PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

NEMix: Single-cell Nested Effects Models for Probabilistic Pathway Stimulation

Juliane Siebourg-Polster[1,2], Daria Mudrak[3], Mario Emmenlauer[4], Pauli Rämö[4], Christoph Dehio[4], Urs Greber[3], Holger Fröhlich[5], Niko Beerenwinkel[1,2]*

1 Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, **2** SIB Swiss Institute of Bioinformatics, Basel, Switzerland, **3** Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland, **4** Biozentrum, University of Basel, Basel, Switzerland, **5** Algorithmic Bioinformatics, Bonn-Aachen International Center for IT, University of Bonn, Bonn, Germany

* niko.beerenwinkel@bsse.ethz.ch

Citation: Siebourg-Polster J, Mudrak D, Emmenlauer M, Rämö P, Dehio C, Greber U, et al. (2015) NEMix: Single-cell Nested Effects Models for Probabilistic Pathway Stimulation. PLoS Comput Biol 11(4): e1004078. doi:10.1371/journal.pcbi.1004078

Editor: Satoru Miyano, University of Tokyo, Japan

Received: June 26, 2014

Accepted: December 8, 2014

Published: April 16, 2015

Data Availability Statement: Data are freely available as part of the R/Bioconductor package 'nem' at http://www.cbg.ethz.ch/software/NEMix.

Funding: CD, UG, and NB were supported by SystemsX.ch (www.systemsx.ch), the Swiss initiative in systems biology, under grant No. 51RT-0_126008 (InfectX), and CD and NB under grant No. 51RTP0_151029 (TargetInfectX). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Nested effects models have been used successfully for learning subcellular networks from high-dimensional perturbation effects that result from RNA interference (RNAi) experiments. Here, we further develop the basic nested effects model using high-content single-cell imaging data from RNAi screens of cultured cells infected with human rhinovirus. RNAi screens with single-cell readouts are becoming increasingly common, and they often reveal high cell-to-cell variation. As a consequence of this cellular heterogeneity, knock-downs result in variable effects among cells and lead to weak average phenotypes on the cell population level. To address this confounding factor in network inference, we explicitly model the stimulation status of a signaling pathway in individual cells. We extend the framework of nested effects models to probabilistic combinatorial knock-downs and propose NEMix, a nested effects mixture model that accounts for unobserved pathway activation. We analyzed the identifiability of NEMix and developed a parameter inference scheme based on the Expectation Maximization algorithm. In an extensive simulation study, we show that NEMix improves learning of pathway structures over classical NEMs significantly in the presence of hidden pathway stimulation. We applied our model to single-cell imaging data from RNAi screens monitoring human rhinovirus infection, where limited infection efficiency of the assay results in uncertain pathway stimulation. Using a subset of genes with known interactions, we show that the inferred NEMix network has high accuracy and outperforms the classical nested effects model without hidden pathway activity. NEMix is implemented as part of the R/Bioconductor package 'nem' and available at www.cbg.ethz.ch/software/NEMix.

Author Summary

Experiments monitoring individual cells show that cells can behave differently even under same experimental conditions. Summarizing measurements over a population of cells can lead to weak and widely deviating signals, and subsequently applied modeling approaches, like network inference, will suffer from this information loss. Nested effects models, a method tailored to reconstruct signaling networks from high-dimensional read-outs of gene silencing experiments, have so far been only applied on the cell population level. These models assume the pathway under consideration to be activated in all cells. The signal flow is only disrupted, when genes are silenced. However, if this assumption is not met, inference results can be incorrect, because observed effects are interpreted wrongly. We extended nested effects models, to use the power of single-cell resolution data sets. We introduce a new unobserved factor, which describes the pathway activity of single cells. The pathway activity is learned for each cell during network inference. We apply our model to gene silencing screens, investigating human rhino virus infection of single cells from microscopy imaging features. Comparing the learned network to the known KEGG pathway of the genes shows that our method recovers networks significantly better than classical nested effects models without capturing of hidden signaling.

This is a PLOS Computational Biology Methods paper.

Introduction

Network inference benefits substantially from perturbation experiments, such as RNA interference (RNAi) screens. Monitoring high-dimensional effects of gene silencing enables inference of non-transcriptional network structures that cannot be learned on observational data alone [1]. Nested effects models (NEMs) are a class of probabilistic graphical models that aim at learning hierarchical dependencies from such intervention experiments. Upon perturbing nodes in a signaling graph, their connectivity is inferred from the nested structure of observed downstream effects. The concept was first introduced in [2]. Since then, many further additions concerning, for example, parameter inference, structure learning, and data integration, were developed [3, 4]. In addition, dynamic models for time series data have been developed [5–7]. In [5], a first application of dynamic nested effects models to time laps microscopy data has been described, but the model can not handle single-cell data. A Bayesian network representation of NEMs in [8] introduces a probabilistic notation for signal propagation, but in practice the signaling is kept deterministic. In all previous NEM models and applications, the signaling pathway under observation is assumed to be active and the signal flow disrupted by silencing the signaling genes one by one.

In principle, RNAi experiments are a highly informative for learning NEMs. Perturbations are introduced by gene silencing in cells through RNA interference using siRNAs [9, 10]. Effects of the knock-downs are then captured by high-dimensional down-stream observations. The screening data analyzed here, comprises imaging data of thousands of individual cells for genome-wide gene silencing. However, the experiments come at the cost of high noise levels, as well as biological and technical biases, including off-target effects [11, 12]. These confounding factors complicate the analysis and interpretation of the screening results. On the other hand, RNAi screens currently reach very high resolution. Per knock-down, the present data sets comprise about 300 image features for several hundred individual cells, which allows for a very detailed analysis of a knock-down event. However, it has been shown that measurements from individual cells of the same experiment can differ widely, for example, due to local environmental differences [13, 14]. Such variation on the single cell level needs to be accounted for. Otherwise, an ambiguous signal is obtained, when averaging over the cell population of a knock-down.

Here, we specifically investigate single-cell observations of pathogen infection screens [15–17]. The experiments monitor cells with an siRNA knock-down during infection with human rhinovirus (HRV). After siRNA knock-down, the pathogen is added to the cells, and the success of infection as well as many other cellular features are extracted from microscopy images taken of the cells from each experiment [18–20]. The aim is to infer a signaling cascade involved in pathogen entry in to the host cell. However, a challenge in the analysis of data from this experimental setup is that by experimental design even in mock controls (i.e., infection without knock-down) the infection rate is far from complete. In fact, the multiplicity of infection (MOI) of the assay was optimized to reach 30 to 50% infected cells, such that both infection-decreasing and infection-increasing hits can be detected. Which cells in the population finally get infected is, at least to some extent, the result of stochastic effects, since cellular processes can be differently manifested in different cells. The multi-functional nature of proteins, for instance, enables a single host factor to enhance a signaling cascade, and at the same time may antagonize other processes that support or inhibit infection. Obviously, infected cells were reached by a pathogen triggering some signal to get internalized. However, for uninfected cells, it is unknown whether a pathogen actually attempted to infect them, which is crucial for determining the effect that the gene knock-down had on these cells. Wrongly assuming that the pathway is active, even though it is not, can result in conflicting knock-down schemes. In the original NEM setting, individual cell observations are summarized for each signaling gene.

To address the problem of network learning when the activation state of the signaling pathway is unknown we introduce a new model, called NEMix, extending the existing NEM framework in several ways. First, we do not summarize the data across cells, but rather perform network inference using the single-cell observations directly. Furthermore, we model the unknown pathway activation with an additional hidden random variable in the graph of signaling genes. The activation state is then estimated for each individual cell. The pathway activity can be regarded as an additional hidden silencing event in the signaling graph. We introduce a general theoretical framework for probabilistic combinatorial knock-downs in NEMs. We develop our model for the most general case, not making any assumptions about the signal propagation. We have implemented the special case of one hidden variable with probabilistic knock-down, where the remaining network is kept deterministic. For inference of the hidden pathway state, we developed an EM algorithm [21]. This step is repeated for each proposal structure during the network search.

