Bioinformatics, 35, 2019, i605–i614
doi: 10.1093/bioinformatics/btz334
ISMB/ECCB 2019

Learning signaling networks from combinatorial perturbations by exploiting siRNA off-target effects

Jerzy Tiuryn and Ewa Szczurek*

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland

*To whom correspondence should be addressed.

Abstract

Motivation: Perturbation experiments constitute the central means to study cellular networks. Several confounding factors complicate computational modeling of signaling networks from this data. First, the technique of RNA interference (RNAi), designed and commonly used to knock-down specific genes, suffers from off-target effects. As a result, each experiment is a combinatorial perturbation of multiple genes. Second, the perturbations propagate along unknown connections in the signaling network. Once the signal is blocked by perturbation, proteins downstream of the targeted proteins also become inactivated. Finally, all perturbed network members, either directly targeted by the experiment, or by propagation in the network, contribute to the observed effect, either in a positive or negative manner. One of the key questions of computational inference of signaling networks from such data are, how many and what combinations of perturbations are required to uniquely and accurately infer the model?

Results: Here, we introduce an enhanced version of linear effects models (LEMs), which extends the original by accounting for both negative and positive contributions of the perturbed network proteins to the observed phenotype. We prove that the enhanced LEMs are identified from data measured under perturbations of all single, pairs and triplets of network proteins. For small networks of up to five nodes, only perturbations of single and pairs of proteins are required for identifiability. Extensive simulations demonstrate that enhanced LEMs achieve excellent accuracy of parameter estimation and network structure learning, outperforming the previous version on realistic data. LEMs applied to *Bartonella henselae* infection RNAi screening data identified known interactions between eight nodes of the infection network, confirming high specificity of our model and suggested one new interaction.

Availability and implementation: https://github.com/EwaSzczurek/LEM

Contact: szczurek@mimuw.edu.pl

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

1 Introduction

Signaling networks consist of interconnected proteins that transmit information inside the cell. External or internal stimuli trigger signaling cascades, which are implemented by consecutive kinases post-transcriptionally modifying one another, for example by phosphorylation. The signal is in this way transmitted down to the nucleus, where the transcriptional machinery regulates adequate functional response of the cell to the signal. Technological advances such as gene silencing using RNA interference (RNAi) (Agrawal *et al.*, 2003), gene knock-out using CRISPR-Cas (Hsu *et al.*, 2014) or perturbations by the means of drug treatment (Molinelli *et al.*, 2013) allow experimental interventions on cellular networks and measurement of their effects. For example, a small interfering RNA (siRNA) perturbation screen in human cells was performed in a study of pathogen infection (Ramo *et al.*, 2014). The aim of the study was to identify genes of the host that are involved in the networks that become activated when the cells are infected by pathogens. The measured phenotype was the level of infectivity by the pathogen upon siRNA-mediated knock-down. The ultimate goal in understanding cellular networks, however, is not only to recognize the genes involved in the network, but also to (i) resolve its structure of interconnections (network edges), and to (ii) understand the way the network members contribute to the observed phenotype, for

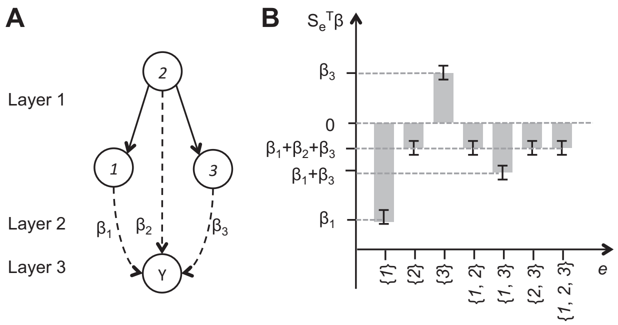

Fig. 1. Enhanced linear effects model. (**A**) Three layers of the system: (1) perturbed signaling network, (2) intermediate regulatory layer, (3) observed effects *Y*. Genes (circles, here 1, 2, 3) are directly or indirectly (via propagation in the network) perturbed in experiments. Bold arrows indicate how perturbations propagate within the network. 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 β_1, β_2, β_3), with weights set to perturbation states. The difference between the enhanced LEM and the previous model is that the contributions do not have to be positive, and can take any real value but not zero. (**B**) Example means (*y*-axis) for all possible perturbation experiments (*x*-axis), as expected in the enhanced LEM with network structure as in (A), for $\beta_1 = -4$, $\beta_2 = 1$ and $\beta_3 = 2$. Whiskers indicate example error

example how the individual kinases regulate the infectivity. Such perturbation studies have the potential to address both these questions. Despite, however, the perturbation technology becoming ever more advanced and increasingly better understood (Lisitskaya *et al.*, 2018; Mohr and Perrimon, 2012; Terns, 2018), computational inference of signaling networks from perturbation data remains a challenge.

The first problem faced by signaling network modeling is the so-called perturbation-effect gap (Markowetz, 2010). Although the interventions target the network (layer 1 in Fig. 1A), the resulting states of the network nodes are not measured. The observed variables reside in the layer of the measured phenotype downstream (layer 3 in Fig. 1A), only indirectly connected with the network via the middle regulatory layer (layer 2 in Fig. 1A). Second, the network nodes are not only targeted directly by the experiments, but also via propagation in the network. Once a given network node is inactivated by the experiment, the signal is blocked and the nodes downstream also become inactivated. Thus, perturbations propagate in the network along its unknown edges, in the same way as the signal. The third problem is the complexity of the hidden and unknown regulatory layer. The measured phenotype is rarely an effect of a single member of the network. Instead, it is a combination of contributions from all perturbed (directly or via propagation in the network) nodes. These contributions may be positive and enlarge the measured phenotype, or negative and decrease it. Fourth, interventions using common technique of siRNA-mediated knock-down, although intended only to target single genes, the so-called on-targets, at the same time perturb multiple other genes, the so-called off-targets (Jackson *et al.*, 2003). A given siRNA, it affects its off-target genes using the microRNA pathway, binding by complementarity of its seed region (positions 2–8) to the 3′ untranslated regions of the transcripts of the genes (Sigoillot and King, 2011). Consequently, the measured phenotype is confounded by the combinatorial effects from the off-targets and cannot be interpreted as the result of perturbing the on-target alone. Finally, there is the problem of model identifiability from perturbation data. The key question here is, how many and what combinations of perturbations are required to uniquely and accurately infer the model?

Nested effects models (NEMs) (Fröhlich *et al.*, 2008, 2009; Markowetz *et al.*, 2005, 2007; Tresch and Markowetz, 2008) and their extensions (Anchang *et al.*, 2009; Fröhlich *et al.*, 2011; Pirkl *et al.*, 2016; Siebourg-Polster *et al.*, 2015; Srivatsa *et al.*, 2018) specifically address the perturbation-effect gap problem. NEMs represent the network structure by directed graphs. The crucial assumption behind NEMs is that perturbation effects show a nested subset hierarchy, which reflects the hierarchy of nodes in the signaling network. As a graphical model of signaling networks, probabilistically inferred from the observed effects, NEMs constitute an attractive approach for solving problem of learning network structures (layer 1 in Fig. 1A). These models, however, have a simplified representation of the regulatory layer (2 in Fig. 1A), and assume that each effect is regulated only by a single gene in the network. Other previous computational approaches to signaling networks concentrated solely on solving problem of elucidating the link between the network and the observed effects (layer 2 in Fig. 1A) (Gat-Viks and Shamir, 2007; Szczurek *et al.*, 2009, 2011). These approaches assume they are given a known network graph, and aim at either small refinements to the given graph, or resolving the detailed mechanisms governing the regulation of the downstream targets by the network components.

