of the 10-node networks learned perfectly using the greedy hill climbing heuristic drops by around 40% compared with exhaustive search for small networks (Fig. 4D versus Fig. 4A). The sensitivity of greedy search remains, however, very high, with median close to one in all experimental settings apart from the large noise case ($\sigma = 0.5$) where the median drops to around 0.8 (Fig. 4E). Median specificities are lower compared with exhaustive search, but remain above 0.85 for all experimental setups, including large noise and low experimental repeat numbers (Fig. 4F).

Finally, we again repeated the 10-node network simulations, but setting one noise level ($\sigma = 0.05$) and adding effects violating the linear model assumptions. To this end, we generated the data as a linear combination of all pathway gene contributions and products of contributions of a selected number (0, 5, 10 and 15) of pairs of these genes, and applied greedy network search and LEM inference. Regardless of the number of experimental repeats, nonlinearity in the data substantially decreases the correlation of inferred to true $\beta$ values (Fig. 3E), and the fraction of perfectly learned networks. For the chosen noise level $\sigma = 0.05$, the median correlation dropped from around 1 to between 0.6 and 0.9, while the fraction of perfectly learned networks decreased from 0.5 to around 0.1. This deviation from model assumption affects also sensitivity (Fig. 4H) and specificity (Fig. 4I) to a larger extent than the increase of noise levels (compare Fig. 4E and F). Still, even with interaction terms contributed by 15 pairs of nodes (here, 30% of all possible pairs), median sensitivity around 0.75 and specificity around 0.9 is reached.

In summary, tests on simulated data demonstrated high accuracy as well as robustness to noise and violations to model assumptions of the LEM model search procedures.

## 4 Application to the MAPK HOG pathway

We next utilized our approach to infer the direction of the signal flow in the MAPK HOG pathway in yeast. LEM was applied to compare the marginal likelihood for the two alternative models, one where Pbs2 activates Hog1 and another where Hog1 activates Pbs2. Here, we utilized wild type and single mutants Δ*pbs*2 and Δ*hog*1 in 0.5M KCl-mediated osmotic stress conditions, for which expression of 2684 genes was monitored 40 min after treatment (O'Rourke and Herskowitz, 2004). This data allowed us to resolve the direction of the relation between the two kinases using LEM. To compare to the third possible model, where Pbs2 and Hog1 are independent, we would need an additional experiment where both are perturbed at the same time and gene expression changes are recorded, which was not included in the collection (O'Rourke and Herskowitz, 2004).

It has been long established that Δ*pbs2* deletion has phenotypically the same effects as Δ*hog1*. Exposing either of the two mutants to elevated osmolarity causes shift in expression of ribosomal genes, following the inability to grow and failure to proliferate as well as acquisition of unusual morphology, resembling mating projections or pseudohyphae (Hohmann, 2002). The latter phenotype is due to inappropriate activation of the pheromone response pathway and the pseudohyphal development pathway in Δ*pbs2* and Δ*hog1* mutants. Similarly to the phenotype, gene expression effects of the mutations in osmotic stress, when inspected by eye, also do not clearly indicate the direction of the signal flow between the two kinases (Fig. 5A). Effect magnitudes for most genes are high for both mutants, a lower fraction of genes shows large effects only for Δ*pbs2*, a similarly low fraction only for Δ*hog1*, and a final group of genes shows relatively subtle effects for both mutants. Thus, the fact that Pbs2p is upstream of Hog1p was historically derived not from deletion screens, but by the direct experimental observation that an osmotic upshift caused phosphorylation of Hog1 in a Pbs2p-dependent manner.

From all 2684 genes measured in the experiments, we selected 698 reporter genes which showed a 2-fold expression change between either mutant and wild type in osmotic stress. To this end, we selected those genes for which the ratio of expression in either mutant with 0.5 M KCl added and measured after 40 min was either less than 0.5 or larger than 2 when compared with WT with 0.5 M KCl after 40 min. We next transformed the ratios to effect

