Supplementary material for "Improved pathway reconstruction from RNA interference screens by exploiting off-target effects"

Sumana Srivatsa[1,2], Jack Kuipers[1,2], Fabian Schmich[1,2], Simone Eicher[3], Mario Emmenlauer[3], Christoph Dehio[3], and Niko Beerenwinkel[1,2]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland*
[2] *SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland*
[3] *Biozentrum, University of Basel, 4056 Basel, Switzerland*

1 Identifiability of pc-NEMs

Here we provide a detailed proof of the identifiability of pc-NEMs. As defined in the main text, $\mathcal{S}$ is a set of N signalling genes, $\mathcal{E}$ is a set of L effects genes, and $\mathcal{K}$ is a set of K knockdown experiments.

Theorem: Let a perturbation map ρ and a pair of error rates $\alpha, \beta \in [0, 0.5)$ be given. If there exists a subset $\mathcal{U} \subseteq \mathcal{K}$ of size at least N, such that $\rho_{us} \in [0, 1)$, for all $(u, s) \in \mathcal{U} \times \mathcal{S}$, and the matrix P defined by $P_{us} = \log(1 - \rho_{us})$ has rank N, then the pc-NEM $F^{\text{pc}} = \Pi\Theta$ is identifiable.

Proof: A model is identifiable if the mapping between the parameter space and the probability space of the data given by the likelihood function is one to one. For given ρ, α, β, and Θ, the likelihood function $\mathcal{L}(\Phi; D)$ is only a function of the DAG Φ and the data $D \in \mathcal{D} = \{0, 1\}^{L \times K}$ generated from K knockdown experiments with L phenotypic effects. Then, the pc-NEM is identifiable if for all $(\Phi_1, \Phi_2) \in \mathcal{G}$ and for all $D \in \mathcal{D}$

$$\mathcal{L}(\Phi_1; D) = \mathcal{L}(\Phi_2; D) \iff \Phi_1 = \Phi_2,$$

where $\mathcal{G}$ is the set of all DAGs.
Given a perturbation map ρ, an effects graph Θ, the error rates $\alpha, \beta \in [0, 0.5)$, and a pair of DAGs $(\Phi_1, \Phi_2) \in \mathcal{G}$, we start from $\mathcal{L}(\Phi_1; D) = \mathcal{L}(\Phi_2; D)$ and will prove that this implies $\Phi_1 = \Phi_2$. From the likelihood defined in Eq. 6, we have, for $j = 1, 2$,

$$\mathcal{L}(\Phi_j; D) = P(D \mid \rho, \Phi_j, \Theta, \alpha, \beta) = \prod_{l=1}^{L} \prod_{k=1}^{K} P(D_{lk} \mid \rho, \Phi_j, \Theta_{i(l)l}, \alpha, \beta)$$

Since $D_{lk} \in \{0, 1\}$ we can write

$$\mathcal{L}(\Phi_j; D) = \prod_{l=1}^{L} \prod_{k=1}^{K} D_{lk} P(D_{lk} = 1 \mid \rho, \Phi_j, \Theta_{i(l)l}, \alpha, \beta) \; + (1 - D_{lk}) P(D_{lk} = 0 \mid \rho, \Phi_j, \Theta_{i(l)l}, \alpha, \beta)$$

Based on Eq. 5 this is further equal to

$$\mathcal{L}(\Phi_j; D) = \prod_{l=1}^{L} \prod_{k=1}^{K} D_{lk}[\Pi_{j_{ki(l)}}(1 - \beta) + (1 - \Pi_{j_{ki(l)}})\alpha] \; + (1 - D_{lk})[(1 - \Pi_{j_{ki(l)}})(1 - \alpha) + \Pi_{j_{ki(l)}}\beta)],$$

where Π_j, for $j = 1, 2$, are the corresponding propagation matrices for the two DAGs Φ_1 and Φ_2, respectively.
Since the likelihoods are equal, we have