Recently, we have introduced linear effects models (LEMs), aiming to address the above-mentioned problems (Szczurek and Beerenwinkel, 2016). LEM is a model where layer 1 of the signaling network is represented by a graph with nodes corresponding to the signaling genes. Edges of that graph correspond to the way the perturbation effects propagate in the network. For two nodes 1 and 2, an edge (1, 2) indicates that perturbation of gene 1 affects also gene 2. The downstream regulation layer 2 is modeled by a vector of model parameters, with entries corresponding to individual contributions to the observed effects. LEMs are inferred from the data, which is the phenotype (layer 3) measured under the perturbation experiments. With this formulation, LEMs aim both at learning the structure of interactions within the network and at deconvolution of the contributions of its components to the observed perturbation effects. The main drawback of the previously introduced model, however, is the limited expressiveness of the allowed contribution values. The model parameters are assumed to be strictly positive. As such, they are interpreted as only the magnitudes (absolute values) of the contributions to the measured effects. Thus, the original LEM cannot model both positive and negative contributions, such as down- or up-regulation of the measured phenotype.

Here, we extend the previously proposed LEM by allowing negative contributions for the network genes, which are now assumed to take any real, nonzero values. This allows realistic modeling of the phenotype as a combination of up-and down-regulatory contributions from the perturbed genes. At the same time, this enhancement of the model raises nontrivial question of model identifiability. For the original LEMs we proved that a set of experiments perturbing all single and all pairs of nodes is required for their identifiability (Szczurek and Beerenwinkel, 2016). The main contribution of this work is a proof that enhanced LEMs are identifiable from data measured under experiments where single, double and triple genes are perturbed. Moreover, we show that small enhanced LEMs, with less than six nodes, are identifiable from the same set of experiments, targeting only single and pairs of nodes, as the original LEMs. In this manuscript, two inference approaches for the enhanced LEMs are compared, namely Bayesian linear regression, referred to as Bayesian approach, and its time-efficient approximation, referred to as the Bayesian Information Criterion (BIC) approach. We perform comprehensive simulations to demonstrate that both approaches yield excellent accuracy of parameter estimation and network structure recovery of the enhanced LEMs, and to track the run times of the two approaches. Although the Bayesian approach performs slightly better in model inference, its run times are much longer than of the BIC approach.

The curse of siRNA off-targets can be turned into a blessing with the use of computational tools of microRNA target predictions such as TargetScan (Lewis *et al.*, 2005). Using these tools, we can identify which genes are off-targeted by the siRNA, and treat both the known on-target and the predicted off-targets as a set of genes that are combinatorially perturbed within the same experiment. Recently, Srivatsa *et al.* (2018) demonstrated the applicability of the siRNA on- and off-targets as combinatorial perturbations, allowing to learn signaling networks using their pc-NEMs models. They did not, however, account for the fact that the set of network genes is only a subset of all genes that are perturbed by the siRNA, and that the perturbation of the remaining genes may also have their effect on the phenotype. Importantly, Schmich *et al.* (2015) showed that a phenotype measured under siRNA screens can successfully be modeled as a linear combination of contributions of the on- or off-targeted genes. Their approach was applied to model the infectivity phenotype in the siRNA screen of Ramo *et al.* (2014), demonstrating dramatically increased correlation of the inferred gene contributions compared to the confounded raw phenotypes between different siRNA libraries. Their model, however, did not account for any possible network connections between the perturbed genes, treating them as isolated nodes. Since LEM can be thought of an extension of this simple model, which accounts for the structure of the signaling network and the way perturbations propagate along its edges, we reasoned that LEMs can be particularly well suited to model signaling networks from combinatorial perturbations in siRNA screens. To this end, we correct the phenotype for the contributions of perturbed genes which are not part of the modeled network. Indeed, application of LEMs to inference of the *Bartonella henselae* (shortly B. *henselae*) infection network from infection kinome screen (Ramo *et al.*, 2014) demonstrates excellent recovery of known interactions. In summary, the present work introduces a more realistic and highly expressive model for learning signaling networks and their regulation of downstream phenotypes, and comes with proven identifiability constraints ensuring accurate model inference.

2 Enhanced linear effects models

A *linear effects model* (LEM) is defined by a triple $M = (\mathcal{G}, \beta, c)$, where $\mathcal{G}$ is a finite, transitively closed, directed acyclic graph (DAG) with n nodes, $\beta \in \{R \setminus 0\}^n$ is a vector of nonzero real values, henceforth called *admissible vectors*, and $c > 0$ is a real number called *precision parameter*. The graph $\mathcal{G} = (V, W)$ is defined by the set of nodes $V = \{1, \ldots, n\}$ and a directed, and transitively closed set of edges W. The assumption $\beta_g \neq 0$, for all $g \in \{1, \ldots n\}$, and $\mathcal{G}$ being a transitively closed DAG is motivated by model identifiability, as shown below. We write $a \to_{\mathcal{G}} b$ to indicate that there is an edge in $\mathcal{G}$ from vertex a to vertex b, and we call this edge *outgoing* from a and *incoming* to b. By a *root* in $\mathcal{G}$ we mean any node a with no incoming edges. By a *leaf* in $\mathcal{G}$ we mean any node with no outgoing edges. The graph $\mathcal{G}$ corresponds to a signaling network, with nodes interpreted as genes and the edges representing the way perturbations of nodes propagate within the network. For a node a of $\mathcal{G}$ we let $V_a^{\mathcal{G}} = \{a\} \cup \{b \in V \,|\, a \to_{\mathcal{G}} b\}$ to be the set of all nodes with edges in $\mathcal{G}$ outgoing from a, plus node a. Moreover for $X \subseteq V$, we let $V_X^{\mathcal{G}} = \cup_{a \in X} V_a^{\mathcal{G}}$. The interpretation of the set $V_X^{\mathcal{G}}$ is that if X is a set of genes that are targeted directly by a given perturbation experiment, then $V_X^{\mathcal{G}}$ consists of all genes that are perturbed directly or via propagation in the network $\mathcal{G}$.