Results

Network inference under unknown pathway activity

We developed NEMix, a new model based on NEMs, which allows to estimate activity of a pathway in individual cells. A NEM is a graphical model, consisting of two graphs. The transitively closed graph Φ encodes dependencies among signaling gene nodes $S_s \in \mathcal{S}$, which are silenced one by one. The bipartite graph Θ connects a set of observable feature nodes $E_e \in \mathcal{E}$ uniquely to the signaling genes (Fig. 1A). We seek the structure of Φ, i.e., the topology of the signaling pathway, by inferring it from the nested structure of observed effects. For a data set $\mathcal{D} = (d_{ek})$ of a set of knock-down experiments $k \in \{1, \ldots, K\}$ and observed features

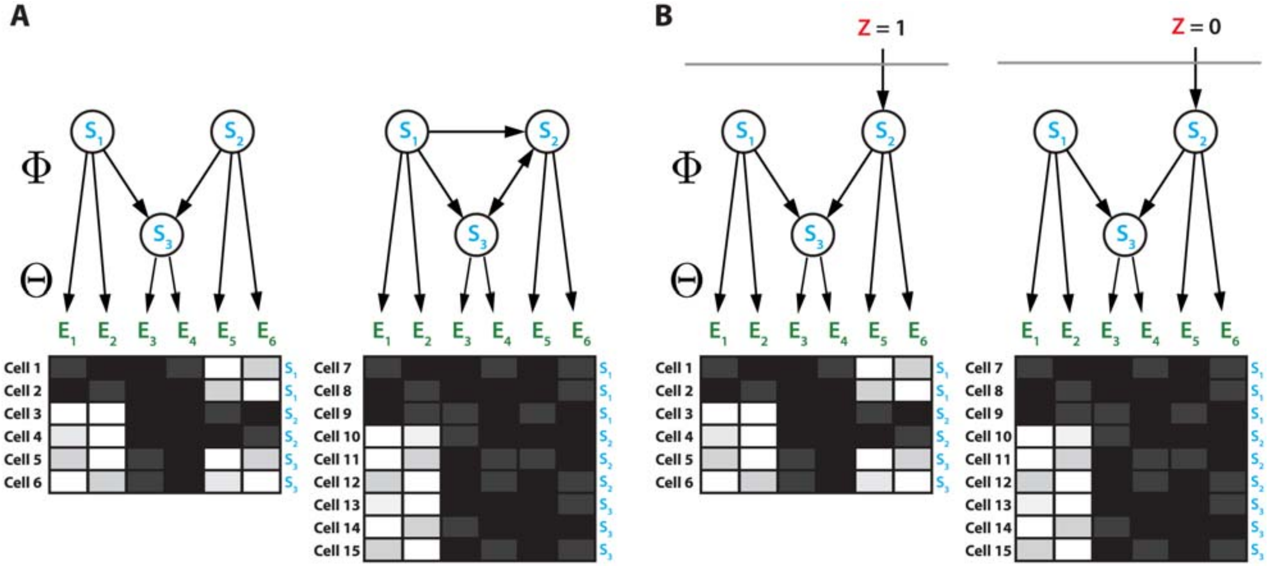

Fig 1. NEM versus NEMix. A schematic example is shown comparing the classical nested effects model (NEM; panel **A**) with the new nested effects mixture model (NEMix; panel **B**) on six features observed in 15 individual cells. Blue nodes in the graph depict the signaling genes S_1, S_2, and S_3 that have been silenced and whose dependency structure is sought. The observed features $E_1, \ldots, E_6$ are shown in green. Each box below the graphs indicates the observed (noisy) features (e.g., image-based read-outs) for a single cell. Within each box, dark entries indicate an effect of the knock-down on the feature, light entries indicate no effect. In cells 1 and 2 (left in both **A** and **B**), the pathway has been activated via S_2, whereas in cells 3, 4, and 5 (right in both **A** and **B**) it has remained inactivated. In the latter case, the effects of silencing S_2 are masked and the resulting silencing scheme then differs from the one where the pathway is stimulated. Classic NEMs (**A**) could explain such a heterogeneous cell population only by two different signaling graphs Φ. By contrast, with the NEMix model proposed in this work (**B**), both observed patterns can be explained by the same signaling graph Φ, because the hidden pathway stimulation Z (shown in red) is modeled explicitly. In the NEMix model, Z is a hidden binary random variable indicating pathway activation ($Z = 1$), which occurs with probability $P(Z = 1) = p_1$.

doi:10.1371/journal.pcbi.1004078.g001

$e \in \{1, \ldots, m\}$, the likelihood function given Φ and θ is

$$P(\mathcal{D} \mid \Phi, \theta) = \prod_{e=1}^{m}\prod_{k=1}^{K} P(d_{ek} \mid \Phi, \theta_e = s), \tag{1}$$

where $\theta_e = s$ indicates that feature e is connected to signaling gene $s \in \mathcal{S}$.

The NEMix model consists of the same two graphs Φ and Θ, but has an additional binary hidden variable Z added to the signaling graph Φ. Its connections to the signaling genes, as well as its overall knock-down probability $p_0 = P(Z_{kc} = 0)$, are unknown and inferred for each individual cell during the network reconstruction process. Given single cell data $\mathcal{D} = (d_{ekc})$ with $c = 1, \ldots, c_k$ cells in knock-down experiment k, the likelihood function of the NEMix model, given Φ and θ, is

$$P(D \mid \Phi, \theta) = \prod_{e=1}^{m}\prod_{k=1}^{K}\prod_{c=1}^{c_k}\sum_{j\in\{0,1\}} p_j P(d_{ekc} \mid \Phi, \theta_e = s, Z_{kc} = j). \tag{2}$$

A detailed derivation of the model and its implementation are given in the Models section. If a signal is activating a pathway, or parts of it, the signal flow is the same as in the NEM. Also the observed knock-down effects for the features E_e are the same. However, when the pathways input signal is inactivated, the knock-down pattern of the features changes (Fig. 1A and B, cells 7 to 15). Not accounting for the pathway disruption can mislead inference of the structure Φ (Fig. 1A, left model).

The connectivity of Z is learned in a greedy fashion during structure inference. For the knock-down probability of the hidden variable, p_0, we implemented an EM algorithm, which estimates jointly p_0 from each cell's observation and the connections of observations to signaling genes, θ. In the following, we show improved network inference with NEMix in simulations and then infer networks of high accuracy, from single cell gene silencing experiments.

Simulation study

To test our model, we performed a large simulation study. We generated 30 network structures with 5 signaling genes, randomly sampled from KEGG pathway maps [22] as previously described in [6]. To each network the hidden input signal was attached randomly. The resulting 30 sample networks are shown in supplementary S1 Fig. From each network, we sampled 50 data sets on 300 observed features in the following way. For each gene, we simulated single knock-downs in 200 cells. To the observed features we added another 30 noise features, not attached to any signaling gene. The data sets were generated in the following way. We sampled effects from a normal distribution with mean $m_e = 1$ and non-effects from a normal distribution with mean $m_n = 0$. The standard deviation for each experiment was sampled uniformly between 2 and 2.5. We furthermore sampled 200 cells for control experiments. The negative control cells do not show any effects and are therefore drawn from the non-effect distribution. The positive control cells always show effects and hence are drawn from the effect distribution. The whole simulation process was repeated for five different fractions of pathway disruption, $p_0 \in \{0, 0.3, 0.5, 0.8, 1\}$. NEMix inference was restarted for 16 initial networks. Each of them consists of the empty graph Φ plus a unique attachment of Z to the signaling genes. Setting the maximal out-degree of Z to two, there are 16 possible such attachments of Z. This regularization on the edges of Z reduces the search space significantly. During structure search we also imposed this restriction, but additionally allowed transitive edges that had to be added as a consequence of the insertion of any edge connecting Z to a gene (see Models section).

We compared NEMix to two other NEM models and, for a baseline comparison, to a random approach, where network edges are sampled uniformly with probability $1/n$, where $n|\Phi|$ is the number of signaling nodes. This probability was chosen as it creates networks with approximately the same number of edges as in the original graphs. To assess the impact that pathway disruption has on the cell population level, we ran the simulations on a standard NEM using the log-likelihood model introduced in [23]. For the NEM approach we had to summarize the single cell observations to the gene level. For these gene-level data sets we used p-values of a Wilcoxon test comparing the cell population of a knock-down to the control distribution. From the p-value distributions a Beta-Uniform-Mixture model was estimated. For each feature a density value is calculated from this model, indicating the effect strength of the knock-down. These density values are used as the input data, as previously introduced in [23]. The third approach, called single-cell NEM (sc-NEM). is a NEMix model on individual cell observations, but with fixed $p_0 = 0$, i.e., a single-cell observation-based NEM without considering uncertain pathway activity. For all three models, we applied a uniform prior on the feature attachments θ, and no prior knowledge was added for the network structures Φ. The NEMix parameter p_0 was initialized by drawing from a uniform distribution in each EM restart. As NEMix and sc-NEMs infer networks on single-cell observations, we calculated log odds ratios from each observation based on the positive and negative control distributions (see 'Modeling the effect likelihoods' in S1 Text). For NEMs and sc-NEMs, we used maximum likelihood estimation to infer θ and in the NEMix it is estimated by in an EM algorithm. Structure learning is performed using a greedy hill climbing algorithm, initialized with an empty network.