$$\begin{aligned} &\prod_{l=1}^{L} \prod_{k=1}^{K} D_{lk}[\Pi_{1_{ki(l)}}(1 - \beta) + (1 - \Pi_{1_{ki(l)}})\alpha] + (1 - D_{lk})[(1 - \Pi_{1_{ki(l)}})(1 - \alpha) + \Pi_{1_{ki(l)}}\beta)] \\ &= \prod_{l=1}^{L} \prod_{k=1}^{K} D_{lk}[\Pi_{2_{ki(l)}}(1 - \beta) + (1 - \Pi_{2_{ki(l)}})\alpha] + (1 - D_{lk})[(1 - \Pi_{2_{ki(l)}})(1 - \alpha) + \Pi_{2_{ki(l)}}\beta)] \end{aligned} \tag{S1}$$

This equality is true for all $D \in \mathcal{D}$. Now, imagine changing one l, k entry in the data. Then Eq. S1 with $D_{lk} = 1$ divided by Eq. S1 with $D_{lk} = 0$ would result in cancellation of all terms apart from the l, k terms corresponding to the changed entry. That is, the ratio of Eq. S1 with $D_{l,k} = 1$ to Eq. S1 with $D_{l,k} = 0$ is

$$\frac{\Pi_{1_{ki(l)}}(1-\beta)+(1-\Pi_{1_{ki(l)}})\alpha}{(1-\Pi_{1_{ki(l)}})(1-\alpha)+\Pi_{1_{ki(l)}}\beta)} = \frac{\Pi_{2_{ki(l)}}(1-\beta)+(1-\Pi_{2_{ki(l)}})\alpha}{(1-\Pi_{2_{ki(l)}})(1-\alpha)+\Pi_{2_{ki(l)}}\beta)}$$
$$\frac{\Pi_{1_{ki(l)}}(1-\alpha-\beta)+\alpha}{1-(\Pi_{1_{ki(l)}}(1-\alpha-\beta)+\alpha)} = \frac{\Pi_{2_{ki(l)}}(1-\alpha-\beta)+\alpha}{1-(\Pi_{2_{ki(l)}}(1-\alpha-\beta)+\alpha)}$$

This is true for any $l, k \in \mathcal{L} \times \mathcal{K}$. Let $A_1 = \Pi_{1_{ki(l)}}(1-\alpha-\beta)+\alpha$ and $A_2 = \Pi_{2_{ki(l)}}(1-\alpha-\beta)+\alpha$. Then the above equation can be written as

$$\frac{A_1}{1-A_1} = \frac{A_2}{1-A_2}$$

It follows that $A_1 = A_2$, that is,

$$\Pi_{1_{ki(l)}}(1-\alpha-\beta)+\alpha = \Pi_{2_{ki(l)}}(1-\alpha-\beta)+\alpha$$
$$\implies \Pi_{1_{ki(l)}} = \Pi_{2_{ki(l)}}$$

The step above relied on the restricton that $\alpha + \beta \neq 1$ which is guaranteed by the assumption on the range of α and β. Therefore, given $\alpha, \beta \in [0, 0.5)$, and since the above step holds for all l, k terms, the equality of the likelihoods for all data implies $\Pi_1 = \Pi_2$.

From Eq. 9 it follows that for each experiment $k \in \mathcal{K}$ and S-gene $s \in \mathcal{S}$ we have,

$$\begin{aligned} 1-\prod_{i=1}^{N}(1-\rho_{ki})^{C_{1_{is}}} &= 1-\prod_{i=1}^{N}(1-\rho_{ki})^{C_{2_{is}}} \\ \sum_{i=1}^{N} C_{1_{is}}\log(1-\rho_{ki}) &= \sum_{i=1}^{N} C_{2_{is}}\log(1-\rho_{ki}), \end{aligned} \tag{S2}$$