LEMs are inferred from data $Y \in R^m$ measured under perturbation experiments described by a $m \times n$ binary *perturbation* matrix E. For an experiment e and a gene g, entry $E_{e,g} = 1$ indicates that e directly perturbs gene g, and $E_{e,g} = 0$ otherwise. The perturbation matrix E, specifying the experiments, and the network graph $\mathcal{G}$, together determine a binary matrix $S(E, \mathcal{G})$, called a *design matrix*. We drop the arguments when they are clear from the context. For a given experiment e and gene g, entry $S_{e,g} = 1$ if gene g is perturbed directly or via propagation along the network, and $S_{e,g} = 0$ otherwise. Thus, for $e \in \{1, \ldots, m\}$, if X_e denotes the set of genes that are targeted as a result of performing the experiment e, then for every $g \in \{1, \ldots, n\}$

$$S_{e,g} = \begin{cases} 1 & \text{if } g \in V_{X_e}^{\mathcal{G}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The data Y quantify the perturbation effects, with Y_e recording the effect of experiment e. For example, Y could measure expression of a certain gene regulated by the network. In contrast to the original LEM formulation (Szczurek and Beerenwinkel, 2016), the values of Y are not restricted to the absolute magnitudes of effects, but correspond to the actual values of these effects, which can be positive or negative. Finally, an admissible vector of parameters $\beta = [\beta_1, \ldots, \beta_n]^T$, represents the individual contributions of the members of the network to the observed perturbation effects. In contrast to the original formulation of LEMs, the here proposed version of the model allows that the contributions need not be only positive, but also can take negative values. For the gene regulation example, $\beta_g > 0$ would indicate that the perturbation of a given node g in the network contributes to activation of the regulated gene. On the other hand, $\beta_g < 0$ would indicate that perturbation of g contributes to repression of the measured gene. Note, that this should be interpreted as that g, when not perturbed, individually has a positive contribution to regulation of the measured gene. Moreover, it is important to distinguish the individual contribution β_g from the total effect of perturbing gene g, which is given by the sum of its own contribution and the contributions from the nodes downstream of g in the network, i.e. $\sum_{a \in V_g^{\mathcal{G}}} \beta_a$. In particular, even when the individual contribution β_g is negative, the total effect of perturbing g can be positive. Remaining assumptions are the same as in the original LEM, namely, 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)

$$Y_e = \sum_g S_{e,g}\beta_g + \epsilon_e = S_{(e,-)}\beta + \epsilon_e, \qquad (2)$$

where ϵ_e stands for measurement error, $\epsilon \sim N(0, c^{-1}I)$, with c denoting the precision parameter (inverse variance), and where $S_{(e,-)}$ denotes the eth row of matrix S. Equation (2) can be read as the linear regression equation with design matrix S and coefficients β. With these assumptions, the log-likelihood function for the LEM $M = (\mathcal{G}, \beta, c)$ and the data Y with perturbation matrix E is given by (Szczurek and Beerenwinkel, 2016)

$$\mathcal{L}(M) = \ln\Big(p(Y|S, \beta, c)\Big) = \sum_{e=1}^{m} \ln\Big(f\big(Y_e|S_{(e,-)}\beta, c^{-1}\big)\Big), \qquad (3)$$

where $f(Y_e|S(_{e,-})\beta,\ c^{-1})$ is the gaussian probability density function with mean $S(_{e,-})\beta$ and variance c^{-1}.

For any two positive integers $k \leq n$, we consider a special perturbation matrix $E^{(k,n)}$ that includes all possible experiments that target directly at most k genes out of given n-element set of genes. Hence $E^{(k,n)}$ has $m = \sum_{i=1}^{k} \binom{n}{i}$ rows and n columns, and for each i-element subset X of genes, $E^{(k,n)}$ has a row with 1s in exactly positions that correspond to genes in X, and 0s in the remaining positions.

In reality it may happen that the modelled biological network contains a cycle. When we take a transitive closure of such a graph $\mathcal{G}$, the cycle turns into a clique. It follows that the design matrix S has identical columns that correspond to the genes belonging to this clique. Consequently, Theorem 3 stated below fails for graphs $\mathcal{G}$ that contain such cliques, causing a problem with identifiability of LEMs. Such genes belonging to the same clique behave in an identical way under all perturbation experiments. Therefore we collapse the whole clique into a single node. This procedure eventually leads to a DAG over a smaller set of nodes. The node that corresponds to the collapsed clique represents a set of genes of the clique, rather than a single gene.

2.1 Learning the network structure

We are given a $m \times n$ perturbation matrix E and a vector $Y \in R^m$ as the observed data. To learn a LEM $(\mathcal{G}, \beta, c)$ from observed data, we need to infer the graph $\mathcal{G}$, corresponding to the signaling network structure, and the vector of contributions β. To search the graph space we use the two procedures developed and described previously (Szczurek and Beerenwinkel, 2016), and either exhaustively evaluate all possible small graphs of up to five nodes, or greedily search for DAGs that maximize the evaluation score by iteratively adding or removing edges from currently considered graph. In the case when adding an edge results in a cycle, the cycle is collapsed into a single node and in the corresponding matrix S the columns for the nodes in the cycle are replaced by a single column. This latter step is sound since, by transitivity, all columns of S that correspond to nodes from the cycle are equal to each other. The greedy search is initialized from an empty graph with unconnected nodes, and an additional number of initializations from randomly sampled initial graphs can be set as a parameter by the user. For enhanced LEMs proposed here, we propose two alternative procedures for evaluating the candidate graphs, referred to as *Bayesian* and *Bayesian Information Criterion* approach, respectively.

The Bayesian graph evaluation procedure is the same as proposed for original LEMs. We score the graphs using marginal likelihood for Bayesian linear regression (Bishop, 2006). 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 β parameters, denoted $p(\beta|b)$, is a zero mean isotropic Gaussian with precision b, $\beta \sim N(0, b^{-1}I)$. Assuming an empirical Bayes approximation, we take point estimates $\hat{b}, \hat{c}$ of the hyper parameters, and for a candidate graph $\mathcal{G}$ as its evaluation score we compute the marginal likelihood function $p(Y|S, \hat{b}, \hat{c})$, which involves integrating over only the parameters β. The point estimates are obtained by maximizing the marginal likelihood in an iterative procedure described previously (Szczurek and Beerenwinkel, 2016). The marginal likelihood used in the Bayesian approach allows comparing models with different number of nodes.

For BIC evaluation, we use least squares to solve the linear regression problem [Equation (2)] for a given candidate graph $\mathcal{G}$. Equivalently, we estimate the parameters $\hat{\beta}$ and $\hat{c}$ that maximize the likelihood $p(Y|S, \beta, c)$ [Equation (3)]. Due to the fact that cycles are collapsed into single nodes, we need to assure that the evaluation score does not favor larger models, with a larger number of parameters. Each candidate graph is thus scored using the negative BIC,

$$2\ln\Big(p(Y|S, \hat{\beta}, \hat{c})\Big) - \log(m)k,$$

where m is the number of experiments and k is the number of nodes after collapsing. Notably, the BIC score is an approximation of the marginal likelihood used in the Bayesian approach (Bishop, 2006). Since its computation requires evaluation of the likelihood only once, it is more time efficient than the iterative procedure required to compute the marginal likelihood. Both scores are maximized in the search.

2.2 Parameter inference

For the Bayesian approach, we estimate the contributions β as the mean of their posterior distribution inferred using the Bayesian procedure described previously (Szczurek and Beerenwinkel, 2016). For the BIC approach, we use the least squares estimates.

In the case when more than one effect is measured in the experiment (e.g. expression changes of many genes), the evaluation procedure can easily be extended to deal with multidimensional data $Y = \{Y_1, \ldots, Y_D\}$ by assuming independence of the parameters for the D different phenotype vectors $Y_d \in Y$. Each Y_d is assumed to be generated from a shared network structure but with a different contribution vector, following the procedure of Szczurek and Beerenwinkel (2016).

2.3 Enhanced model identifiability

Before proving identifiability conditions for LEMs, we show several important properties for pairs of DAGs $(\mathcal{G}_1, \mathcal{G}_2)$ over the same set V of nodes. Given such a pair and a positive integer k, we consider a system $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$ of equations of the form

$$\sum_{x \in V_X^{\mathcal{G}_1}} v_x = \sum_{x \in V_X^{\mathcal{G}_2}} u_x,$$

where $X \subseteq V$ ranges over all subsets with at most k elements. Here variables v_x and u_x have indices x ranging over genes.

