




# Supplementary Text and Figures

*on manuscript*

## Learning causal networks with latent variables from multivariate information in genomic data


Louis Verny<sup>1,2</sup>, Nadir Sella<sup>1,2</sup>, Séverine Affeldt<sup>1,2</sup><sup>⊞a</sup>,  
Param Priya Singh<sup>1,2</sup><sup>⊞b</sup> & Hervé Isambert<sup>1,2\*</sup>

<sup>1</sup>Institut Curie, PSL Research University, CNRS, UMR168, 26 rue d'Ulm, 75005 Paris, France

<sup>2</sup>Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France

<sup>⊞a</sup> Current address: LIPADE, University of Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France

<sup>⊞b</sup> Current address: Department of Genetics, Stanford University, Palo Alto, USA

 These authors contributed equally to this work.

\*To whom correspondence should be addressed; E-mail: herve.isambert@curie.fr

## Contents

<b>1</b>	<b>Information-theoretic approach to network reconstruction</b>	<b>2</b>
1.1	Signature of causality <i>versus</i> indirect contributions to information in graphs . . . . .	2
1.2	Finite size effect and most likely contributor score . . . . .	3
<b>2</b>	<b>Algorithmic pipeline of the information-theoretic approach <code>miic</code></b>	<b>6</b>
2.1	Algorithm 1: Learning skeleton taking into account latent variables . . . . .	6
2.2	Algorithm 2: Confidence estimation and sign of retained edges . . . . .	8
2.3	Algorithm 3: Probabilistic orientation and propagation of remaining edges . . . . .	9
<b>3</b>	<b>Algorithmic implementation and tools</b>	<b>11</b>
3.1	<code>miic</code> R-package . . . . .	11
3.2	<code>miic</code> and <code>FCI</code> executables . . . . .	12
<b>4</b>	<b>References for Supplementary Text</b>	<b>16</b>
<b>5</b>	<b>Supplementary Figures</b>	<b>18</b>

# 1 Information-theoretic approach to network reconstruction

## 1.1 Signature of causality *versus* indirect contributions to information in graphs

We first discuss the rationale of the information-theoretic method to learn ancestral graphs with the assumption that an infinite amount of data is available with a distribution  $P(\mathbf{X})$  faithful to an underlying graph model  $\mathcal{G}$ , before discussing in the next section the necessary corrections needed, in practice, to account for the finite size of the dataset.

We will thus assume that the measured distribution  $P(\mathbf{X})$  is stable or faithful to the underlying graph model  $\mathcal{G}$ , implying that each structural independency under  $m$ -separation criterion [1] (*i.e.* each excluded edge  $XY$  in  $\mathcal{G}$ ) corresponds to a vanishing conditional mutual information as,

$$\begin{aligned} (X \perp_m Y|\{A_i\})_{\mathcal{G}} &\iff (X \perp\!\!\!\perp Y|\{A_i\})_P \\ &\iff I(X; Y|\{A_i\}) = 0 \end{aligned} \tag{S1}$$

**Theorem 1** [Signature of causality vs indirect contributions in  $\mathcal{G}$ , Affeldt & Isambert 2015 [2]] *Given some data with a distribution  $P(\mathbf{X})$  faithful to a graph  $\mathcal{G}$ ,*

- i) [Signature of causality] *If  $\exists X, Y, Z \in \mathbf{V}$  and  $\{A_i\} \subseteq \mathbf{V} \setminus \{X, Y, Z\}$  s.t.  $I(X; Y|\{A_i\}) = 0$  and  $I(X; Y; Z|\{A_i\}) < 0$ , then  $\mathcal{G}$  is necessarily causal, *i.e.* it has at least one  $v$ -structure.*
- ii) [Indirect contribution]  *$\forall X, Y, Z \in \mathbf{V}$  and  $\forall \{A_i\} \subseteq \mathbf{V} \setminus \{X, Y, Z\}$  s.t.  $I(X; Y; Z|\{A_i\}) > 0$ , then  $I(X; Y|\{A_i\}) = I(X; Y; Z|\{A_i\}) + I(X; Y|Z, \{A_i\}) > 0$  and  $I(X; Y; Z|\{A_i\}) > 0$  can be seen as the positive contribution to the remaining conditional mutual information  $I(X; Y|\{A_i\}) > 0$  (and equivalently to  $I(X; Z|\{A_i\}) > 0$  and  $I(Y; Z|\{A_i\}) > 0$  by symmetry of  $I(X; Y; Z|\{A_i\})$ ).*