where C_1 and C_2 are the path count matrices for the two DAGs Φ_1 and Φ_2, respectively. Based on the condition on ρ, there exists a subset of knockdown experiments $\mathcal{U}$ with $|\mathcal{U}| \geq N$ and $\rho_{us} \in [0, 1)$ for $(u, s) \in \mathcal{U} \times \mathcal{S}$. Further, we have a matrix P as a function of experiments in $\mathcal{U}$ with $P_{us} = \log(1-\rho_{us})$. Rewriting Eq. S2 only for experiments $u \in \mathcal{U}$ as a matrix product

$$\sum_{i=1}^{N} C_{1_{is}}\log(1-\rho_{ui}) = \sum_{i=1}^{N} C_{2_{is}}\log(1-\rho_{ui})$$
$$P\,C_1 = P\,C_2$$

Since P is full rank it follows that $C_1 = C_2$.

The $N \times N$ path count matrix C is a function of the signal graph Φ, where each entry C_{ij} denotes the total number of paths (both direct and indirect) from node i to node j. It can be written as

$$C = f(\Phi) = \sum_{i=0}^{N}\Phi^i = \sum_{i=0}^{\infty}\Phi^i = (I-\Phi)^{-1}$$

because the adjacency matrix of any DAG Φ can be permuted and represented as a strictly upper triangular matrix by sorting the rows and columns accordingly. The matrix $(I - \Phi)$ is then an upper triangular matrix with 1's on the diagonal. Hence, all of its eigenvalues are equal to 1 and $(I - \Phi)$ is invertible. Thus, we can solve the equation above for Φ,

$$\Phi = I - C^{-1}$$

and see that $f : \Phi \to C$ is bijective, which proves that $\Phi_1 = \Phi_2$.

■

The condition on the range of knockdown probabilities of experiments in $\mathcal{U}$ ensures that the experiments can distinguish between any two DAGs. If $\rho_{ks} = 1$ then $\Pi_{ks} = 1$, reducing the model to classical NEM which is identifiable only in the transitively closed space. This condition in combination with the full rank condition guarantee unique likelihoods for each DAG. All the perturbation maps used in this paper complied with this condition illustrating that this assumption is generally fulfilled in practice.