If the set V has n elements, then a solution of the system of equations $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$ is a pair of vectors $v, u \in R^n$ that satisfies these equations. A solution v, u is said to be *admissible* if for all $a \in V$, $v_a \neq 0$ and $u_a \neq 0$. Let $k \geq 1$. We say that two DAGs $\mathcal{G}_1, \mathcal{G}_2$ are *k-distinguishable* if the system $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$ has no admissible solutions.

The reason for considering systems of equations $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$ is the following. Consider two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ over the same n-element set of vertices and a perturbation matrix E that addresses at most k element sets of genes being directly targeted by perturbation experiments. In the case when two vectors $\beta, \gamma \in R^n$ form an admissible solution of $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$, we have that $S(E,\mathcal{G}_1)\beta = S(E,\mathcal{G}_2)\gamma$, and the LEMs $(\mathcal{G}_1, \beta, c)$ and $(\mathcal{G}_2, \gamma, c)$ would not be distinguished by the given data. This follows from the fact that LEMs assume that Y is normal distributed around the mean given by the product of the design matrix and the contribution vector, and in this case it would be the same for the two different models. Thus, for a shared parameter c, these two models would obtain identical likelihood (3).

For the sake of space, all proofs of the results presented here are moved to Supplementary Material. We have the following immediate observation which is used throughout the proof of the main result.

Proposition 1. For a subset $X \subseteq V$, let $\Lambda^{\mathcal{G}}_X = \cap_{a \in X} V^{\mathcal{G}}_a$. For each k, the system $\Sigma^k_{(\mathcal{G}_1,\mathcal{G}_2)}$ is equivalent to the system that consists of all equations of the form

$$\sum_{x \in \Lambda^{\mathcal{G}_1}_X} v_x = \sum_{x \in \Lambda^{\mathcal{G}_2}_X} u_x,$$

where $X \subseteq V$ ranges over all subsets with at most k elements.

Example 1. In contrast to the situation when the contributions are strictly positive (and all DAGs are then 2-distinguishable, as it was proved for previous LEMs), when we relax this assumption, we have to consider perturbation experiments that target at most three element sets of genes. The following example justifies this claim and the following theorem states it in general. Consider the two DAGs $\mathcal{G}_1, \mathcal{G}_2$ on a six element set of vertices $V = \{1, 2, 3, 4, 5, 6\}$ depicted in Figure 2. The system $\Sigma^2_{(\mathcal{G}_1,\mathcal{G}_2)}$ has $6 + \begin{pmatrix} 6 \\ 2 \end{pmatrix} = 21$ equations, where each equation corresponds to one perturbation experiment and includes variables perturbed in this experiment (directly or via propagation in the pathway graph) with coefficient 1. Substituting the following values into these equations, it is easy to verify that it is a solution of the system $\Sigma^2_{(\mathcal{G}_1,\mathcal{G}_2)}$:

$$\begin{aligned} v_1 = v_2 = v_5 = v_6 = u_1 = u_2 = u_5 = u_6 = 1, \\ \text{and} \quad v_3 = v_4 = u_3 = u_4 = -1. \end{aligned} \tag{4}$$

Since all these values are non-zero, it is an admissible solution. In order to distinguish these two DAGs, we need to consider perturbations of three genes. Here we have $V^{\mathcal{G}_1}_5 \cap V^{\mathcal{G}_1}_6 \cap V^{\mathcal{G}_1}_1 = \{1\}$, but $V^{\mathcal{G}_2}_5 \cap V^{\mathcal{G}_2}_6 \cap V^{\mathcal{G}_2}_1 = \emptyset$. Hence, by Proposition 1, we conclude that every solution has to satisfy $v_1 = 0$. Therefore it is not admissible.

Theorem 1. For every $n \geq 1$, any pair of two different DAGs on the same n-element set of vertices is 3-distinguishable.

The reader may find it interesting that the above example is the smallest possible. Namely it can be proved that for all $n \leq 5$, all DAGs over n-vertex set of genes are 2-distinguishable.

Theorem 2. For every $5 \geq n \geq 1$, any pair of two different DAGs on the same n-element set of vertices is 2-distinguishable.

We need the following result on uniqueness of solutions.

Theorem 3. For every $k \geq 1$, for every finite DAG $\mathcal{G}$ having n vertices, and for every perturbation matrix $E^{(k,n)}$, the design matrix $S = S(E^{(k,n)}, \mathcal{G})$ has rank n. Therefore, if $m = \sum_{i=1}^{k} \begin{pmatrix} n \\ i \end{pmatrix}$, then for every vector $\gamma \in R^m$, the system of equations $Sx = \gamma$ has at most one solution.

We previously showed that the originally introduced LEMs, which assume all contributions β are strictly positive, are identifiable from data measured under experiments $E^{(2,-)}$, i.e. with perturbations of all single and all pairs of nodes (Szczurek and Beerenwinkel, 2016).

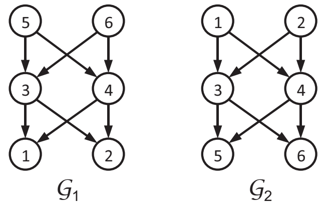

Fig. 2. An example pair of two DAGs that are not 2-distinguishable. The DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are transitive closures of the shown graphs (the transitive edges are not drawn for clarity)

We say that a LEM M is *identifiable* from data Y if any other LEM M' over the same set of vertices has a different likelihood, i.e. we have $M \neq M'$ iff $\mathcal{L}(M) \neq \mathcal{L}(M')$.

Theorem 4. Every LEM is identifiable from data Y measured under perturbation experiments $E^{(3,-)}$. LEMs over less than 6 vertices are identifiable from data measured under perturbation experiments $E^{(2,-)}$. Moreover, for identifiability of LEMs it is necessary to assume that the underlying graphs are transitively closed and acyclic.

3 Sensitivity of enhanced LEMs to noise and experimental setup

We previously demonstrated on simulated data that only extreme levels of noise and for few experimental repeats are an issue for parameter estimation and graph structure learning of original LEMs (Szczurek and Beerenwinkel, 2016). There, the data were simulated according to the assumption that the individual contributions of the perturbed genes to the measured phenotype are strictly positive. Here, we perform similar simulations for enhanced LEMs, but simulating much more realistic, both positive and negative contributions. The aim of this experiment is to show first, that despite the significantly enlarged space for allowed parameter β values, the enhanced LEMs can also be accurately inferred from data. Second, to compare the performance of the two alternative approaches to parameter estimation and graph scoring, Bayesian and BIC (Sections 2.1 and 2.2). Finally, to motivate the new approach by demonstrating dramatically decreased performance of previous LEMs when both positive and negative contributions are allowed.