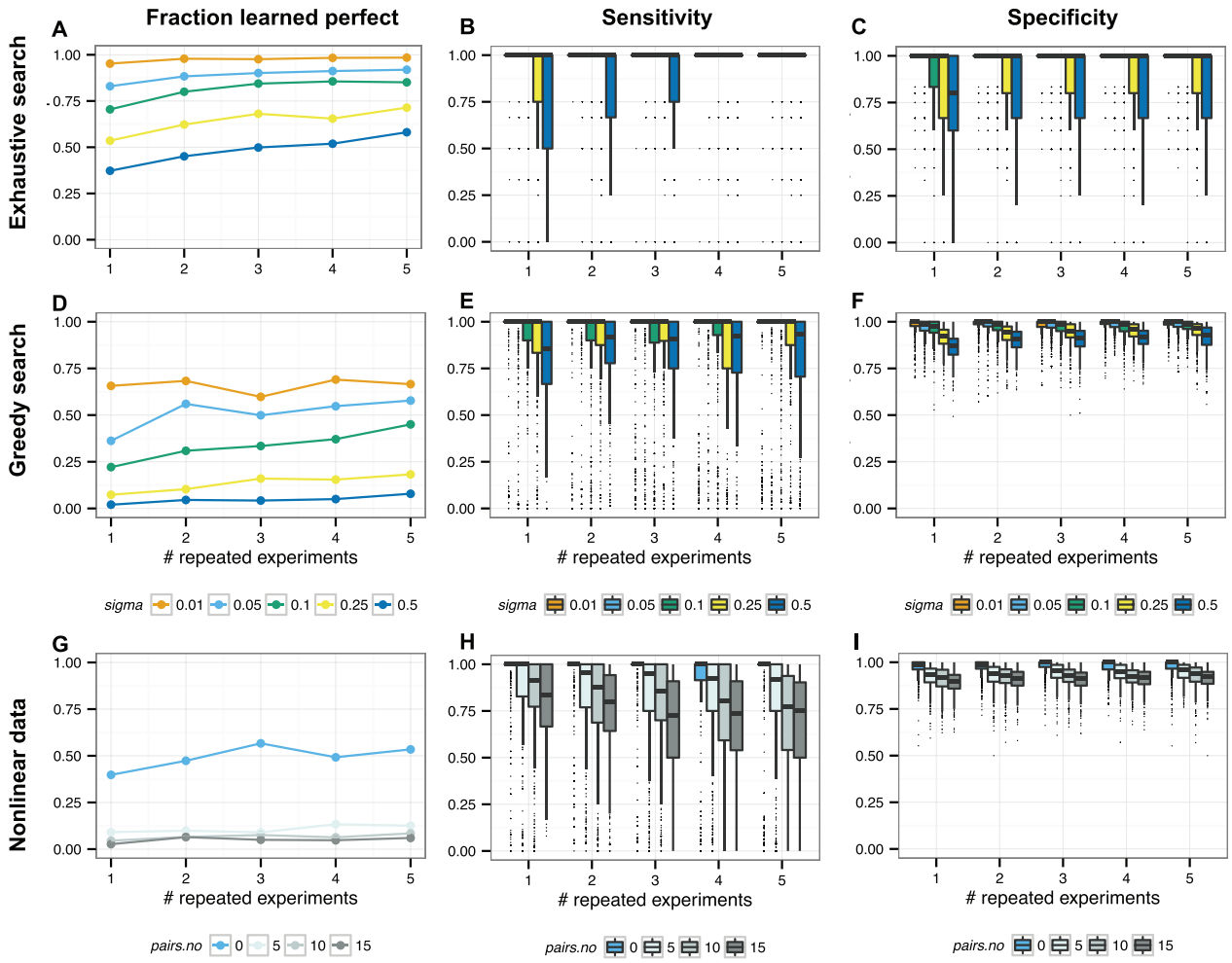


**Fig. 4.** LEMs allow accurate structure learning. Fraction of perfectly learned pathways (*y*-axis **A**, **D**), as well as sensitivity (*y*-axis **B**, **E**) and specificity (*y*-axis **C**, **F**) of edge learning, decrease with growing noise levels (indicated by colours), and increase with the number of times the experiments are repeated (*x*-axis), both for the exhaustive search over 3-node networks (**A–C**) and greedy search over 10-node networks (**D–F**). The addition of an increasing number of non-linear effects (colours **G**, **H**, **I**) decreases the fraction of perfectly learned pathways (*y*-axis **G**), as well as sensitivity (*y*-axis **H**) and specificity (*y*-axis **I**) of edge learning, regardless of the number of times the experiments are repeated (*x*-axis; here plotted for greedy network search). The performance of structure learning is most affected by high noise values and addition of non-linear effects.

magnitudes by taking absolute natural logarithm values (Fig. 5A). In application to this data, LEM correctly identified the true model structure (Fig. 5B). The distribution of the per-gene log marginal likelihoods (Fig. 5C) is clearly three-modal, with a pronounced low-end mode for genes with small log marginal likelihood for the model, and high-end mode of genes with high log marginal likelihood. We focused on the genes in the high-end mode (with log marginal likelihood $> 20$). Not surprisingly, these effect genes are enriched for functionality associated with the above-mentioned mutant phenotypes, including pheromone response, ribosomal protein functions (translation), as well as stress response (Fig. 5B). After removal of several ORFs now annotated as dubious, we further subdivided those genes into three groups. One group contained 126 genes with large estimated contributions ($\beta_{Hog1} > \ln(2)$) for Hog1 and lower ($\beta_{Pbs2} < \ln(2)$) for Pbs2. Second contained five genes with large estimated contributions for Pbs2 and lower for Hog1. The third was composed of two genes with high contributions estimated for both pathway components.