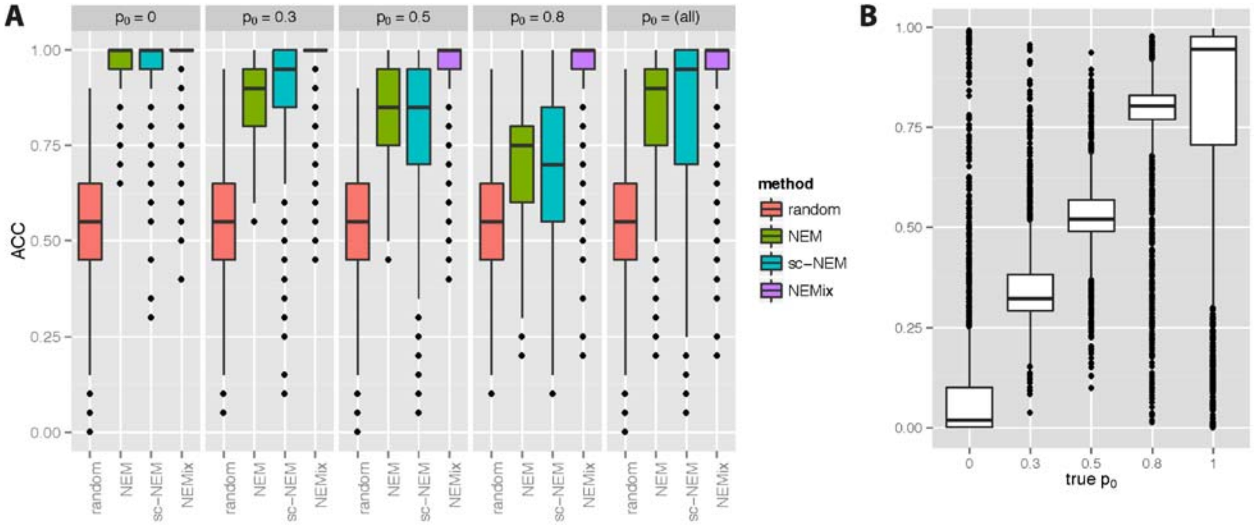

Fig 2. Performance comparison of the simulations. (A) Simulation results are summarized based on the accuracy of recovered edges for the compared methods. The methods are random, random edge sampling with rate $\frac{1}{n}$; NEM, the normal NEM inference; sc-NEM, the cell level NEM and NEMix, using the NEMix inference with the hidden pathway state. All methods were run on 50 simulated data sets from 30 sample networks, repeated for different knock-down probabilities of the pathway state p_0. (B) For the NEMix model, the distributions of inferred p_0 values are compared to the true p_0.

doi:10.1371/journal.pcbi.1004078.g002

Fig. 2A summarizes the overall performance for all methods and the different fractions of pathway signal perturbation p_0. We display accuracy of the edge recovery, for varying p_0. We also calculated the area under the ROC curve (AUC) based on the edge frequencies of the 50 replicate data sets, which yielded similar results in terms of accuracy (see supplementary S2 Fig). As expected, all methods performed equally well when there is no signal disruption ($p_0 = 0$). However, when p_0 is moderate to high, NEMix performs significantly better than the other methods. If the triggering signal is always turned off, performance of all methods drops drastically. Intuitively, this is because in such a special case, all features downstream of Z always show an effect and hence they cannot be used for structure learning. For example, if, in Fig. 1B, Z is inactive for each cell, we could not infer the structure among S_2 and S_3. In reality though, permanent shut down of the pathway is very unlikely. For the infection screens $p_0 = 1$ would mean that no cell is ever infected. Pathway activity estimates are also of overall high accuracy (Fig. 2B). Although simulation results demonstrate that the performance of learning Z and θ varies, depending on the network structure, the average performance is very good (S3 Fig, S4 Fig, S5 Fig).

Currently, one of the main obstacles for learning larger NEMix models is the fast growing run-time for $n > 5$ network nodes. Run-time is further increased by a factor of n, when initiating the algorithm with each possible connection of Z to one of the knock-down genes. To assess its performance on larger networks, we ran a reduced simulation study on $n = 5$, 10, and 15 genes. The setup and results of the study are described in detail in S6 Fig. Larger networks of 15 nodes can still be estimated very well (S6 Fig. A) and estimation of the parameter p_0 even improves (S1 Fig. D). However, the average time to estimate a 15-node network was 9.5 hours. This is substantially more than the average 1.9 hours needed for 10-node networks. Thus, in a highly parallelized computing environment, even larger networks can be estimated.

We also assessed the connection of features to the signaling genes in the inferred graph Θ. There can be situations, where attachment of features is equally likely for several signaling genes. In these cases, where no single gene is preferred, we counted a feature as correctly attached if it was connected to any of the signaling genes with equal likelihood. Accuracy of the θ

estimates is high ($> 80\%$) for small p_0 values and decreases with increasing p_0. For small p_0, also performance of the sc-NEMs is good, which shows the advantage of learning on the single-cell data level. However, NEMix stands out from the other methods for higher p_0. Recovery of noise features, i.e., correct filtering of the additionally added uninformative features, is not strongly affected by the hidden signal (see supplementary S7 Fig). Analyzing individual networks, one again observes high variation in performance (see supplementary S8 Fig).

Application to pathogen infection experiments

We applied NEMix in the context of infection signaling, using the RNAi screening data monitoring HRV infection, mentioned in the introduction. Briefly, viruses were added to the siRNA transfected cells and after an incubation time, cells were fixated, stained, and then imaged. Subsequently, 360 cell features were extracted from the 9 images per knock-down experiment using the software CellProfiler [24]. For the whole experimental procedure the protocols of [17] were followed. The HRV assay is rather short with an infection time of only seven hours, resulting in measurements proximal to the infection event. The short time range is advantageous, because it leaves less room for confounding developments in the cells. Furthermore, the used antibody resulted in clean readouts, well to extract from the images.

Before using the data for network inference, we performed two additional filtering steps. For each knock-down, the well is split into 9 images. They are arranged in three rows and three columns. We used only the middle image, because it is of the highest quality. In this way we avoided too many out-of-focus cells, which bias especially the cell texture features. After this filtering step, we had around 200 to 300 cells per knock-down. A second filtering step concerns siRNA off-targets [25]. We sought to avoid confounding by this effect and therefore selected only genes with low predicted off-target effects as described in ‘siRNA filtering for off-targets’ of S1 Text.

We applied NEMix to a small subset of the screened genes, in order to recover a known pathway. We decided on the well-known MAP-Kinase signaling cascade as a proof of principle, for several reasons. First, it has been studied and validated in great detail [26–28], such that the available signaling network from the KEGG database [22] can be used as a reliable source to compare to. Second, the pathway is known to be involved in HVR infection signaling, where it is associated with asthmatic and COPD exacerbation [29–31]. Finally, we observed an enrichment for low off-target siRNAs in this pathway when performing a gene set enrichment analysis [32] (see supplementary S9 Fig). We then selected a small subset of 8 MAP-Kinase pathway genes for analysis based on the derived score for predicted off-target effects. Nodes of KEGG pathways can contain several genes. We selected genes such that they are all assigned to different KEGG nodes using a weighted maximum bipartite matching of low off-target siRNAs and unique KEGG nodes. After gene selection, we inferred networks for the 5 and 8 genes with lowest off-target score.

Like in the simulation study above, we compared the NEMix model to the NEM and the sc-NEM approach. As input data sets, the local effect likelihoods from the selected knock-down gene experiments were computed as follows. As the experiments lack reliable controls, we instead used a random sample of cells from the plate on which the gene was located, assuming that the majority of knock-downs will not have an effect. Like for the simulation study, we derived the cell population effects for the NEM from Wilcoxon tests, comparing the knock-down experiment to the control. From the resulting p-value distributions, effect strengths for the features were estimated using the Beta-Uniform-Mixture model. Log odds ratios for sc-NEMs and NEMix in this case are calculated only based on one control distribution (see Models section). NEMix inference again is repeated for the 16 initial networks of all possible