To this end, we generated two test datasets. First, we generated all 28 possible network structures $\mathcal{G}$ and their corresponding contribution vectors β for LEMs over a set of three nodes {1, 2, 3}. These networks include also graphs where there is a cycle involving two nodes, and they are thus clumped into one node. Consequently, the data simulated in this dataset allows testing whether structure learning in LEMs has the ability to detect cycles and to return graphs with correctly collapsed nodes. To assess the variability of parameter estimates due to different values of β, for each graph we simulated 50 different random β vectors. Each entry β_g was drawn from the Gamma distribution $\Gamma(10, 9)$ (with parameters shape equal 10 and rate equal 9), and multiplied by −1 or 1, each with probability 0.5. This assured that the simulated β contributions can be both positive and negative, with most values close to 1 or −1 and not 0. For all simulated models, we simulated five versions of the phenotype data vectors Y, each with a different level of noise $[\sigma = \sqrt{c^{-1}} \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$, where σ denotes standard

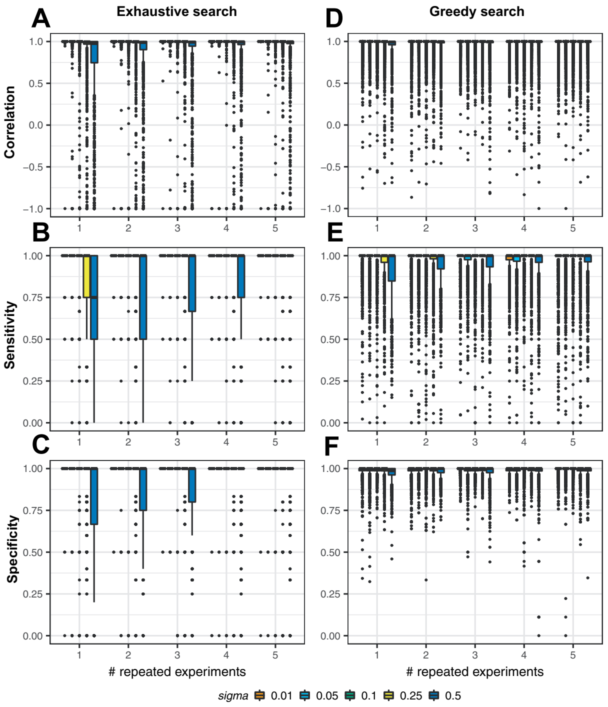

Fig. 3. Accurate parameter estimation and network structure learning for enhanced LEMs using the Bayesian approach. (**A**–**C**) Performance for three-node LEMs and exhaustive search in graph space. (A) Box plots summarizing distribution (showing 25th, 50th and 75th percentiles: horizontal bars, and 1.5 interquartile ranges: vertical line ends) of the correlation between the true β values used to simulate the data and the estimated values (y-axis) for increasing number of experimental repeats (x-axis) and for increasing noise (colors). The estimated β values are very close to the true ones for almost all simulations, with only a few outliers. Both sensitivity (B) and specificity (C) of true edge recovery are close to 1 for almost all simulated graphs, and are lowered only for extreme noise values and for few experimental repeats. (**D**–**F**) The same performance analysis as for 3-node LEMs in (A–C) but for 10-node LEMs using greedy search in graph space. For larger graphs and greedy search, the performance of parameter estimation decreases only slightly, with median correlation remaining close to 1. Compared to exhaustive search, the sensitivity and specificity of edge recovery in graphs is also only slightly lowered, and has more outliers. The median values of both sensitivity and specificity are close to 1

deviation of error terms in Equation (2)], for five different experimental setups, where the number of times each experiment was repeated was equal $1, 2, 3, 4$ or 5. Note that simulating multidimensional $Y = \{Y_1, \ldots, Y_D\}$ would not change the results of parameter estimation, as different parameters are estimated for each $Y_d \in Y$. The performance of structure learning is expected to increase, since the different Y_d are assumed to be generated from the same structure and increasing D would increase the power of structure learning. Hence, one-dimensional vector Y can be considered the worst case scenario for structure learning. For the first dataset, we simulated experiments $E^{(2,3)}$ (with all possible experiments targeting up to two from the three nodes in the graph), which by Theorem 4 allows model identifiability in this case. This dataset was used to evaluate the two enhanced LEM approaches and the previous LEMs performance when model learning is performed using exhaustive search, where all possible model structures can be evaluated.

Second, we simulated graphs and their contribution vectors for sets of 10 nodes $\{1, \ldots, 10\}$. The graphs were sampled at random using the network generation function from the nem R package as follows: for each node, the number k of outgoing edges between 0 and the total number of nodes (here, 10) was chosen according to a power law distribution with parameter $\gamma = 2.5$. We then selected k nodes having at most 10 ingoing edges and connected the node to them. Finally the graph was transitively closed. For each graph, we simulated 50 corresponding β vectors with entries from the same $\pm\Gamma(10, 9)$ distribution as in the first dataset. For such simulated models, one-dimensional data Y was simulated using the same procedure as for the first dataset, with five different noise levels and five different numbers of experimental repeats. Here, however, we simulated experiments $E^{(3,10)}$ (with all possible experiments targeting directly up to three from the set of 10 nodes in the graph). By Theorem 4, this assured model identifiability from the data. The second dataset was used to assess the performance of learning larger models, which in LEM is performed using the heuristic of greedy search over the large space of possible graphs (exhaustive search is computationally intractable for graphs with 10 nodes). Greedy search was performed with three random initializations in addition to the initialization from an empty graph.

Figure 3 summarizes performance of the Bayesian approach of enhanced LEMs to parameter estimation and network structure learning on both datasets (using exhaustive and greedy structure learning). When simulated LEMs with three nodes and reasonable noise levels ($\sigma < 0.25$) are inferred, the correlation between the estimated and true β values is concentrated at 1 for almost all simulated models, with just a few outliers with lower correlation. Larger noise does not change the distribution of obtained correlations, but only results in a larger number of outliers. Increasing the number of repeated experiments reduces the number of outliers with low correlation for all levels of noise (Fig. 3A). Both sensitivity (fraction of edges present in the simulated network graph that are correctly identified as such; Fig. 3B) and specificity (fraction of absent edges that are correctly identified as such; Fig. 3C) of exhaustive search are almost perfect already for one experimental repeat; with just a few outliers for extreme noise levels. Again, the number of outliers with lower sensitivity and specificity values decreases with more experimental repeats. For large noise, sensitivity is a bit lower and is more affected by repeat number than specificity. Compared to these results, for simulated LEMs with 10 nodes and graph inference using greedy search, the distribution of correlation between true to estimated β values over simulations has more outliers with low correlation. Still, the median of this distribution, even for very large noise and low number of repeats, is close to 1 (Fig. 3D). Similarly, the distributions of sensitivity (Fig. 3E) and specificity (Fig. 3F) have more outliers with lower values, but the medians remain close to 1.