**Sketch of proof** [The full proof is given in [2]: Theorem 4 and Corollary 2] *i)* is proven via its contrapositive showing that non-causal graphs with structural independence  $I(X; Y|\{A_i\}) = 0$  have necessarily vanishing conditional three-point information terms,  $I(X; Y; Z|\{A_i\}) = 0, \forall Z \in \mathbf{V} \setminus (\{X, Y\} \cup \{A_i\})$ . Moreover, since  $I(X; Y; Z|\{A_i\}) \leq 0$  whenever  $I(X; Y|\{A_i\}) = 0$  for any graph in general (from Eq. 5 in main text with  $A_n = Z$ ), it means that  $I(X; Y; Z|\{A_i\}) < 0$  indeed implies a necessary causal graph. *ii)* is simply Eq. 5 in main text with  $A_n = Z$  as well.  $\square$

Theorem 1 *i)*, which characterizes the signature of causality in observational data, will be used to orient  $v$ -structures, once Theorem 1 *ii)* has been used to learn structural independences by collecting one-by-one the significant contributors  $\{A_i\}$  and partitioning iteratively mutual information terms into positive contributions from indirect paths as,

$$\begin{aligned} I(X; Y) &= I(X; Y; A_1) + I(X; Y|A_1) \\ &= I(X; Y; A_1) + I(X; Y; A_2|A_1) + I(X; Y|A_1, A_2) \\ &= I(X; Y; A_1) + I(X; Y; A_2|A_1) + \dots \\ &\quad \dots + I(X; Y; A_n|\{A_i\}_{n-1}) + I(X; Y|\{A_i\}_n) \end{aligned} \tag{S2}$$

with  $I(X; Y; A_k | \{A_i\}_{k-1}) > 0$  for all  $k$ . Hence, conditional independence,  $I(X; Y | \{A_i\}_n) = 0$ , is eventually retrieved (if it holds) after subtracting successive significant positive three-point conditional information from the original two-point conditional information [2, 3] as,

$$I(X; Y | \{A_i\}_n) = I(X; Y) - I(X; Y; A_1) - \dots - I(X; Y; A_n | \{A_i\}_{n-1}) \quad (\text{S3})$$

The robustness of the approach hinges on picking the most likely contributors first to avoid a later accumulation of incorrect contributors in an attempt to compensate for early errors. Choosing the most likely contributors requires, however, to take into account the finite size of the dataset as detailed in the next section.

## 1.2 Finite size effect and most likely contributor score

This section, adapted from [2], addresses finite size corrections to multivariate information and introduce a heuristic score to collect the most likely contributors  $\{A_i\}_n$  in Eq. S3.

Given  $N$  independent samples from some available data  $\mathcal{D}$ , the Maximum Likelihood,  $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ , that they might have been generated by the graphical model  $\mathcal{G}$ , is given by [4],

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-NH(p,q)}}{Z_{\mathcal{D},\mathcal{G}}} = \frac{e^{N \sum_{\mathbf{x}} p(\mathbf{x}) \log q(\mathbf{x})}}{Z_{\mathcal{D},\mathcal{G}}} \quad (\text{S4})$$

where  $H(p, q) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log q(\mathbf{x})$  is the cross entropy between the empirical probability distribution  $p(\mathbf{x})$  of the data  $\mathcal{D}$  and the theoretical probability distribution  $q(\mathbf{x})$  of the model  $\mathcal{G}$ , and  $H(p) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$  is the entropy of the data and  $Z_{\mathcal{D},\mathcal{G}}$  a data- and model-dependent factor ensuring proper normalization condition. The structural constraints of the model  $\mathcal{G}$  are included in the factorization form of the theoretical probability distribution,  $q(\mathbf{x})$ .