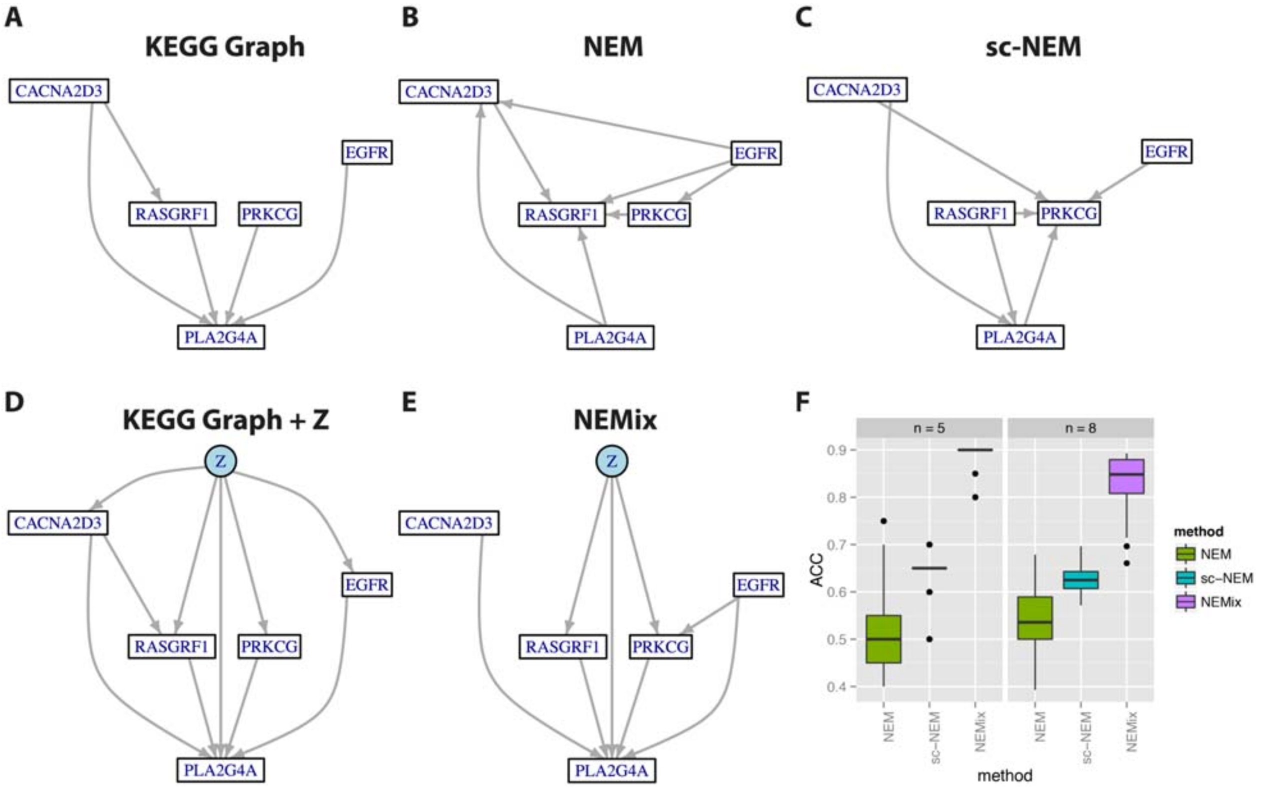

Fig 3. Inferred MAPK networks on HRV infection data. Best networks of the 5 top scoring siRNAs from the MAPK pathway for HRV infection for the different compared methods are displayed. (A) shows the known KEGG pathway. (B) is the inferred NEM and (C) the sc-NEM. (D) left shows the known network with the most likely attachment of the hidden variable *Z* (blue) and (E) is the inferred NEMix. For all networks their performance is summarized in Table 1. Subfigure (F) summarizes robustness of the MAPK network inference. For the inferred MAPK signaling networks on the HRV infection data, we assessed robustness of the accuracy for edge recovery. Box-plots display the result of 50 bootstrap samples for the three compared methods, on the 5 gene ($n = 5$) and 8 gene ($n = 8$) network.

doi:10.1371/journal.pcbi.1004078.g003

connections of Z with maximal out-degree 2 to the empty graph Φ. Like in the simulation study, p_0 was initialized by drawing randomly from a uniform distribution. Again we used uniform priors for θ and imposed no priors for the signaling networks other than the maximal out-degree of Z (plus the transitive edges that need to be added).

Table 1. Performance summary of the 5 gene MAPK network.

Network	Likelihood	ACC	AUC	p0	Sub-figure
True Graph	2641.47	1	1		A
NEM	2809.83	0.5	0.23		B
sc-NEM	29410.81	0.65	0.47		C
True Graph + Z	31768.94	1	1	0.48	D
NEMix	34982.87	0.9	0.84	0.42	E

The first column gives the log-likelihood for each model, showing that the true network is much less likely than the inferred networks. The second and third column show performance of the networks in terms of accuracy (ACC) and area under curve (AUC). The inferred p_0 for the NEMix models is displayed in column four. Column five indicates the corresponding sub-figure of Fig. 3. The network 'KEGG Graph + Z' denotes the structure of the known KEGG network, where only the position of *Z*, p_0, and θ are inferred.

doi:10.1371/journal.pcbi.1004078.t001

The known KEGG network and the inferred results for the top 5 signaling genes are displayed in Fig. 3A-E. Results for the top-8 gene network are given in S10 Fig. To assess robustness of the learned networks, we repeated the inference on 50 bootstrap samples of the original data set. Both networks show high AUC values and even better accuracy (see Table 1). As can be seen from Fig. 3F, network inference was very robust for the top-5 gene network. For the top-8 gene network, performance had a slightly higher variation. Individual plots for sensitivity and specificity are given in supplementary S11 Fig. A, B. Also the estimate of p_0 shows only little variation (S11 Fig. C). In all cases, the likelihood score of the known KEGG network is much lower than for the best inferred networks, indicating that under the assumptions of our model, the data and the KEGG database do not perfectly agree. Possible reasons for this observation include our model missing to explain part of the data correctly, the KEGG database being incomplete, and inaccuracies in the data generating process. Nevertheless, the accuracy value of 0.85 for the learned NEMix outperforms all other methods. All edges contained in the learned NEMix models are of high robustness ($> 80\%$ for 5 genes, and $> 70\%$ for 8 genes). Consensus networks of the bootstrap results are shown in supplementary S12 Fig.

Furthermore, the hidden root Z is attached to the same nodes in both the known KEGG graph and the estimated network for 5 genes. Also the inferred 8 node network connects Z to the same three genes. As genes were selected based on small off-target effects of their targeting siRNAs, they are not necessarily hits for HRV infection. However, of the selected genes EGFR [33], TAB2 [34] and CACNA2D3 [35] have been shown to be involved in this process.

All models have a built-in filter for uninformative features, which has been previously introduced in [36]. A comparison shows that averaged over the bootstrap samples, for all three methods, the set of used features largely agrees (supplementary S13 Fig and S14 Fig). The maximum likelihood attachments of features to the knock-down genes and the null node are shown in supplementary S15 Fig and S16 Fig, together with a detailed description of the different feature types. The inferred signaling disruption of $p_0 = 0.42$ seems rather high. We compared this to the average infection rate in mock experiments, i.e., cells without siRNA knock-down. These resemble cases, where Z can be perturbed but none of the other signaling genes in the network. Mock wells from plates of the 8 genes used here, actually have a much higher percentage of uninfected cells, roughly in the range of 75 to 81%. However, this comparison should be taken with caution since control wells of these screens might have suffered from strong plate location bias, as they were located on the margins of the plate.

As a general observation, NEMix-inferred networks were sparser than those obtained from NEMs, because spurious edges introduced in the latter are correctly explained by hidden pathway activity Z in NEMix. Therefore, NEMix networks have increased specificity, which might come at the cost of some missing true edges. Especially the 8-gene networks inferred by NEM and sc-NEM are much denser than the known KEGG network. A sparse network is beneficial in the sense that it allows to focus on a small set of highly specific edges. For validation experiments, it is desirable to have a low false positive rate in the predicted interactions as usually only very few of these dependencies can be experimentally tested.

Discussion

RNAi screens are known to be prone to many sources of noise and bias such that their analysis is highly challenging. Here, we have identified one confounding factor, namely heterogeneous signaling pathway activation within a cell population, and incorporated it directly into a novel probabilistic model for pathway reconstruction. To address the problem of unknown activation of signaling pathways during network inference, we have introduced a general framework, building on NEMs, to handle hidden combinatorial knock-downs in a probabilistic manner.

With NEMix we provide an implementation for inference under unknown pathway stimulation. For the first time, image features are explicitly used on the single-cell level for NEM inference, acknowledging large cell-to-cell variation. We have demonstrated the advantages of NEMix over current NEMs in simulations and inferred highly accurate networks in a case study on HRV infection. Especially, when the underlying true signaling networks are expected to be sparse, NEMix is beneficial. It removes spurious edges introduced due to confounding factors and therefore reduces the false positive rate, a desired property when it comes to validation of edges.

A limitation of the current model formulation is the assumption of independent single cell observations. In reality, this assumption might not be met as cells can be biased due to their location and neighbors. Removing this bias either by normalization or explicit modeling, as for example in [14], could further improve the model. Furthermore, in the current data sets cells can be in different cell cycle states. Grouping them according their states may remove further biases, but this clustering task is itself very challenging.

Another general limitation of NEMs and NEMix models is that they cannot learn certain pathway features. From static data, NEMs cannot resolve any loop structures by construction. This is a general problem for network inference without time resolved data. Therefore, only performance statements based on comparing transitively closed pathways can be made. The sampled graphs in the simulation are already transitively closed and since the transitive closure is a feature inherent to all the models we compare, it should not influence the ranking based on their performance. Before comparing a network to the corresponding KEGG pathway, we also built its transitive closure. This fact should be considered when interpreting the inferred models. For example, the model does not allow for distinguishing a feed forward loop from a sequential cascade; however, the hierarchical order of genes in the network would remain the same, and this piece of information does already provide considerable insight into the biological processes. The way we have assessed performance here puts particular emphasis on this hierarchical structure of the network nodes.

Further improvements could be achieved during data preparation. Image segmentation is not always perfect and might introduce technical biases into data sets, adding more confounding factors. If data is not curated carefully, we risk to capture technical biases with the additional hidden variable in NEMix models. Another interesting aspect of the data sets deserving a more thorough analysis, is the nature of the image features themselves. Here, readouts have been used to infer the graph of signaling genes. However, one could investigate in more detail how features are grouped when attaching them to the signaling genes. Some features might not contribute useful information and could be filtered in advance, others might be redundant. Future projects could use the output of NEMix models and seek for biological interpretation of feature correlations.