2 Supplementary Figures

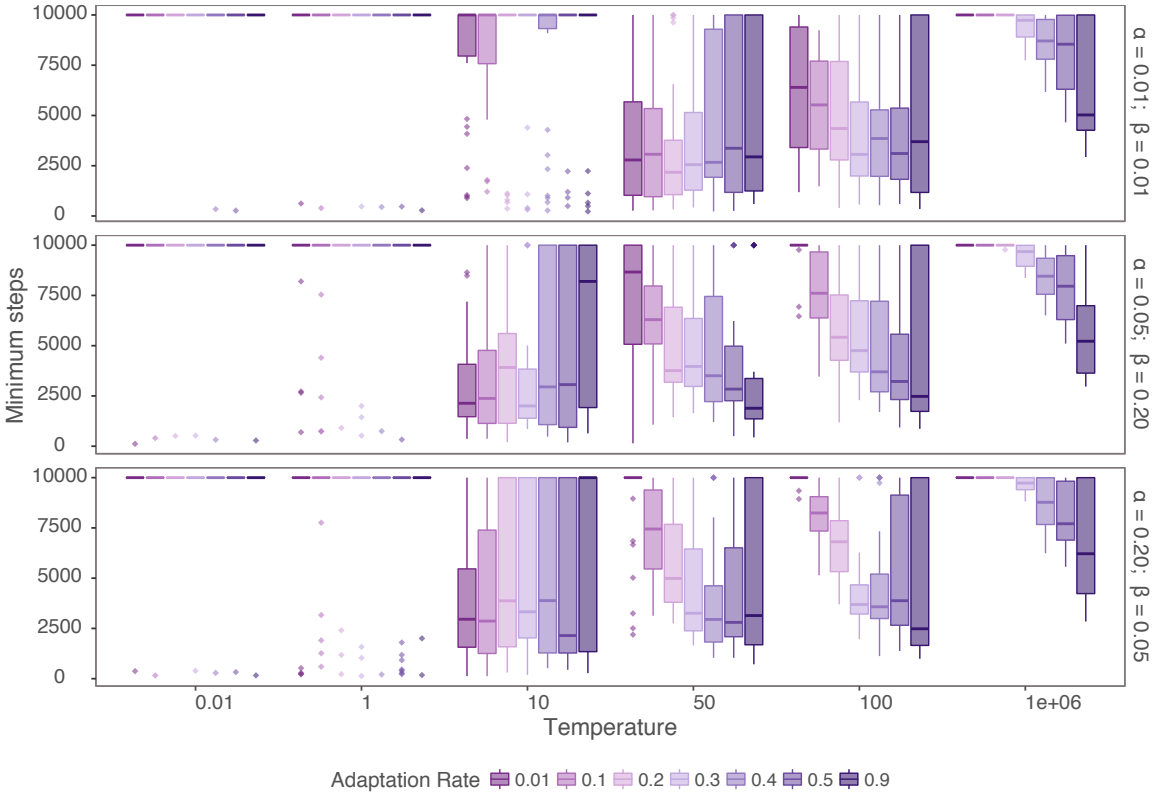

Figure S1: Parameter optimization for adaptive simulated annealing. The minimum number of steps (y-axis) required to attain the maximum likelihood at different start temperatures (x-axis) and adaptation rates for a fixed ideal acceptance rate of $\frac{1}{N} = 0.125$. Data simulated from 30 different networks with 320 phenotypic effects and noise levels of $\alpha = 0.01$ and $\beta = 0.01$, $\alpha = 0.05$ and $\beta = 0.20$, and $\alpha = 0.20$ and $\beta = 0.05$.

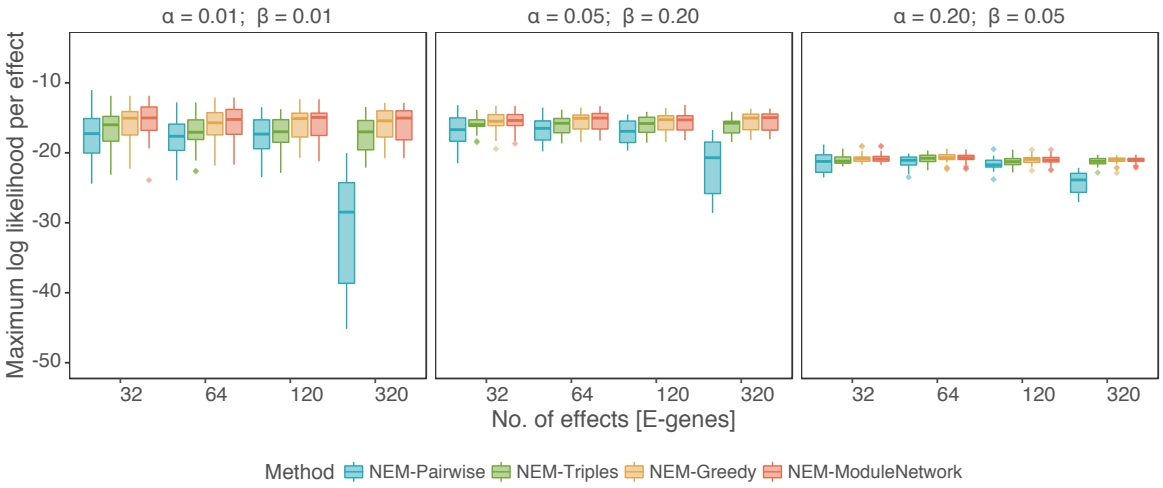

Figure S2: Comparison of different algorithms developed for network inference in NEMs. Maximum log likelihood per effect (y-axis) of the inferred network using pairwise (blue), triples (green), greedy (orange) and module network (red) algorithms. Each panel corresponds to inference from data with 32, 64, 120, and 320 effects and noise levels of $\alpha = 0.01$ and $\beta = 0.01$, $\alpha = 0.05$ and $\beta = 0.20$, and $\alpha = 0.20$ and $\beta = 0.05$.