Compared to the Bayesian approach, the BIC approach shows similarly excellent performance of parameter estimation, as well as sensitivity and specificity of graph learning, for both datasets (Supplementary Fig. S2). Only a very detailed comparison would reveal that the distribution of sensitivity values for the exhaustive search across three node graphs extends toward slightly lower values and has more outliers with low values in the case of the BIC approach. The difference between the Bayesian and the BIC approaches, however, becomes more apparent when fraction of perfectly learned simulated graphs is compared, and becomes very important when the run time is considered (Fig. 4). We define a graph to be learned perfectly, when the set of inferred edges is identical to the set in the simulated graph. Compared to the BIC approach, using exhaustive search on small graphs, a larger fraction of graphs is learned perfectly with the Bayesian approach (Fig. 4A and B). The advantage of the Bayesian approach is especially prominent for low numbers of experimental

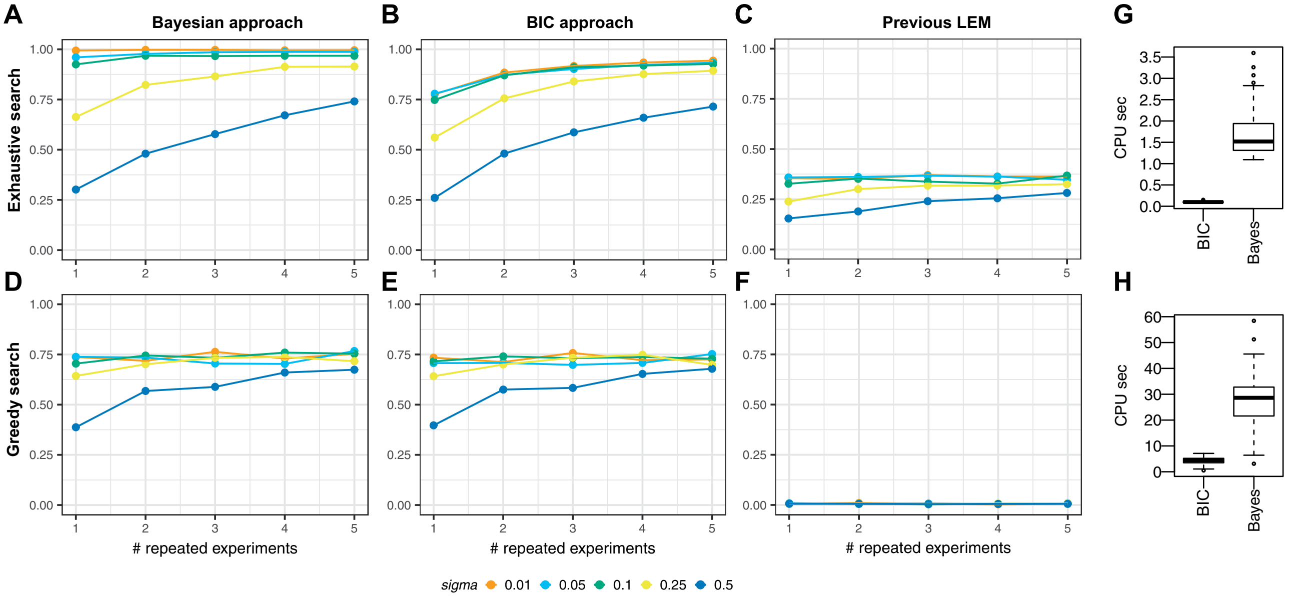

Fig. 4. Comparison of the Bayesian and the BIC approaches of enhanced LEMs, and the previous LEMs. Fraction of simulated three-node graphs that were recovered perfectly (with no missing and no added edges) using exhaustive search and the Bayesian (**A**) approach of enhanced LEMs is larger than when the BIC approach (**B**) was applied, especially for low number of experimental repeats (*x*-axis) and large noise (marked with colors). Fraction of simulated 10-node graphs that were recovered perfectly using greedy search and the Bayesian (**D**) approach is similar to the fraction when the BIC approach (**E**) was applied. The fraction of both 3-node graphs (**C**) and the 10-node graphs (**F**) that were perfectly recovered from the same data using previous LEMs is significantly lower than for the enhanced LEMs. Both the run time in CPU sec (*y*-axis) of 3-node model inference using exhaustive search across 50 simulations (**G**) and the run time of 10-node enhanced LEM inference using greedy search (**H**) is much larger for the Bayesian approach, compared to the BIC approach

repeats. The difference between the approaches is not apparent when greedy search over larger graphs is performed (Fig. 4D and E). In this case for both approaches and all experimental setups the fraction of perfectly learned graphs is decreased compared to exhaustive search over small graphs. Still, very high sensitivity and specificity results visualized in Figure 3 demonstrate that the identified graphs are close to the true simulated graphs, also for greedy search. Thus, overall, the differences in performance between the Bayesian and the BIC approaches are due to the BIC approach more often missing or inserting only a small number of edges.

The run time comparison, on the other hand, shows a huge advantage of the BIC approach (Fig. 4G and H). To compare the run times, we simulated data for and inferred 50 random LEMs with 3 nodes, and 50 random LEMs with 10 nodes. The median CPU time used for exhaustive search over the small models with the Bayesian approach is 1.5 CPU seconds and is 15.4 times longer than the median time (0.0985 s) used when the BIC approach is applied (Fig. 4G). For greedy search (with one initialization from empty graph) and the Bayesian approach, the median CPU time is 28.612 s and is 6.4 times longer than the median 4.466 s of CPU time needed for the BIC approach (Fig. 4H). In summary, the two approaches perform similarly well in terms of parameter estimation and structure learning, with slight advantage of the Bayesian approach, but their run times largely differ, with the advantage of the BIC approach.

Finally, we assessed the performance of previous LEMs (Szczurek and Beerenwinkel, 2016) using exhaustive search over 3-node graphs and greedy search over 10-node graphs, on the two simulated datasets, respectively (Supplementary Fig. S3). Importantly, both datasets were generated allowing both positive and negative values of the parameters corresponding to the gene contributions to the phenotype, which is against the assumptions of the previous version of LEMs. Previous LEMs assumed that these values are strictly positive and that they can be interpreted as absolute magnitudes of the contributions instead of their actual values. Consequently, given the simulated data vectors Y, which contained both positive and negative entries, we first transformed them into their absolute values $|Y|$ prior to application of previous LEMs. Compared to excellent results of enhanced LEMs on this data, the previous LEMs perform poorly (Supplementary Fig. S3). As can be expected, with wrong assumption about the parameter values, parameter estimation using previous LEMs fails, with median correlation between the true and the estimated contribution values equal 0 (Supplementary Fig. S3A and D). The performance of structure learning is also poorer than that of the enhanced LEMs, with the greedy search performing worse than exhaustive search. When exhaustive search over three-node graphs is applied, median sensitivities are close to 1 for lower values of noise ($\sigma \leq 0.1$;Supplementary Fig. S3B). The specificity, however, is considerably lower than obtained by enhanced LEMs, regardless of noise and experimental repeat values (Supplementary Fig. S3C). The fraction of three-node graphs learned perfect by previous LEMs from this data is much lower than obtained when enhanced LEMs were applied (Fig. 4C). Compared to these results, when greedy search over 10-node networks is performed, sensitivity drops drastically (Supplementary Fig. S3E), while specificity increases (Supplementary Fig. S3F), indicating that for previous LEMs the greedy search identifies structures with too few edges. Consequently, the fraction of structures learned perfect drops down to zero (Fig. 4F). These results indicate very clearly the need for introduction of enhanced LEMs when more realistic data are to be modeled accurately.