In case of cell infection screens, infection efficacy was an obvious factor that needed to be addressed. However, the same idea could be applied to other sources of noise. For example, transfection efficacy of the knock-downs could be considered. Quality and efficacy of a knock-down can be quantified by mRNA levels (qPCR) or protein level (western blot analysis) of a gene. However, for high-throughput assays, such confirmation is not available for most gene knock-downs. In order to account for different siRNA transfection efficacies further hidden variables could be introduced. In contrast to the global Z variable introduced here, hidden knock-down rates would then be estimated for each gene individually. As a consequence, the complexity of the problem would increase substantially. Instead of one parameter, n (number of genes) parameters would have to be estimated. Furthermore, knock-down probabilities could only be estimated from a fraction of the observations (e.g., cells under the specific knock-down). Another drawback is that the increased number of hidden variables gives rise to identifiability problems when estimating infection efficacy in combination with the knock-down rates. For example, if the hidden variable Z was only attached to one signaling gene, effects of Z and a failed transfection could not be distinguished. Although extending the NEMix model to this situation would be an interesting future project, we believe that problems in the transfection process play an overall minor role. For the current experiments, KIF11 siRNAs (cell killers) were used to control transfection quality on the plate level. For the plates containing the cells used in our analysis, these controls show very high penetrance, i.e., out of an average of 2000 cells per well, on average only 7% of cells survive in these wells. Although this test does not make a statement about the efficacy of individual siRNAs, it ensures the general functioning of the transfection process. Additionally, the library vendor claims the knock-down efficacies achieved with their smart-pool siRNAs to be in the range of 70–95%. This proportion is a result of many possible sources of imperfect gene silencing, including non-transfected cells and off-target effects. Given the above facts in combination with our off-target filtering strategy, we are convinced that the analyzed data are of high quality.

We tried to minimize the general problem of confounding siRNA off-targets by considering only genes targeted by siRNAs with low predicted off-target effects. This selection step helps to achieve reasonably unbiased results with our model, but it also limits the gene sets we can analyze. Ideally, we want to be able to select any gene of interest. This scenario calls for models that can correct the off-target effects on the single-cell level. A potential solution to this issue could be delivered by NEMs directly. We could still learn the networks based on siRNA knock-downs directly, but handle the signal propagation differently. With NEMix it is already possible to use each siRNA as a combinatorial knock-down. In reality however, individual genes are knocked-down to different degrees by an siRNA. In a NEM, this would mean to split up the silencing signal of an siRNA into partial knock-downs of several genes. Then, signal propagation would have to be formulated in a fully probabilistic fashion and NEMs would have to be reformulated such that their nodes do not have binary states anymore. Further developing NEMix, by integrating the above mentioned shortcomings, will make the models more powerful for future network reconstruction tasks.

Especially in the light of single cell data sets, which show large heterogeneity among individual observations, our approach is beneficial. Such data sets are becoming more and more available, and they reveal that the high cell-to-cell variation has severe consequences when summarizing such heterogeneous observations. On the population level, the signal is potentially confounded as it is only contained in part of the observations. NEMix uses the full power of single-cell experiments, as it is applied on the single-cell level directly, avoiding any data averaging. Only at this data resolution, the heterogeneity within a cell population can be accounted for and it becomes possible to investigate potentially confounding factors, such as, for example, pathway activity. NEMix is the first NEM-based method with additional unknown components in the signaling graph Φ. It is capable of inferring these missing data and provides an estimate for the fraction of signal disruption. We find such ambiguous signaling in RNAi infection screens and we have demonstrated that NEMix can improve network inference substantially by accounting for the confounding factor.

Models

The NEM framework

A NEM, as introduced in [2], aims to infer the hidden dependency structure among a set of n binary signaling variables $\mathcal{S}$ from the nested structure of m observed effect variables $\mathcal{E}$ (features). It therefore consists of two directed graphs, one describing the dependencies among the signaling genes and one connecting the features to the genes.

The binary adjacency matrix of signaling genes is denoted $\Phi = (\phi_{ks})$, with $\phi_{ks} = 1$ if gene k propagates its effects to gene s and using the convention $\Phi_{k,\, k} = 1$, for all k. The signaling graph Φ is thus always transitively closed. If a gene is silenced, the effect is propagated deterministically along the edges of Φ. The connection of features $\mathcal{E}$ to the genes $\mathcal{S}$ is given by parameters θ_e, where $\theta_e = s$ indicates that feature e is linked to gene s. For a gene k and a feature e, a NEM predicts an effect of k on e if there is a gene s such that $\phi_{ks} = 1$ (i.e., k and s are connected), and $\theta_e = s$ (i.e., s has an effect on e). The observed data are denoted $D = (d_{ek})$, where each d_{ek} is the measurement of feature e under perturbation of k (Fig. 1A).

Given an external signal which affects one or more of the signaling genes, each of them will have a binary signaling state. The state value is 0 if the signaling is interrupted, i.e., does not reach the node, and 1 if the signal reaches the node, i.e., the natural state of a stimulated pathway.

For inferring the structure Φ among the signaling genes, we consider its posterior

$$P(\Phi \mid D) = \frac{P(D \mid \Phi)P(\Phi)}{P(D)}, \tag{3}$$

where the marginal likelihood $P(D|\Phi)$ can be obtained by integrating out the connections of features to the genes,

$$P(D \mid \Phi) = \int_{\theta} P(D \mid \Phi, \theta)P(\theta \mid \Phi)d\theta, \tag{4}$$

with prior distribution $P(\theta|\Phi)$. In the absence of further knowledge, the prior is usually set to the uniform distribution. Given the network structure and assuming conditional independence of the parameters θ_e and of the silencing experiments k, the marginal likelihood becomes

$$P(D \mid \Phi) = \prod_{e=1}^{m}\sum_{s=1}^{n}\left[\prod_{k=1}^{K}P(d_{ek} \mid \Phi, \theta_e = s)\right]P(\theta_e = s). \tag{5}$$

The local effect likelihoods $P(d_{ek}|\Phi, \theta)$ denote the probability of observing an effect in feature e under knock-down of gene k. They can usually be pre-computed from the data and different approaches have been proposed [2, 23, 36]. For the results presented below, log-odds ratios as introduced in [36] were used (see 'Modeling the effect likelihoods' in S1 Text for details).

The NEMix model

We first define the NEMix model and then derive it in detail. A NEMix consists of a nested effects model with effects graph Θ and an extended signaling graph Φ. The signaling graph Φ describes the dependency structure among the signaling genes and has an additional binary hidden variable Z indicating pathway activity. Z is a root of Φ, i.e., it can be connected to any of its nodes and does not have any direct connections to features in θ. The silencing probability of Z is denoted by p_0 and is a priory not known. For a set knock-down experiments $k \in \{1, \ldots, K\}$, with single cell observations $c \in \{1, \ldots, c_k\}$ of signaling gens $s \in \{1, \ldots, n\}$ and features $e \in \{1, \ldots, m\}$, the marginal likelihood of a NEMix is

$$P(D \mid \Phi) = \prod_{e=1}^{m}\sum_{s=1}^{n}P(\theta_e = s)\prod_{k=1}^{K}\prod_{c=1}^{c_k}\sum_{j\in\{0,1\}}p_jP(d_{ekc} \mid \Phi, \theta_e = s, Z_{kc} = j), \tag{6}$$

where $p_j = P(Z_k = j)$.

Probabilistic combinatorial knock-downs. We first extend the model to cope with several silenced genes at the same time. To achieve this goal, we condition the perturbation state of each gene on the states of its parents. For each knock-down experiment $k \in \{1, \ldots, K\}$, let $\mathcal{S}_k \subseteq \mathcal{S}$ be the set of genes knocked down at the same time in experiment k. We assume for combinatorial silencing events that we observe an effect on feature e if it can be reached by either of the genes in $\mathcal{S}_k$ through a path in Φ. Let furthermore S_{sk} be the hidden binary random variable for the silencing state of gene s under knock-down of $\mathcal{S}_k$. Then $S_{sk} = 0$ if the gene is perturbed and $S_{sk} = 1$ if it is not. If $s \in \mathcal{S}_k$, then S_{sk} is set to zero but otherwise its value depends on the states of the parents of s, $S_{pa(s)k} \in \{0, 1\}^{|pa(s)|}$, through the conditional probability $P(S_{sk}|S_{pa(s)k})$.

The local effect likelihoods are then given by the marginalization over S_{sk},

$$P(d_{ek} \mid \Phi, \theta_e = s) = \sum_{x \in \{0,1\}} P(d_{ek} \mid \theta_e = s, S_{sk} = x) P(S_{sk} = x), \tag{7}$$

where $P(S_{sk})$ is the probability of gene s being active in experiment k. If the state S_{sk} depends on the states of the parent nodes, one can deduce the marginal $P(S_{sk})$ from the joint distribution $P(S) = P(S_{1k}, \ldots, S_{nk})$ of the signaling graph Φ. The joint probability factorizes when conditioning on the parent nodes,

$$P(S_{sk}) = \sum_{S_{1,k}, \ldots, S_{s-1,k}, \, S_{s+1,k}, \ldots, S_{n,k}} P(S_{1k}, \ldots, S_{nk}) \tag{8}$$

However, for all signaling genes in $\mathcal{S}_k$, we know that their state is 0, independent of their parents. So in fact, we only need to sum over all genes $\mathcal{S}_{a,\,k} = \{S_{1,\,k}, \ldots, S_{s-1,\,k}, S_{s+1,\,k}, \ldots, S_{n,\,k}\}$ that are not in the set of knock-down genes $\mathcal{S}_k$, except for gene s itself:

$$P(S_{sk}) = \sum_{\mathcal{S}_a} \prod_{i \notin \mathcal{S}_k} P(S_{ik} \mid S_{pa(i)k}) \underbrace{\prod\nolimits_{i \in \mathcal{S}_k} P(S_{ik} = 0)}_{=1} . \tag{9}$$

If $s \in \mathcal{S}_k$, then $P(S_{sk} = 0) = 1$. Substituting (7) and (9) into the marginal likelihood (5) leads to

$$P(D \mid \Phi) = \prod_{e=1}^{m} \sum_{s=1}^{n} P(\theta_e = s) \prod_{k=1}^{K} \sum_{x \in \{0,1\}} P(d_{ek} \mid \theta_e = s, S_{sk} = x) \sum_{\mathcal{S}_{a,k}} \prod_{i \notin \mathcal{S}_k} P(S_{ik} \mid S_{pa(i)k}). \tag{10}$$

The conditional local effect likelihoods $P(d_{ek}|\theta_e = s, S_{sk} = x)$ can usually be pre-computed (see 'Modeling the effect likelihoods' in S1 Text).