In particular, the conditional mutual information,  $I(X; Y | \{A_i\})$ , for structural independence, Eq. S3, cannot be exactly zero, given a finite dataset of  $N$  independent samples, and has to be compared to a finite threshold,  $I(X; Y | \{A_i\}) < k_{X;Y|\{A_i\}}/N$ , where  $k_{X;Y|\{A_i\}} > 0$  is related to the likelihood normalization ratio between graphs including or excluding edge  $XY$  with separation set  $\{A_i\}$  [2],

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = \frac{e^{-NI(X;Y|\{A_i\})}}{Z_{\mathcal{D},\mathcal{G}_{\setminus XY|\{A_i\}}}/Z_{\mathcal{D},\mathcal{G}}} = e^{-NI(X;Y|\{A_i\})+k_{X;Y|\{A_i\}}} \quad (\text{S5})$$

$$k_{X;Y|\{A_i\}} = \log(Z_{\mathcal{D},\mathcal{G}}/Z_{\mathcal{D},\mathcal{G}_{\setminus XY|\{A_i\}}}) \quad (\text{S6})$$

where  $k_{X;Y|\{A_i\}}$  tends to limit the complexity of the models by favoring fewer edges. A common complexity criterion in model selection is the Bayesian Information Criterion (BIC) or Minimum Description Length (MDL) criterion [5, 6], which is simply related to the maximum likelihood normalization constant reached in the asymptotic limit of a large dataset  $N \rightarrow \infty$  (Laplace approximation). However, this limit distribution is only reached for very large datasets in practice. Alternatively, the normalization of the maximum likelihood can also be done over all possible datasets including the same number of samples to yield a (universal) Normalized Maximum Likelihood (NML) criterion [7, 8] and its decomposable

version [9, 10]. All application results presented in this paper are obtained with the  $XY$ -symmetric decomposable NML criterion introduced in [3], which was shown to yield significantly better results than BIC/MDL criterion on benchmark networks.

Thus, finite size effects in graphical model comparison can be included by redefining two-point and three-point conditional multivariate information as,

$$I'(X; Y|\{A_i\}) = I(X; Y|\{A_i\}) - \frac{k_{X;Y|\{A_i\}}}{N} \quad (\text{S7})$$

$$I'(X; Y; Z|\{A_i\}) = I(X; Y; Z|\{A_i\}) - \frac{k_{X;Y;Z|\{A_i\}}}{N} \quad (\text{S8})$$

where conditional three-point information including finite size corrections,  $I'(X; Y; Z|\{A_i\})$ , and their associated complexity terms,  $k_{X;Y;Z|\{A_i\}}$ , are defined with respect to two-point information including finite size corrections and their associated complexity terms, using the same Eq. 5 in main text with  $A_n = Z$ ,

$$I'(X; Y; Z|\{A_i\}) = I'(X; Y|\{A_i\}) - I'(X; Y|\{A_i\}, Z) \quad (\text{S9})$$

$$k_{X;Y;Z|\{A_i\}} = k_{X;Y|\{A_i\}} - k_{X;Y|\{A_i\}, Z} \quad (\text{S10})$$

Hence, Eq. S3 including finite size corrections becomes,

$$I'(X; Y|\{A_i\}_n) = I'(X; Y) - I'(X; Y; A_1) - \dots - I'(X; Y; A_n|\{A_i\}_{n-1}) \quad (\text{S11})$$

where the conditional two-point and tree-point multivariate information are related to the following maximum likelihood ratios, using Eq. S6,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X; Y|\{A_i\})} \quad (\text{S12})$$

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}} = e^{NI'(X; Y; Z|\{A_i\})} \quad (\text{S13})$$

with conditional independence including finite size effect corresponding to  $I'(X; Y|\{A_i\}) \leq 0$ .

Hence, learning, iteratively, the most likely edge to be removed  $XY$  and its corresponding separation set  $\{A_i\}$  will imply to simultaneously minimize two-point information (Eq. S12) while maximizing three-point information (Eq. S13). In fact, the sign and magnitude of conditional three-point information including finite size corrections,  $I'(X; Y; Z|\{A_i\})$ , determine the probability that  $Z$  should be included in or excluded from the sepset candidate  $\{A_i\}$  as:

- If  $I'(X; Y; Z|\{A_i\}) > 0$ ,  $Z$  is more likely to be included in  $\{A_i\}$  with probability,

$$P_{\text{nv}}(X; Y; Z|\{A_i\}) = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}}} = \frac{1}{1 + e^{-NI'(X; Y; Z|\{A_i\})}} \quad (\text{S14})$$

- If  $I'(X; Y; Z|\{A_i\}) < 0$ ,  $Z$  is more likely to be excluded from  $\{A_i\}$ , suggesting obligatory causal relationships in the form of a v-structure between  $X, Y, Z$  with probability,

$$P_v(X; Y; Z|\{A_i\}) = 1 - P_{nv}(X; Y; Z|\{A_i\}) = \frac{1}{1 + e^{NI'(X; Y; Z|\{A_i\})}} \quad (\text{S15})$$