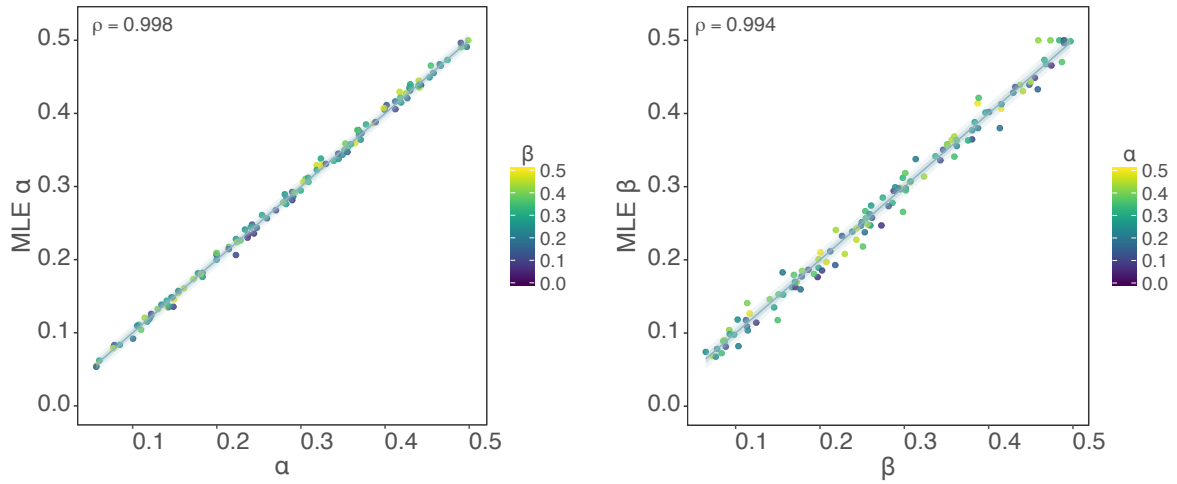

Figure S3: Noise estimation given the true network structure. Maximum likelihood estimates (y-axis) of 120 different α and β values learned using pc-NEMs given the true network structure, against true error rates (x-axis). The blue bands correspond to one and two standard deviations of inferring the error rates. In estimation of $\alpha(\beta)$, each point is coloured based on the corresponding $\beta(\alpha)$ value used to generate the data.