Deterministic combinatorial knock-downs. If we assume deterministic signaling and all knock-downs are known, then $P(S_{sk})$ will either be 0 or 1. As mentioned above, we assume a gene to be perturbed if at least one of its parents is perturbed, and unperturbed if none of the parents are perturbed. For transitively closed NEMs this also means $S_{sk} = 0$ if and only if $\mathcal{S}_k \cap pa(s) \neq \emptyset$, i.e., only if the parents of s contain one of the knocked-down genes, then s itself can be perturbed. Therefore, the conditional probabilities $P(S_{sk}|S_{pa(s)k})$ are in this case

$$P(S_{sk} = 1 \mid S_{pa(s)k}) = \begin{cases} 1 & \text{if } (S_{pa(s)k} = (1, 1, \ldots, 1) \text{ or } pa(s) = \emptyset) \text{ and } s \notin \mathcal{S}_k \\ 0 & \text{otherwise} \end{cases}$$

$$P(S_{sk} = 0 \mid S_{pa(s)k}) = 1 - P(S_{sk} = 1 \mid S_{pa(s)k}). \tag{11}$$

Since these probabilities are either 0 or 1, we use the following indicator function

$$\delta_{sk} := P(S_{sk} = 1) = P(S_{sk} = 1 \mid S_{pa(s)k}). \tag{12}$$

The last equation holds, because for

$$P(S_{sk}) \quad = \sum_{S_{pa(s)k}} P(S_{sk}, S_{pa(s)k}) = \sum_{S_{pa(s)k}} P(S_{sk} \mid S_{pa(s)k}) P(S_{pa(s)k}) \tag{13}$$

all terms $P(S_{pa(s)k})$ except for one parent configuration are zero. The local effect likelihoods can then be written as

$$\begin{aligned} P(d_{ek} \mid \Phi, \theta_e = s) &= \sum_{x \in \{0,1\}} P(d_{ek} \mid \theta_e = s, S_{sk} = x) \delta_{sk} \\ &= \begin{cases} P(d_{ek} \mid \theta_e = s, S_{sk} = 0) & \text{if} \ \ \delta_{sk} = 1 \\ P(d_{ek} \mid \theta_e = s, S_{sk} = 1) & \text{if} \ \ \delta_{sk} = 1, \end{cases} \end{aligned} \tag{14}$$

which resembles the situation introduced in [6] for dynamic NEMs. However, in the following we make use of the more general case.

Hidden pathway stimulation. We now turn to a special case, where exactly one (root) node of the network has probabilistic signaling and the others follow the deterministic rules above. Silencing experiments can be noisy for many different reasons and it might be unknown whether the signaling pathway of interest is actually activated during knock-down of a gene. To model this uncertainty, we consider an additional hidden binary random variable Z_k, indicating the state of an external signal that activates the pathway, where $Z_k = 1$ means active and $Z_k = 0$ means inactive in experiment k. The random variable Z can be viewed as an additional node in Φ that has only outgoing edges and can not have any observables directly attached to it (see Fig. 1B for an example). Let furthermore $p_0 = P(Z_k = 0)$ be the probability that the signaling pathway has not been activated, and $p_1 = P(Z_k = 1) = 1 - p_0$ the probability that it is active. The silencing of genes $\mathcal{S}_k$ together with a unknown pathway stimulation can then be regarded as a hidden combinatorial knock-down event, where signaling genes $\mathcal{S}_k$ are silenced deterministically and the external signal Z_k is inactivated with probability p_0. Fig. 1B illustrates a simple NEMix with the additional pathway state variable. Since Z has no parents, we can easily factorize p_j out of the joint probability of the states $P(S)$ in (9) to obtain

$$P(S_{sk}) = \sum_{j \in \{0,1\}} p_j \cdot P(S_{sk} \mid Z_k = j) = \sum_{j \in \{0,1\}} p_j \cdot \delta^j_{sk}, \tag{15}$$

where $\delta^j_{sk} = P(S_{sk} \mid Z_k = j)$ is again an indicator function for the state of s in experiment k, given that the pathway is in state j. Substituting this expression into the local effect likelihoods (7) leads to

$$\begin{aligned} P(d_{ek} \mid \Phi, \theta_e = s) &= \sum_{S_{sk} \in \{0,1\}} P(d_{ek} \mid \Phi, \theta_e = s, S_{sk}) \textstyle\sum_{j \in \{0,1\}} p_j \cdot \delta^j_{sk} \\ &= \sum_{j \in \{0,1\}} p_j \cdot P(d_{ek} \mid \Phi, \theta_e = s, Z_k = j), \end{aligned} \tag{16}$$

and the marginal likelihood (10) becomes

$$P(D \mid \Phi) = \prod_{e=1}^{m} \sum_{s=1}^{n} P(\theta_e = s) \prod_{k=1}^{K} \sum_{j \in \{0,1\}} p_j P(d_{ek} \mid \Phi, \theta_e = s, Z_k = j). \tag{17}$$

For the RNAi infection experiments described in the introduction and the results, pathways of interest are those involved in infection signaling, i.e., pathways which are activated upon signals triggered by a pathogen. However, infection of a cell is a stochastic event, depending on many factors, for example, whether at all a pathogen docked on successfully to the cell. Consequently, in a cell with a silenced gene, there can be several explanations for why it stayed uninfected. It could be because the knocked-down gene was important for the infection signaling, but there is also a chance that other factors account for this, for example, no pathogen came within reach of the cell. In case a pathogen triggered a signal, the pathway is considered active ($Z_k = 1$), corresponding to a normal NEM. When no infection attempt was made, the infection pathway is inactive ($Z_k = 0$).

The population of cells in knock-down experiment k can be divided into infected and uninfected cells. For infected cells, the external input signal from the pathogen reached the cell and the signaling pathway is active ($Z_k = 1$). In these cases Z is observed. For uninfected cells, however the state of Z_k is unknown and no longer deterministic. So, for an infected cell, we have $P(Z_k = 0) = 0$ and $P(Z_k = 1) = 1$, whereas for an uninfected cell, we have $P(Z_k = 0) = p_0$ and $P(Z_k = 1) = p_1$. Here, p_0 is the probability that the signaling pathway has not been activated by the pathogen and $p_0 + p_1 = 1$. This is exactly the above situation of a hidden combinatorial knock-down and the additional model parameter $p_j, j \in \{0, 1\}$ either needs to be estimated for each observation, or it has to be integrated out.

Learning from single-cell data. As illustrated in the infection experiment example above, in general, the signaling state Z_k can be different for each individual cell c in an experiment k, and therefore we have to treat each cell as an individual observation. Regarding single cells as independent, for a given network structure Φ, the local likelihoods further decompose into

$$P(d_{ek} \mid \Phi, \theta_e = s) = \prod_{c=1}^{c_k} P(d_{ekc} \mid \Phi, \theta_e = s), \tag{18}$$

where c_k is the number of cells in experiment k. Instead of a single number, now $S_{sk} = (S_{skc})_{c = 1, \ldots, c_k}$ is a vector where each S_{skc} is the state of gene s in cell c under knock-down k. With this modification, the marginal likelihood expands to

$$P(D \mid \Phi) = \prod_{e=1}^{m} \sum_{s=1}^{n} P(\theta_e = s) \prod_{k=1}^{K} \prod_{c=1}^{c_k} \sum_{j \in \{0,1\}} p_j P(d_{ekc} \mid \Phi, \theta_e = s, Z_{kc} = j). \tag{19}$$

In other words, for each cell, there will be an effect of the perturbation set $\mathcal{S}_k$ on feature e if any of the perturbations reach gene s and e is connected to s.

Model identifiability. As shown in [36], NEMs have unidentifiable components. If two nodes share the same set of parents, then these two nodes are indistinguishable. Furthermore, NEMs are unique only up to reversals, i.e., different parametrizations can exist for the same model that explain the data equally well. Such equivalent representations are related by cyclic node permutations, $\Phi\Theta = \Phi'\Theta'$ with $(\Phi', \Theta') = (\Phi S^{-1}, S\Theta)$ and permutation matrix S reversing cycles in Φ. This result still holds when adding the additional hidden pathway state Z.