But, in the case  $I'(X; Y; Z|\{A_i\}) > 0$ , Eq. S13 can also be interpreted as quantifying the likelihood increase that the edge  $XY$  should be removed from the model by extending the candidate sepset from  $\{A_i\}$  to  $\{A_i\} + Z$ , i.e.  $\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}} = \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}} \times \exp(NI'(X; Y; Z|\{A_i\})) > \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}$ , as  $\exp(NI'(X; Y; Z|\{A_i\})) > 1$ . Yet, as the three-point information,  $I'(X; Y; Z|\{A_i\})$ , is actually symmetric with respect to the variables,  $X, Y$  and  $Z$ , the factor  $\exp(NI'(X; Y; Z|\{A_i\}))$  provides in fact the same likelihood increase for the removal of the three edges  $XY, XZ$  and  $ZY$ , conditioned on the same initial set of nodes  $\{A_i\}$ , namely,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}, Y}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\}, X}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\}}}} = e^{NI'(X; Y; Z|\{A_i\})} \quad (\text{S16})$$

However, despite this symmetry of three-point information,  $I'(X; Y; Z|\{A_i\})$ , the likelihoods that the edges  $XY, XZ$  and  $ZY$  should be removed are not the same, as they depend on different 2-point information,  $I'(X; Y|\{A_i\})$ ,  $I'(X; Z|\{A_i\})$  and  $I'(Z; Y|\{A_i\})$ , Eq. S12. In particular, the likelihood ratio between the removals of the alternative edges  $XY$  and  $XZ$  is given by,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}, Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}, Y}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}}} = \frac{e^{-NI'(X; Y|\{A_i\})}}{e^{-NI'(X; Z|\{A_i\})}} \quad (\text{S17})$$

and similarly between edges  $XY$  and  $ZY$ .

Hence, for  $XY$  to be the most likely edge to be removed conditioned on the sepset  $\{A_i\} + Z$ , not only  $Z$  should contribute through  $I'(X; Y; Z|\{A_i\}) > 0$  with probability  $P_{nv}(X; Y; Z|\{A_i\})$  (Eq. S14), but  $XY$  must also correspond to the ‘weakest’ edge of  $XY, XZ$  and  $ZY$  conditioned on  $\{A_i\}$ , as given by the lowest conditioned 2-point information, Eq. S17. Note that removing the edge  $XY$  with the lowest conditional 2-point information is consistent, as expected, with the Data Processing Inequality,  $I(X; Y|\{A_i\}) \leq \min(I(X; Z|\{A_i\}), I(Z; Y|\{A_i\}))$ , in the limit of large datasets. However, quite frequently,  $XZ$  or  $ZY$  might also have low conditional 2-point information, so that the edge removal associated with the symmetric contribution  $I(X; Y; Z|\{A_i\})$  will only be consistent with the Data Processing Inequality (DPI) with probability,

$$\begin{aligned} P_{\text{dpi}}(XY; Z|\{A_i\}) &= \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\}}}} \\ &= \frac{1}{1 + \frac{e^{-NI'(X; Z|\{A_i\})}}{e^{-NI'(X; Y|\{A_i\})}} + \frac{e^{-NI'(Z; Y|\{A_i\})}}{e^{-NI'(X; Y|\{A_i\})}}} \end{aligned} \quad (\text{S18})$$

In practice, taking into account this DPI-consistency probability  $P_{\text{dpi}}(XY; Z|\{A_i\})$ , as detailed below, significantly improves the results obtained by relying solely on the ‘non-v-structure’ probability

$P_{\text{nv}}(X; Y; Z|\{A_i\})$ . Conversely, the DPI-consistency probability  $P_{\text{dpi}}(XY; Z|\{A_i\})$  is not sufficient on its own to uncover causal relationships between variables, which require to compute three-point information  $I(X; Y; Z|\{A_i\})$  and the probability  $P_{\text{nv}}(X; Y; Z|\{A_i\})$  (see Proposition 1 and Proposition 2, below).