In summary, the LEM correctly pointed at the true model structure, and provided a clear interpretation of perturbation data, associating almost all of the control of the downstream effects to Hog1, implying that the effects seen in Δ*pbs2* mutant are almost solely due to propagation down to Hog1.

## 5 Conclusion

To our knowledge, LEM is the first approach which uses observed effects of perturbations of hidden variables to tackle two tasks at the same time. First, it resolves the interconnections between the perturbed pathway components (layer 1 in Fig. 1A), and second, it derives their individual contributions to the observed effects (layer 2). Results on simulated data clearly show that LEMs are indeed capable of accurately solving both tasks. One group of existing approaches, with the most prominent class of models based on nested effects assumption (NEMs and extensions) (Markowetz *et al.*, 2005; Markowetz *et al.*, 2007; Tresch and Markowetz, 2008; Fröhlich *et al.*, 2008, 2009b, 2011; Anchang *et al.*, 2009; Siebourg-Polster *et al.*, 2015), is mostly concerned with the first task. The output brought by LEMs from perturbation data is richer than what can be inferred with the NEM-based methods, but also requires more input.

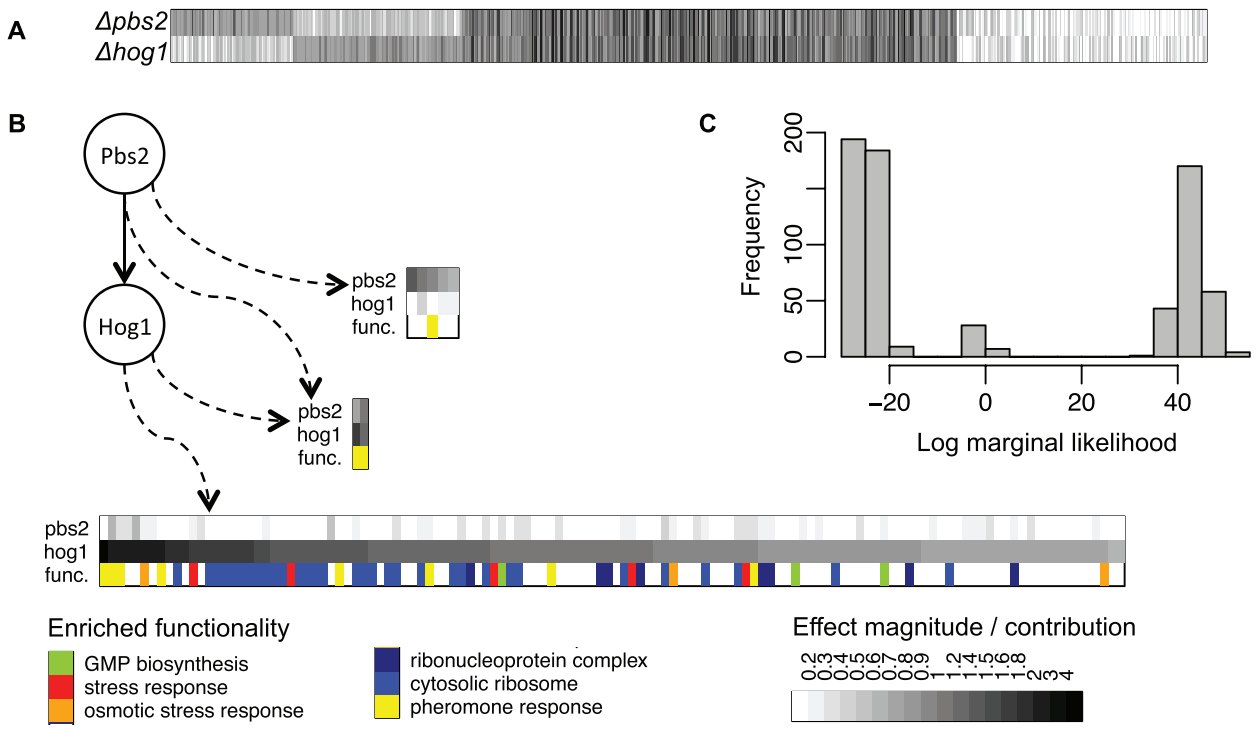


**Fig. 5.** Application to the MAPK HOG pathway. (**A**) The magnitudes of observed effects for 698 responsive genes. Rows correspond to experiments, columns to genes and entries to absolute log expression changes (shades of gray legend at the bottom of the figure). (**B**) The inferred pathway structure and estimated individual contributions to effects. For the purpose of this illustration, we divided the set of presented effect genes into three subsets and with dashed arrows indicated which of the pathway genes had a large estimated contribution to the observed expression changes of those genes. For the genes in each set, we represent their estimated contributions of Pbs2 and Hog1 as coloured matrices, with columns standing for the genes and the first two rows for Pbs2 and Hog1, respectively. The third rows are coloured according to the functional annotation of the genes. (**C**) The distribution of the log marginal likelihoods for the per-gene effects.