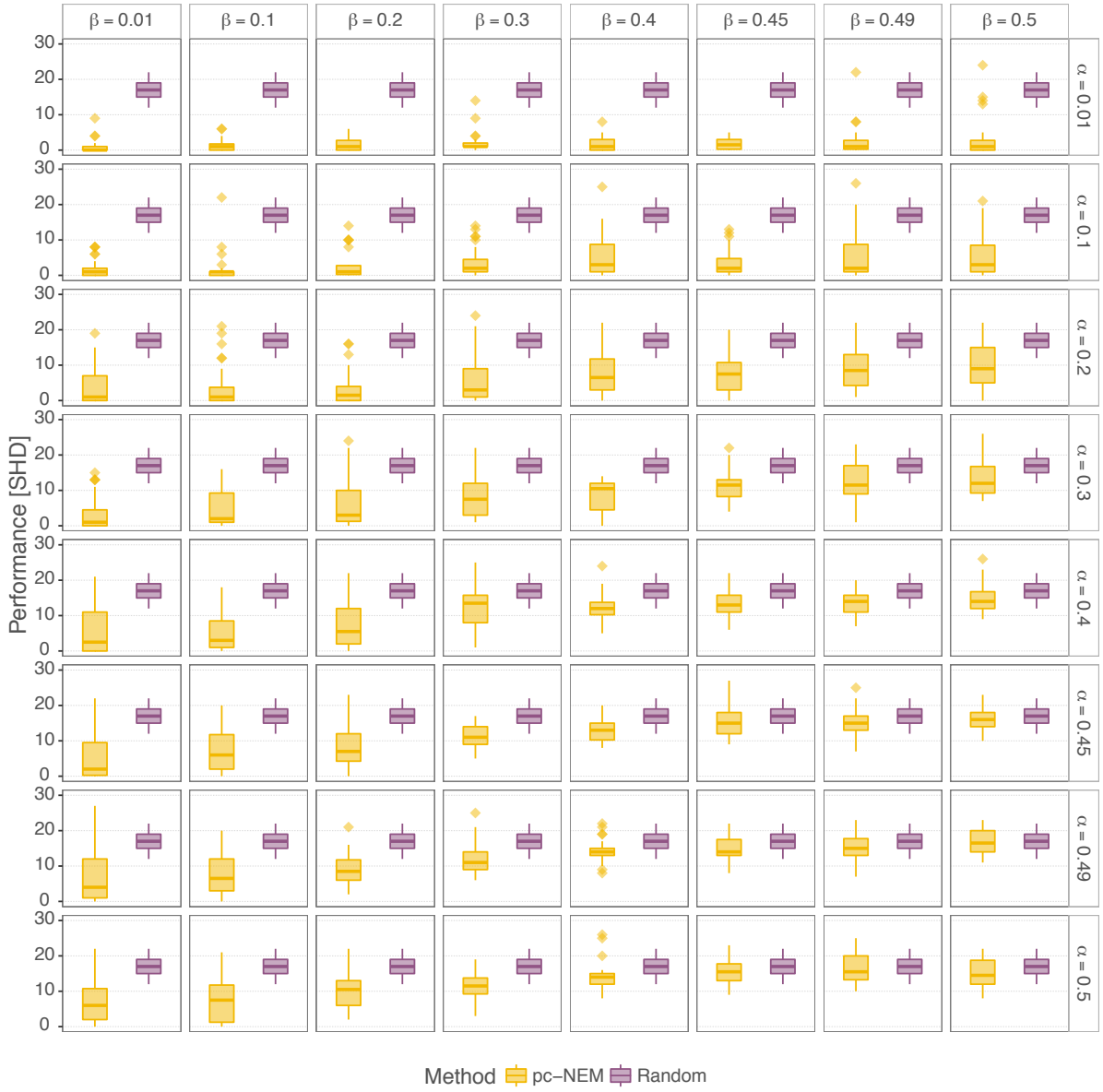

Figure S4: Performance summary of network and noise parameters inference. Each panel reports the performance (y-axis) of pc-NEMs (yellow) for network inference without *a priori* knowledge of error rates, on simulated data with 320 phenotypic effects at different noise levels. Each column (row) defines the performances on data with fixed β (α) value and varying α (β) values. The purple box-plots correspond to SHD of 30 uniformly sampled random DAGs.