To optimize the likelihood that the edge  $XY$  can be accounted for by the additional contribution of  $Z$  conditioned on previously selected  $\{A_i\}$ , we propose to combine the maximum of three-point information (Eq. S14) and the minimum of 2-point information (Eq. S18) by defining the score  $S_{\text{lb}}(Z; XY|\{A_i\})$  as the lower bound of  $P_{\text{nv}}(X; Y; Z|\{A_i\})$  and  $P_{\text{dpi}}(XY; Z|\{A_i\})$ , since both conditions need to be fulfilled to warrant that edge  $XY$  is likely to be absent from the model  $\mathcal{G}$ ,

$$S_{\text{lb}}(Z; XY|\{A_i\}) = \min \left[ P_{\text{nv}}(X; Y; Z|\{A_i\}), P_{\text{dpi}}(XY; Z|\{A_i\}) \right] \quad (\text{S19})$$

Hence, the pair of nodes  $XY$  with the most likely contribution from a third node  $Z$  and likely to be absent from the model can be ordered according to their rank  $R(XY; Z|\{A_i\})$  defined as,

$$R(XY; Z|\{A_i\}) = \max_Z (S_{\text{lb}}(Z; XY|\{A_i\})) \quad (\text{S20})$$

Then,  $Z$  can be iteratively added to the set of contributing nodes (*i.e.*  $\{A_i\} \leftarrow \{A_i\} + Z$ ) of the top edge  $XY = \operatorname{argmax}_{XY} R(XY; Z|\{A_i\})$  to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph.

## 2 Algorithmic pipeline of the information-theoretic approach `miic`

The implementation of the information-theoretical approach `miic` proceeds in three steps corresponding to the following algorithmic pipeline:

- Algorithm 1: Learning skeleton taking into account latent variables
- Algorithm 2: Confidence estimation and sign of retained edges
- Algorithm 3: Probabilistic orientation and propagation of remaining edges

`miic` is implemented in an R-package and freely available under a General Public License (Supplementary Software).

### 2.1 Algorithm 1: Learning skeleton taking into account latent variables

Using the heuristic score, Eq. S20, to implement the successive subtractions of three-point conditional information terms in Eq. S11, yields Algorithm 1

---

**Algorithm 1:** Skeleton reconstruction in the presence of latent variables

---

**In:** observational data of finite size  $N$ , complexity criterion NML (or MDL)

**Out:** skeleton of ancestral graph  $\mathcal{G}$

**Initiation**

Start with complete undirected graph

**forall the edges  $XY$  do**

**if  $I'(X; Y) < 0$  then**

**$XY$  edge is non-essential and removed**

**separation set of  $XY$ :**  $\text{Sep}_{XY} = \emptyset$

**else**

        find the **most contributing node**  $Z$  and **compute its rank**,  $R(XY; Z|\emptyset)$

        ( $Z$  can be restricted to neighbors of  $X$  and  $Y$  if latent variables are excluded)

**end**

**end**

**Iteration**

**while  $\exists XY$  edge with  $R(XY; Z|\{A_i\}) > 1/2$  do**

**for edge  $XY$  with highest rank  $R(XY; Z|\{A_i\})$  do**

**expand contributing set  $\{A_i\} \leftarrow \{A_i\} + Z$**

**if  $I'(X; Y|\{A_i\}) < 0$  then**

**$XY$  edge is non-essential and removed**

**separation set of  $XY$ :**  $\text{Sep}_{XY} = \{A_i\}$

**else**

            find the **next most contributing node**  $Z$  and **compute rank**,  $R(XY; Z|\{A_i\})$

            ( $Z$  can be restricted to neighbors of  $X$  and  $Y$  if latent variables are excluded)

**end**

**update highest rank edge**

**end**

**end**

---

## 2.2 Algorithm 2: Confidence estimation and sign of retained edges

Once a first skeleton has been obtained using Algorithm 1, the confidence on each retained edge can be estimated through an edge specific confidence ratio  $C_{XY}$  based on the probability  $P_{XY}$  to remove a directed edge  $X \rightarrow Y$  from the graph  $\mathcal{G}$ , as defined by Eq. S12,

$$P_{XY} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X;Y|\{A_i\})} \quad (\text{S21})$$

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \quad (\text{S22})$$

where  $\langle P_{XY}^{\text{rand}} \rangle$  is the average of the probability to remove the  $XY$  edge after randomly permutating the dataset for each observable. Hence, the lower  $C_{XY}$ , the higher the confidence on the  $XY$  edge.