Regarding inference of the new parameter p_0, there are only few situations in which p cannot be learned. For a NEMix model with given graphs Φ and Θ, the pathway inactivation probability p_0 of its hidden pathway activity Z is not identifiable, if and only if either (1) there are no observables from $\mathcal{E}$ attached downstream of Z, or (2) Z is connected only to sub-components of a network that are always perturbed (for a proof, see ‘Unidentifiable parameters’ in S1 Text). Both conditions are rather artificial cases. The first one describes the situation where some signaling genes do not have any features attached and Z is connected only to these. In this case, there are no observations from which p_0 can be estimated. However, usually we assume a uniform attachment of features to the genes and models containing genes without any downstream features are hardly ever observed. The interpretation of the second condition is that only genes which receive a propagated signal from all other genes in the network are affected by the pathway deactivation. Here, p_0 cannot be estimated because all observations downstream of Z will show an effect, independent of the state of Z. Again, it appears rather exceptional that only the final node of a signaling cascade is affected by a pathway deactivation.

NEMix inference

Structure learning is performed using a greedy heuristic to find an optimal network. Similar to the NEM procedure described in [3], edges are incrementally added if the likelihood is increased (see ‘Structure learning’ in S1 Text). In addition, our approach is restricted to structures without incoming edges into the hidden root Z. We initialize the algorithm with a set of initial networks. These consist of the empty graph and one edge connecting Z to one of the knock-down genes. Additionally, we limit the out-degree of Z to two. Here, by out-degree we mean only the non-transitive edges. We still allow the insertion of transitive edges from Z to any signaling gene, which has to be added in order to fulfill the transitivity requirement. This regularization reduces the search space and prevents that too many dependencies between genes are explained by Z alone.

As for classic NEMs, network structure scoring involves the marginal likelihood. For the NEMix model, $P(\mathcal{D}|\Phi)$ cannot be optimized analytically. Marginalization over the feature attachments is omitted in our extended model. Instead, we estimate θ jointly with p during model inference. To do so, we approximate the marginal likelihood (10) by the expectation of the complete data log-likelihood

$$P(D, Z \mid \Phi, \theta, p) = \prod_{k=1}^{K} \prod_{c=1}^{c_k} \prod_{j \in \{0,1\}} \left[p_j \prod_{e=1}^{m} P(d_{ekc} \mid \Phi, \theta_e = s, Z_{kc} = j) \right]^{Z_{kc}^{(j)}}, \tag{20}$$

with respect to Z, where θ and p_0 need to be efficiently estimated. For this task we have developed an EM algorithm. A derivation of the expected hidden log-likelihood and the maximum likelihood estimates is given in ‘Estimating the hidden signal’ of S1 Text. When starting the EM algorithm, p_0 is initialized with a random draw from the uniform distribution and for θ we use a uniform initial configuration.

Implementation

The NEMix model is included as part of the R/Bioconductor package NEM as an additional inference type. It is invoked by calling the package’s main function `NEM(data, inference = ‘NEM.greedy’, control)` and choosing the inference type `control$type = ‘NEMix’`. (See ‘NEMix implementation in NEM package’ in S1 Text for more detailed instructions on the implementation and usage of NEMix in R). To record run-times of NEMix model estimation, simulations were run without any parallelization on a 1.7GHz Intel i7 machine. Only one starting configuration was used, and EM iterations were performed using three restarts to avoid local optima that are globally suboptimal. For realistic data sets of 300 features and 200 cells per knock-down, NEMix estimation took on average nine minutes for 5-gene networks, with an average of 13 iteration steps until convergence of the EM algorithm. For the 8-gene network, the average run-time was 66 minutes, while the average number of iterations per EM round remained 13 also for these larger networks. The longer run-times of NEMix models as compared to NEMs are primarily due to the hidden data estimation. Each structure scored once in a NEM inference, needs to be scored 40 times on average during NEMix estimation. In addition, the input data sets are roughly 200 times larger.

Supporting Information

S1 Text. Supplementary texts. The supplementary text contains additional information regarding the NEMIX model, description and pre-processing of the data sets, as well as a short usage description of the NEMIX code in the R package nem.
(PDF)

S1 Fig. Sample networks. All 30 sample networks were randomly generated from the KEGG graph using a random walk along the edges. Unidentifiable structures were omitted. The blue node marks the randomly added hidden signal.
(EPS)

S2 Fig. Performance for varying p_0. For each of the 30 generated networks, 50 data set were drawn. In (A) the area under the ROC curve (AUC) was calculated based on edge frequencies of the samples. The right most panel displays the result for all values of p_0 jointly. Sub-figure (B) shows the area under the PR curve (AUPRC).
(EPS)

S3 Fig. Accuracy values per network. For each of the 30 generated networks, 50 data sets were drawn. Then, the accuracy (ACC) was calculated based on edge frequencies of the samples.
(EPS)

S4 Fig. Estimated values for p_0. For each of the 30 generated networks, 50 data sets were drawn. The distribution of the estimated p_0 per network is shown. Each row represents a different true signal disruption probability.
(EPS)

S5 Fig. Inferred pathway states Z. For each of the 30 generated networks, 50 data sets were drawn. Percentage of correctly inferred state values of Z for the sample data sets is shown for each of generated networks.
(EPS)

S6 Fig. Performance for larger network sizes. To assess the edge recovery performance for larger NEMix models, we ran a reduced simulation study. We sampled 30 random networks of network size $n = 5, 10,$ and 15 genes. The hidden variable Z was again attached randomly to at most 2 of the signaling genes (plus additional transitive edges). We fixed p_0 to 0.4, which is close to our application example. For each network, we then generated 30 data sets of 300 features from 200 cells per each knock-down. For run-time reasons we only initiated the structure search with the empty network and used just 2 restarts for the EM runs. This reduces performance of network learning but shows how the overall performance scales with growing network size. Even for larger networks performance is still very good as can be seen from the area under ROC curve (A), area under precision-recall curve (AUPRC; B) and accuracy (C). Estimation of p_0 becomes even more precise for larger n, as shown in (D) by the absolute distance of the sampled from the estimated p_0. Run-time on the other hand increases substantially for

larger networks. Panel (E) shows the run-times per network estimation in minutes.
(EPS)

S7 Fig. Performance of feature attachments θ. For each of the 30 generated networks, 50 data sets were drawn. In (A) the percentage of correctly inferred feature attachments are displayed and (B) shows the percentage of correctly filtered uninformative features. Both plots show results for different fractions of signal disruption p_0.
(EPS)

S8 Fig. Network wise performance of feature attachments θ. For each of the 30 generated networks, 50 data sets were drawn. In (A) the percentage of correctly inferred feature attachments are displayed and (B) shows the percentage of correctly filtered uninformative features, for each individual network.
(EPS)

S9 Fig. GSEA for reliable siRNAs. To see if KEGG pathways are affected differently by off-targeting siRNAs, we performed a gene set enrichment analysis [32] on the siRNA scores, using the implementation in the R package 'HTSanalyzR' [37].
(EPS)

S10 Fig. Inferred 8 gene MAPK networks on HRV infection data. Best networks of the 8 top scoring siRNAs from the MAPK pathway for HRV infection for the different compared methods are displayed. (A) shows the known KEGG pathway. (B) is the inferred NEM and (C) the sc-NEM. (D) left shows the known network with the most likely attachment of the hidden variable Z (blue) and (E) is the inferred NEMix. For all networks their performance is summarized in S1 Table.
(EPS)

S11 Fig. Performance of MAPK network inference. We computed the specificity (A) and sensitivity (B) for all compared methods, based on 50 bootstrap samples. Both plots show the results for 5 and 8 signaling genes with top scoring siRNAs, using the HRV infection data. Subfigure (C) shows robustness of inferred pathway activity. The estimated pathway activity for 5 and 8 gene networks, derived from the 50 bootstrap samples is shown. p_0 shows little variation and is similar for both networks.
(EPS)

S12 Fig. Consensus networks for MAPK pathway. Consensus networks for 5 genes (A) and 8 genes (B) are displays. Shown are all edges with frequency of at least 0.7 in the 50 bootstrap samples. NEMix inference was run using the 16 different starting configurations. For each bootstrap sample the best solution was chosen.
(EPS)

S13 Fig. Feature attachments for the 5 MAPK pathway genes. Histograms show the selection frequencies for image features from 50 bootstrap samples on the HRV infection data.
(EPS)

S14 Fig. Shared feature usage for compared methods. The Venn diagrams compare frequently attached features (left) and never attached features (right) based on the 50 bootstrap samples. Results for the 5 gene network are shown in the top row and for the 8 gene network in the bottom row.
(EPS)

S15 Fig. Uninformative features for the 5-gene network. Each row shows the probability for one feature of being attached to each of the genes in the network or the null node. For all features in this plot, the null node had the highest probability, which means, they are filtered out. Features are colored by the channel they were measured from. These channels are, fluorescence of DNA in the nucleus (blue), fluorescence of actin (red), fluorescence of cell internal pathogens (green). Furthermore there are general location and orientation features (black). The measurements themselves then give information on intensity, shape, texture or neighbors of the objects segmented from the images. These objects are 'Cells': the cell body, 'Nuclei': the cell nuclei, 'PeriNuclei': a peripheral area around the nucleus, 'VoronoiCells': the area of the cell from a Voronoi-tessellation of the image. Many of the uninformative features are related to orientation of objects or their location, which are expected not to carry useful information for the network inference.
(EPS)

S16 Fig. Feature attachments in the 5-gene network. Each row shows the probability for one feature of being attached to each of the genes in the network or the null node. Rows are sorted by the gene for which the attachment probability is highest. For a description of the different feature types see caption of supplementary S15 Fig.
(EPS)

S17 Fig. Illustration of a NEMix, with unidentifiable $p = P(Z)$. The hidden variable is attached only to signaling genes that are always perturbed. For models with such structure p cannot be inferred.
(EPS)

S1 Table. Performance summary of the 8 gene MAPK network. The first column gives the log-likelihood for each model, showing that the true network is much less likely than the inferred networks. The second and third column show performance of the networks in terms of accuracy (ACC) and area under curve (AUC). The inferred p_0 for the NEMix models is displayed in column four. Column five indicates the corresponding sub-figure of Fig. 3. The network 'KEGG Graph + Z' denotes the structure of the known KEGG network, where only the position of Z, p_0, and θ are inferred.
(PDF)

Acknowledgments

We thank all members of the InfectX and TargetInfectX consortium for stimulating discussions that helped making this work possible. Furthermore, this work was supported by Sybit (SystemsX) in terms of data management and storage. Finally, we thank Fabian Schmich for contributing the siRNA target prediction matrix.