4 Application to infection network inference from siRNA screening data

To demonstrate the performance of enhanced LEMs on real data, we utilized on- and off-targets of siRNA interventions as

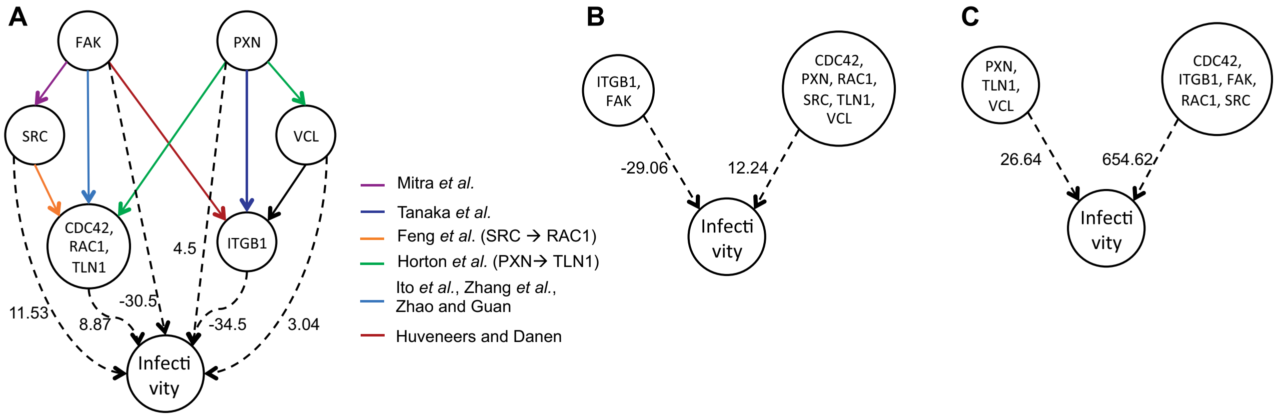

Fig. 5. Structure and parameters of the B. *henselae* infection network inferred using LEMs from the siRNA screen data of Ramo *et al.* (2014). Nodes are labeled with gene names of the network members. The nodes labeled by several gene names represent cliques involving these genes. The color of bold edges indicates support found in the literature. The dashed edges represent the individual contributions of the network nodes to the infectivity phenotype, and are labeled by their values, scaled by 10^3. (**A**) Enhanced LEM model learned from this data using the Bayesian approach. (**B**) Enhanced LEM model learned using the BIC approach. (**C**) Previous LEM model learned from this data

combinatorial perturbations. We analyzed siRNA kinome screen by Ramo *et al.* (2014) carried out in the HeLa ATCC-CCL-2 cell line using Qiagen (Human Kinase siRNA Set V4.1) reagents, with four different siRNAs per on-target gene. The measured phenotype was infectivity of the cells with the pathogen B. *henselae*, corresponding to a rate of infection per well, derived from image features collected under the experiments (Ramo *et al.*, 2014). Before modeling, we removed readouts from control and bad quality wells and filtered out siRNAs which target essential genes (cell killers). Next, the data were normalized using the B-scoring algorithm (Brideau *et al.*, 2003) in order to remove systematic within-plate effects. At the final preprocessing step, the data were Z-scored to eliminate experimentally introduced cross-plate biases. Prediction of off-targets of the siRNAs on all genes was performed using TargetScan (Lewis *et al.*, 2005).

In any siRNA experiment, the measured phenotype can be assumed to be a combination of contributions of all perturbed genes, i.e. also all remaining genes outside of the network. Therefore to be able to apply LEMs to infer a signaling network of interest (here the infection network), the phenotype needs to be corrected by removing the contributions from the remaining genes. To this end, we used the Lasso (Tibshirani, 1994) to estimate contributions of all genes to the phenotype measured in all experiments, assuming the genes were independent (not connected in any network). Since only the infection network was stimulated in the study (Ramo *et al.*, 2014), this assumption is valid for all remaining genes outside of the infection network. Next, the phenotype was corrected by subtracting the estimated contributions of those remaining genes which were perturbed in each experiment. Formally, this procedure is motivated in Supplementary Material. We next constructed the experiments matrix to contain only eight columns, corresponding to the same eight genes in the B. *henselae* infection network studied by Srivatsa *et al.* (2018), and 44 000 rows, corresponding to such siRNAs that either on- or off-targeted any of the eight genes. Note our approach to constricting the experiments matrix is different from Srivatsa *et al.* (2018), who used only data from 35 experiments corresponding to siRNAs on-targeting the genes in the network, and is intended to increase the power of our analysis. Finally, we applied enhanced LEMs using greedy search with 100 random initializations (in addition to the initialization from an empty graph) using Bayesian and BIC approaches to infer interactions in the infection network from the corrected phenotype (Fig. 5A and B). Using the Bayesian approach, the algorithm identified a network containing a cycle involving CDC42, RAC1, and TLN1, with the connection between CDC42 and RAC1 known to be involved in the regulation of actin cytoskeleton (Verma and Ihler, 2002) (Fig. 5A). Transitive closure results in a clique connecting all of these genes together, which is collapsed into a single node. Out of eight identified internode interactions, seven found support in the literature. Interactions between Paxillin (PXN) and talin1 (TLN1) and between Paxillin (PXN) and vinculin (VCL) were assigned high evidence scores in the curated network of integrin anhesome (Horton *et al.*, 2015). Interaction between focal adhesion kinase (FAK) and Cdc42 were determined to play a role in the context of cellular motility by several studies (Ito *et al.*, 1982; Zhang *et al.*, 2004; Zhao and Guan, 2011). Src-dependent activation of Rac1 was studied in the context of glioma tumorigenesis (Feng *et al.*, 2011). The recovered interaction between FAK and integrin $\beta 1$ (ITGB1) is known to be active in adhesion signaling (Huveneers and Danen, 2009). The identified network contains also the established interaction FAK– Src (Mitra *et al.*, 2005). In addition to the previously established interactions, our model contains a novel one, VCL → ITGB1, constituting a novel hypothesis about the mechanisms of B. *henselae* infection.

Enhanced LEMs using the BIC approach identified a network that is similar to the network found using the Bayesian approach, but more cyclic (Fig. 5B). The network has two disjoint cycles, which were collapsed into two unconnected clique nodes due to transitive closure. One cycle contains ITGB1 and FAK, which are known to interact in adhesion signaling (Huveneers and Danen, 2009). An edge between FAK and ITGB1 was also found using the Bayesian approach. The second cycle contains all remaining genes of the B. *henselae* infection pathway. This cycle contains the smaller cycle containing CDC42, RAC1 and TLN1, suggested also by the Bayesian approach. In addition, BIC placed in this cycle also VCL and PXN, which are known to interact with each other. Finally, the previous LEM model applied to the same data found a structure also containing two cycles, but containing different subsets of genes. To be able to apply previous LEMs, we transformed the infectivity phenotype Y, which had both negative and positive values, to its absolute value $|Y|$. Since the Bayesian approach had a slightly better performance in our simulation study (Section 3), and much better performance than previous LEMs, we consider the model found using the Bayesian approach (Fig. 5A) most likely to be the closest to the true biological network.

In contrast to other effects models, LEM infers the contributions each signaling node has to the measured phenotype upon its perturbation. Additionally, in contrast to the previous LEMs, the enhanced LEMs account for the positive or negative direction of these contributions. In application to the B. *henselae* network, using both the Bayesian and the BIC approaches, the enhanced LEMs estimated the individual contributions of FAK and ITGB1 to infectivity are strongly negative and most significant among the genes in this network. This indicates that FAK and ITGB1 play the most important roles for successful infection by the pathogen, in agreement with previous findings (Truttmann *et al.*, 2011).