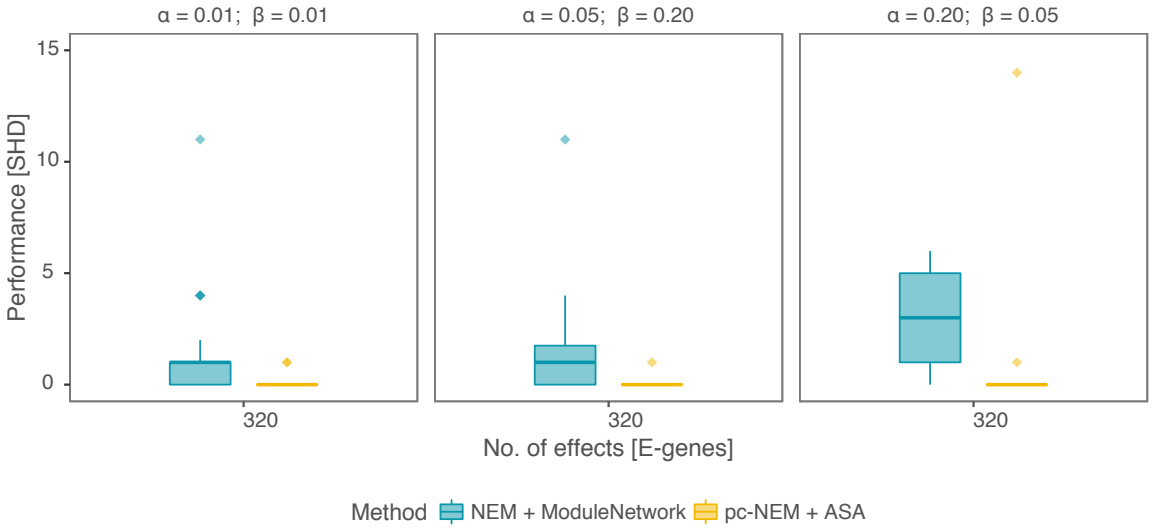

Figure S5: Results from simulation study on randomly sampled transitively closed networks. Structural Hamming distance (SHD) measuring the performance (y-axis) of pc-NEMs (yellow) and NEMs with module network (blue) algorithms, on simulated data from 30 different transitively closed networks of size $N = 8$ and perturbation maps, with 320 phenotypic effects and varying noise levels. The transitively closed networks were sampled at random from a set of all pathways in KEGG. The average number of off-targets per network is 21.37.

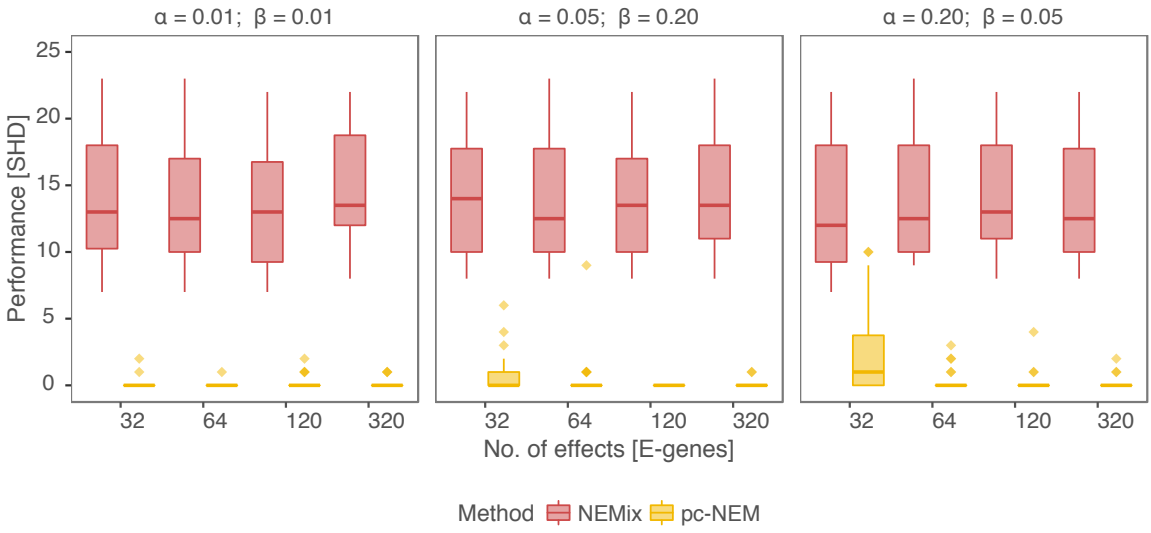

Figure S6: Results from simulation study on transitively closed networks. Structural Hamming distance (SHD) measuring the performance (y-axis) of pc-NEMs (yellow) and NEMix (red) on simulated data from 30 different transitively closed networks of size $N = 8$ and perturbation maps, with varying number of phenotypic effects and noise levels. The networks were sampled from *hsa*05200 pathway. It should be noted that NEMix is designed to perform network inference under unknown pathway stimulation from single-cell observations. Since it depends on single cell observations, it is heavily underpowered when used on gene-level data resulting in poor performance.