Author Contributions

Conceived and designed the experiments: JSP CD UG HF NB. Performed the experiments: DM UG. Analyzed the data: JSP ME PR. Contributed reagents/materials/analysis tools: DM ME PR CD UG HF. Wrote the paper: JS HF NB. Designed and implemented statistical model: JSP HF NB.

References

1. Markowetz F, Spang R (2007) Inferring cellular networks—a review. BMC Bioinformatics 8: S5. doi: 10.1186/1471-2105-8-S6-S5

2. Markowetz F, Bloch J, Spang R (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics 21: 4026. doi: 10.1093/bioinformatics/bti662 PMID: 16159925

3. Fröhlich H, Tresch A, Beissbarth T (2009) Nested effects models for learning signaling networks from perturbation data. Biometrical J 51: 304–323. doi: 10.1002/bimj.200800185

4. Niederberger T, Etzold S, Lidschreiber M, Maier KC, Martin DE, et al. (2012) MC EMiNEM Maps the Interaction Landscape of the Mediator. PLoS Comput Biol 8: 10. doi: 10.1371/journal.pcbi.1002568

5. Failmezger H, Praveen P, Tresch A, Fröhlich H (2013) Learning gene network structure from time laps cell imaging in RNAi Knock downs. Bioinformatics 29: 1534–1540. doi: 10.1093/bioinformatics/btt179 PMID: 23595660

6. Fröhlich H, Praveen P, Tresch A (2011) Fast and efficient dynamic nested effects models. Bioinformatics 27: 238–44. doi: 10.1093/bioinformatics/btq631 PMID: 21068003

7. Anchang B, Sadeh MJ, Jacob J, Tresch A, Vlad MO, et al. (2009) Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. Proc Natl Acad Sci 106: 6447. doi: 10.1073/pnas.0809822106 PMID: 19329492

8. Zeller C, Fröhlich H, Tresch A (2009) A Bayesian network view on nested effects models. EURASIP J Bioinform Syst Biol 2009: 8. doi: 10.1155/2009/195272

9. Hannon GJ (2002) RNA interference. Nature 418: 244–251. doi: 10.1038/418244a PMID: 12110901

10. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411: 494–8. doi: 10.1038/35078107 PMID: 11373684

11. Schultz N, Marenstein DR, De Angelis Da, Wang WQ, Nelander S, et al. (2011) Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-β pathway and reveal microRNA regulation of TGFBR2. Silence 2: 3. doi: 10.1186/1758-907X-2-3 PMID: 21401928

12. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, et al. (2009) Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 6: 569–575. doi: 10.1038/nmeth.1351 PMID: 19644458

13. Snijder B, Sacher R, Rämö P, Liberali P, Mench K, et al. (2012) Single-cell analysis of population context advances RNAi screening at multiple levels. Mol Syst Biol 8: 579. doi: 10.1038/msb.2012.9 PMID: 22531119

14. Knapp B, Rebhan I, Kumar A, Matula P, Kiani Na, et al. (2011) Normalizing for individual cell population context in the analysis of high-content cellular screens. BMC Bioinformatics 12: 485. doi: 10.1186/1471-2105-12-485 PMID: 22185194

15. Jurgeit A, Moese S, Roulin P, Dorsch A, Lötzerich M, et al. (2010) An RNA replication-center assay for high content image-based quantifications of human rhinovirus and coxsackievirus infections. Virol J 7: 264. doi: 10.1186/1743-422X-7-264 PMID: 20937137

16. Jurgeit A, McDowell R, Moese S, Meldrum E, Schwendener R, et al. (2012) Niclosamide is a proton carrier and targets acidic endosomes with broad antiviral effects. PLoS Pathog 8: e1002976. doi: 10.1371/journal.ppat.1002976 PMID: 23133371

17. Rämö P, Drewek A, Arrieumerlou C, Beerenwinkel N, Ben-Tekaya H, et al. (2014) Simultaneous analysis of large-scale RNAi screens for pathogen entry. BMC Genomics in press.

18. Conrad C, Gerlich DW (2010) Automated microscopy for high-content RNAi screening. J Cell Biol 188: 453–61. doi: 10.1083/jcb.200910105 PMID: 20176920

19. Mohr S, Bakal C, Perrimon N (2010) Genomic screening with RNAi: results and challenges. Annu Rev Biochem 79: 37–64. doi: 10.1146/annurev-biochem-060408-092949 PMID: 20367032

20. Mohr SE, Perrimon N (2011) RNAi screening: new approaches, understandings, and organisms. Wiley Interdiscip Rev RNA 3: 145–58. doi: 10.1002/wrna.110 PMID: 21953743

21. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Statistical Methodol 39: 1–38.

22. Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173

23. Fröhlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T (2007) Large scale statistical inference of signaling pathways from RNAi and microarray data. BMC Bioinformatics 8: 386. doi: 10.1186/1471-2105-8-386

24. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7: R100. doi: 10.1186/gb-2006-7-10-r100 PMID: 17076895

25. Seinen E, Burgerhof JGM, Jansen RC, Sibon OCM (2010) RNAi experiments in D. melanogaster: solutions to the overlooked problem of off-targets shared by independent dsRNAs. PLoS One 5: 1–7. doi: 10.1371/journal.pone.0013119

26. Seger R, Krebs E (1995) The MAPK signaling cascade. FASEB J 9: 726–735. PMID: 7601337

27. Murphy LO, Blenis J (2006) MAPK signal specificity: the right place at the right time. Trends Biochem Sci 31: 268–75. doi: 10.1016/j.tibs.2006.03.009 PMID: 16603362

28. Dhillon aS, Hagan S, Rath O, Kolch W (2007) MAP kinase signalling pathways in cancer. Oncogene 26: 3279–90. doi: 10.1038/sj.onc.1210421 PMID: 17496922

29. Wang X, Lau C, Wiehler S, Pow A, Mazzulli T, et al. (2006) Syk Is Downstream of Intercellular Adhesion Molecule-1 and Mediates Human Rhinovirus Activation of p38 MAPK in Airway Epithelial Cells. J Immunol 177: 6859–6870. doi: 10.4049/jimmunol.177.10.6859 PMID: 17082600

30. Hall DJ, Bates ME, Guar L, Cronan M, Korpi N, et al. (2005) The Role of p38 MAPK in Rhinovirus-Induced Monocyte Chemoattractant Protein-1 Production by Monocytic-Lineage Cells. J Immunol 174: 8056–8063. doi: 10.4049/jimmunol.174.12.8056 PMID: 15944313

31. Laza-Stanca V, Stanciu LA, Message SD, Edwards MR, Gern JE, et al. (2006) Rhinovirus replication in human macrophages induces NF-kappaB-dependent tumor necrosis factor alpha production. J Virol 80: 8248–58. doi: 10.1128/JVI.00162-06 PMID: 16873280

32. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–50. doi: 10.1073/pnas.0506580102 PMID: 16199517

33. Zhu L, Lee PK, Lee WM, Zhao Y, Yu D, et al. (2009) Rhinovirus-induced major airway mucin production involves a novel TLR3-EGFR-dependent pathway. Am J Respir Cell Mol Biol 40: 610–9. doi: 10.1165/rcmb.2008-0223OC PMID: 18978302

34. Schuler BA, Schreiber MT, Li L, Mokry M, Kingdon ML, et al. (2014) Major and minor group rhinoviruses elicit differential signaling and cytokine responses as a function of receptor-mediated signal transduction. PLoS One 9: e93897. doi: 10.1371/journal.pone.0093897 PMID: 24736642

35. Triantafilou K, Kar S, van Kuppeveld FJM, Triantafilou M (2013) Rhinovirus-Induced Calcium Flux Triggers NLRP3 and NLRC5 Activation in Bronchial Cells. Am J Respir Cell Mol Biol. doi: 10.1165/rcmb.2013-0032OC PMID: 23815151

36. Tresch A, Markowetz F (2008) Structure learning in nested effects models. Stat Appl Genet Mol Biol 7: 26.

37. Wang X, Terfve C, Rose JC, Markowetz F (2011) HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. Bioinformatics 27: 879–80. doi: 10.1093/bioinformatics/btr028 PMID: 21258062