Both the NEM methods and LEM represent layer 1 in the same way, but unlike NEM, LEM estimates layer 2 as continuous, positive-value contributions of all pathway components to the effects. For unambiguous model inference, LEM requires not only perturbations of all single pathway genes, but also of all pairs of genes. Other previous studies concentrated solely on the second task of learning the links from the pathway to the effects in layer 2 (Gat-Viks and Shamir, 2007; Szczurek *et al.*, 2009). LEM outperforms these approaches with the ability to infer the pathway graph in layer 1, with the price that the description of layer 2 by LEM is by far less involved.

LEMs assume the measured effects are linear combinations of individual, per-pathway gene contributions to these effects. With this assumption, LEM is not suitable for modeling epistatic effects. Therefore, LEM is not aimed at studying redundancy between pathways, and in general nonlinear effects of combinatorial perturbations. Instead, it is tailored to resolve the structure of one pathway at a time, where the perturbations are assumed to propagate from parent regulator nodes to the child nodes whenever any of the parents is perturbed. Moreover, due to identifiability constraints, LEMs cannot take into account positive and negative contributions, as activation and repression, and can consider only the absolute magnitude of the measured effects. The assumption that the gene contributions are only positive is not obviously biologically sound; for example perturbation of one gene may mask effects of perturbation in another. This assumption, however, apart from model identifiability, assures also that the definition of the gene contributions is compatible with the definition of the network graph. Propagation of perturbation effects in the network graph is only 'positive' in the sense that there is no masking or down-regulation of perturbations.

In summary, not accounting for higher-order interactions among gene contributions or for their sign pose limits on the range of model applicability. Still, already with the application to the small example of the signal flow in MAPK HOG pathway, we see that the power of LEM lays in resolving the structure together with distributing the control over effects to the proper players within the pathway.

## Funding

This work was partially supported by the Warsaw Center of Mathematics and Computer Science.

*Conflict of Interest*: none declared.

## References

Anchang,B. *et al.* (2009) Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc. Natl. Acad. Sci. USA*, **106**, 6447–6452.

Bender,C. *et al.* (2010) Dynamic deterministic effects propagation networks: Learning signalling pathways from longitudinal protein array data. *Bioinformatics*, **26**, i596–i602.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Chickering,D.M. (2003) Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, **3**, 507–554.

Ellis,B. and Wong,W.H. (2008) Learning causal Bayesian network structures from experimental data. *J. Am. Stat. Assoc.*, **103**, 778–789.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Fröhlich,H. *et al*. (2008) Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data. *Bioinformatics*, **24**, 2650–2656.

Fröhlich,H. *et al*. (2011) Fast and efficient dynamic nested effects models. *Bioinformatics*, **27**, 238–244.

Fröhlich,H. *et al*. (2009a) Deterministic Effects Propagation Networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics*, **10**, 322.

Fröhlich,H. *et al*. (2009b) Nested effects models for learning signaling networks from perturbation data. *Biom. J*., **51**, 304–323.

Gat-Viks,I. and Shamir,R. (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res*., **17**, 358–367.

Gat-Viks,I. *et al*. (2006) A probabilistic methodology for integrating knowledge and experiments. *J. Comp. Biol*., **13**, 165–181.

Hohmann,S. (2002) Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol. Mol. Biol. Rev*., **66**, 300–372.

Markowetz,F. (2010) How to understand the cell by breaking it: Network analysis of gene perturbation screens. *PLoS Comput. Biol*., **6**, e1000655–e1000.

Markowetz,F. *et al*. (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, **21**, 4026–4032.

Markowetz,F. *et al*. (2007) Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, **23**, i305–i312.

O'Rourke,S.M. and Herskowitz,I. (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol. Biol. Cell*, **15**, 532–542.

Pe'er,D. *et al*. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17(Suppl. 1)**, S215–S224.

Rogers,S. and Girolami,M. (2005) A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, **21**, 3131–3137.

Russell,S.J. and Norvig,P. (2003) *Artificial Intelligence: A Modern Approach*. 2nd edn. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Sachs,K. *et al*. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Siebourg-Polster,J. *et al*. (2015) NEMix: Single-cell nested effects models for probabilistic pathway stimulation. *PLoS Comput. Biol*., **11**, e1004078.

Szczurek,E. *et al*. (2009) Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments. *Mol. Syst. Biol*., **5**, 287.

Tresch,A. and Markowetz,F. (2008) Structure learning in nested effects models. *Stat. Appl. Genet. Mol. Biol*., **7**, Article9.