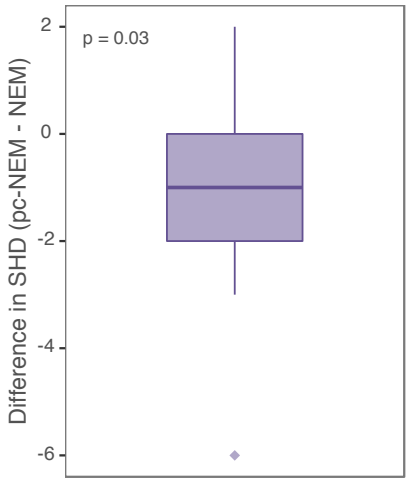

Figure S7: Performance comparison on HRV data. Difference of structural Hamming distance (SHD) between pc-NEMs and NEMs inferred networks on HRV data with high off-target effects. The p-value reported is from one-sample Wilcoxon signed rank test. The plot is based on the values summarised in Table S2.

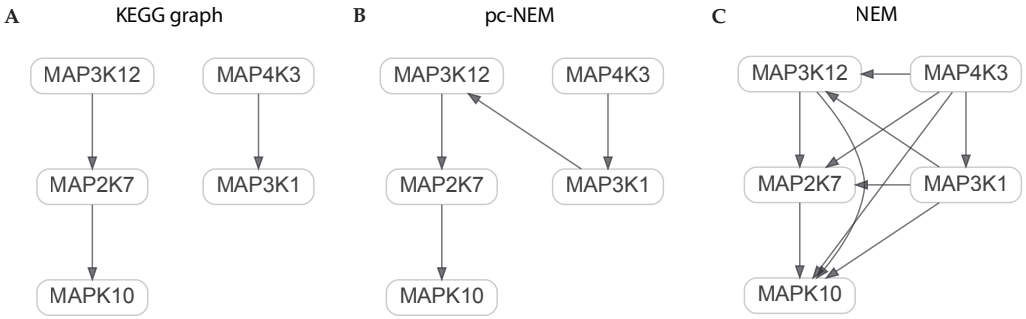

Figure S8: Example of networks inferred using HRV data. Binary gene-level data derived from image-based single-cell data, used for inferring a network of five genes involved in MAPK signaling pathway. **A.**The known KEGG pathway (hsa04010), **B.** The graph resulting from pc-NEM, and **C.** The graph inferred by NEM.

3 Supplementary Tables

Table S1: Performance as a function of network size. Structural Hamming distance (SHD) and runtime measuring performance of pc-NEMs on simulated data with 320 phenotypic effects generated from 30 different networks of size $N = 8$, $N = 12$ and $N = 16$, respectively. The networks were sampled at random from a set of all signalling pathways in KEGG.

No. of S-genes (N)	Median SHD	Median runtime (min)	Median min. iterations
8	0	13	2,348
12	0.5	37	8,918
16	4	86	17,740

Table S2: Performance summary on HRV data. Structural Hamming distance measuring the performance of pc-NEMs and NEMs on HRV infected RNAi screen data with low and high off-target effects. All the networks are of size $N = 5$.

Data	SHD	
	pc-NEM	**NEM**
Low off-target	7	7
High off-target	1	7
High off-target	3	5
High off-target	5	7
High off-target	5	7
High off-target	5	8
High off-target	5	5
High off-target	6	7
High off-target	6	8
High off-target	6	6
High off-target	7	7
High off-target	7	5
High off-target	7	8
High off-target	7	6
High off-target	8	8
High off-target	8	10
High off-target	8	9
High off-target	8	9
High off-target	8	7
High off-target	9	9
High off-target	10	8