5 Discussion and conclusions

This paper contributes two main results. First, it introduces an important extension to the previously proposed LEMs. By allowing both negative and positive contributions of the network members to the measured phenotype, model assumptions are much closer to reality than in the original model formulation. The enhanced model expresses how network members jointly regulate the downstream effects, where some of the members may up- and others may down-regulate these effects. Thus, using the enhanced LEMs, we can now both learn the graph representing the network structure, and a more involved representation of the regulatory layer than before. Second, the paper brings a proof of identifiability of enhanced LEMs from combinatorial experiments where single, pairs, and triplets of network nodes are performed. For small enhanced LEMs, with up to five nodes, we show identifiability with perturbations of only up to two nodes.

We meant here that the comparison of the way the regulatory layer is modeled in NEMs (Markowetz *et al.*, 2005), original LEMs (Szczurek and Beerenwinkel, 2016) and the here introduced enhanced LEMs reveals that the models become increasingly expressive. The binary information of whether a certain effect is regulated by a certain single network gene or not, modeled in NEMs, was replaced by modeling the effects as linear combinations of positive contributions from all network genes in LEMs, with which the current model can be any real values but not zero. Increase in expressiveness, however, clearly comes with a price of larger data required for model learning. While NEMs are inferred from only perturbations of all single network nodes, LEMs already require combinatorial perturbation data of single and double nodes. Enhanced LEMs with more than five nodes need not only single and double, but also triple perturbations. We anticipate that such combinatorial experiments will become increasingly available. Technological advances in automated RNAi screening (Lambeth *et al.*, 2010) make such combinatorial experimental regimes more and more efficient. As our and others (Srivatsa *et al.*, 2018), examples of successful inference of interactions in the B. *henselae* infection network from siRNA screening data suggest, also the off-target perturbations can be utilized as a rich source of combinatorial interventions on signaling networks. This approach, however, depends on the ability to accurately predict the off-targets, which can be very challenging. Currently, our model does not account for false positive and false negative predictions. Finally, it considers only binary perturbation states and is not applicable to pooled siRNA reagents. Taking account of the prediction errors as well as strength of the perturbations would be a valuable extension of the approach.

Although enhanced, the LEMs introduced in this paper still have several limitations. First, LEM assumes that all network members have nonzero contribution to the observed downstream phenotypes, which is biologically less likely especially for the kinases high up in the signaling cascade, or the signaling receptors, corresponding to the roots of the network graph. Second, LEM is applicable only to small networks. Efficiency of the greedy model search procedure could be improved by maintenance of tabu lists and avoiding recently visited neighbors, as well as storage of only the minimum necessary information about the neighbors. Third, the requirement of combinatorial perturbations being available for all triplets of genes in the network is a limiting factor in the applicability of LEM to larger networks. One way to deal with this problem is to explore the equivalence classes of LEMs, which could be obtained when fewer combinations of perturbations would be available, and allow the method to return more than one equally scoring network. Another way, as proposed in this work, is to take advantage of the combinatorial perturbations resulting from multiple off-targets of siRNAs. Still, even with these limitations, the enhanced LEMs do bring significantly improved expressiveness of the model in comparison not only to the original LEM, but also to other methods. With the proved identifiability requirements, it is clear which experiments need to be done to be able to reliably, and, as we show in simulations and in application to infection network recovery from siRNA screening data, accurately infer both network structure and the regulatory layer from the phenotypes measured downstream.

Funding

This work was supported by the National Science Centre, Poland [2015/19/P/NZ2/03780 to J.T. and E.S.]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 665778.

Conflict of Interest: none declared.

References

Agrawal,N. *et al.* (2003) RNA interference: biology, mechanism, and applications. *Microbiol. Mol. Biol. Rev*, **67**, 657–685.

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.

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

Brideau,C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.*, **8**, 634–647.

Feng,H. *et al.* (2011) Activation of Rac1 by Src-dependent phosphorylation of Dock180(Y1811) mediates PDGFR-stimulated glioma tumorigenesis in mice and humans. *J. Clin. Invest*, **121**, 4670–4684.

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

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

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

Horton,E.R. *et al.* (2015) Definition of a consensus integrin adhesome and its dynamics during adhesion complex assembly and disassembly. *Nat. Cell Biol.*, **17**, 1577–1587.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Huveneers,S. and Danen,E.H. (2009) Adhesion signaling - crosstalk between integrins, Src and Rho. *J. Cell. Sci.*, **122**, 1059–1069.

Ito,S. *et al.* (1982) Vinculin phosphorylation by the src kinase: inhibition by chlorpromazine, imipramine and local anesthetics. *Biochem. Biophys. Res. Commun.*, **107**, 670–675.

Jackson,A.L. *et al.* (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.*, **21**, 635–637.

Lambeth,L.S. *et al.* (2010) A direct comparison of strategies for combinatorial RNA interference. *BMC Mol. Biol.*, **11**, 77.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Lisitskaya,L. *et al.* (2018) DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nat. Commun.*, **9**, 5165.

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

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**, i305–312.

Mitra,S.K. *et al.* (2005) Focal adhesion kinase: in command and control of cell motility. *Nat. Rev. Mol. Cell Biol.*, **6**, 56–68.

Mohr,S.E. and Perrimon,N. (2012) RNAi screening: new approaches, understandings, and organisms. *Wiley Interdiscip. Rev. RNA*, **3**, 145–158.

Molinelli,E.J. *et al.* (2013) Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput. Biol.*, **9**, e1003290.

Pirkl,M. *et al.* (2016) Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean nested effect models. *Bioinformatics*, **32**, 893–900.

Ramo,P. *et al.* (2014) Simultaneous analysis of large-scale RNAi screens for pathogen entry. *BMC Genomics*, **15**, 1162.

Schmich,F. *et al.* (2015) gespeR: a statistical model for deconvoluting off-target-confounded RNA interference screens. *Genome Biol.*, **16**, 220.

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

Sigoillot,F.D. and King,R.W. (2011) Vigilance and validation: keys to success in RNAi screening. *ACS Chem. Biol.*, **6**, 47–60.

Srivatsa,S. *et al.* (2018) Improved pathway reconstruction from RNA interference screens by exploiting off-target effects. *Bioinformatics*, **34**, i519–i527.

Szczurek,E. and Beerenwinkel,N. (2016) Linear effects models of signaling pathways from combinatorial perturbation data. *Bioinformatics*, **32**, i297–i305.

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

Szczurek,E. *et al.* (2011) Deregulation upon DNA damage revealed by joint analysis of context-specific perturbation data. *BMC Bioinformatics*, **12**, 249.

Terns,M.P. (2018) CRISPR-based technologies: impact of RNA-targeting systems. *Mol. Cell*, **72**, 404–412.

Tibshirani,R. (1994) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Ser. B*, **58**, 267–288.

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

Truttmann,M.C. *et al.* (2011) *Bartonella henselae* engages inside-out and outside-in signaling by integrin $\beta 1$ and talin1 during invasome-mediated bacterial uptake. *J. Cell. Sci.*, **124**, 3591–3602.

Verma,A. and Ihler,G.M. (2002) Activation of Rac, Cdc42 and other downstream signalling molecules by *Bartonella bacilliformis* during entry into human endothelial cells. *Cell. Microbiol*, **4**, 557–569.

Zhang,Z. *et al.* (2004) The phosphorylation of vinculin on tyrosine residues 100 and 1065, mediated by SRC kinases, affects cell spreading. *Mol. Biol. Cell*, **15**, 4234–4247.

Zhao,X. and Guan,J.L. (2011) Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. *Adv. Drug Deliv. Rev.*, **63**, 610–615.