In practice,  $\langle P_{XY}^{\text{rand}} \rangle$  is not actually evaluated looking for contributors  $\{A_i\}$  as done for  $P_{XY}$  (since there should be no contributors nor edges after randomization of the data) but just computing  $\langle P_{XY}^{\text{rand}} \rangle = \langle e^{-NI'(X^{\text{rand}};Y)} \rangle$ , where the  $X^{\text{rand}}$  variable is assigned randomly permuted values of  $X$  across the different samples (randomizing  $Y$  or both variables is statistically equivalent). As a result,  $C_{XY}$  is slightly overestimated (as ignoring contributors actually underestimates  $\langle P_{XY}^{\text{rand}} \rangle$ ) but can be computed efficiently by averaging over hundreds of permuted values at each vertex. The filtering of retained edges is implemented in Algorithm 2.

---

**Algorithm 2:** Filtering retained edges according to an edge specific confidence ratio  $C_{XY}$

---

**In:** Skeleton obtained from Algorithm 1, confidence level  $C_s < 1$ , nb permutations  $r_{\text{max}}$

**Out:** Revised skeleton, after filtering out lower confidence edges with  $C_{XY} > C_s$

**forall the vertices  $X_i$  do**

**forall the random permutations  $r < r_{\text{max}}$  do**

        Assign  $X_i^{\text{rand}}$  values through **random permutation** of  $X_i$  values

**forall the  $X_j$  adjacent of  $X_i$  with  $j > i$  do**

            | Compute  $I'_r(X_i^{\text{rand}}; X_j) \leftarrow \max(0, I'(X_i^{\text{rand}}; X_j))$

**end**

**end**

**forall the  $X_j$  adjacent of  $X_i$  with  $j > i$  do**

        | Compute  $\langle P_{X_i X_j}^{\text{rand}} \rangle = \langle e^{-NI'_r(X_i^{\text{rand}}; X_j)} \rangle_{r_{\text{max}}}$

        | Compute  $C_{X_i X_j} = P_{X_i X_j} / \langle P_{X_i X_j}^{\text{rand}} \rangle$  and **remove edge**  $X_i X_j$ , if  $C_{X_i X_j} > C_s$

**end**

**end**

---

In addition, the sign of each retained edge,  $X - Y$ , is defined by the sign of the partial correlation coefficient,  $\rho_{XY \cdot \mathbf{A}}$ , between  $X$  and  $Y$  conditioned on its derived contributors  $\mathbf{A} = \{A_i\}$  in Algorithm 1,

with positive coefficients corresponding to partial correlations and negative coefficients corresponding to partial anti-correlations.

The partial correlation coefficient can be computed in  $\mathcal{O}(n^3)$  using matrix inversion of the partial covariance matrix (on  $\{X, Y\} \cup \mathbf{A}$  variables) or by dynamic programming using the recursive formula, for any  $A_o \in \mathbf{A}$ ,

$$\rho_{XY \cdot \mathbf{A}} = \frac{\rho_{XY \cdot \mathbf{A} \setminus \{A_o\}} - \rho_{XA_o \cdot \mathbf{A} \setminus \{A_o\}} \rho_{A_o Y \cdot \mathbf{A} \setminus \{A_o\}}}{\sqrt{1 - \rho_{XA_o \cdot \mathbf{A} \setminus \{A_o\}}^2} \sqrt{1 - \rho_{A_o Y \cdot \mathbf{A} \setminus \{A_o\}}^2}} \quad (\text{S23})$$

Negative partial correlations are represented as blue edges in the predicted network reconstructions, Figures 2-4 and their figure supplements.

### 2.3 Algorithm 3: Probabilistic orientation and propagation of remaining edges

Given the skeleton obtained from Algorithm 1, possibly filtered through Algorithm 2, based on edge specific confidence ratio, Eqs. S14 and S15 can then be used to establish the following Proposition 1 and Proposition 2 for probabilistic orientation and propagation rules of unshielded triples.

To this end, let us first introduce three different endpoint marks associated to edges in mixed graphs: they are the tail ( $-$ ), the head ( $>$ ) and the unspecified ( $\circ$ ) endpoint marks. In addition, we will use the asterisk symbol ( $*$ ) as a wild card denoting any of the three marks and define orientation probabilities at either one or two (underlined> endmarks using Propositions 1 and 2 below.

**Proposition 1** [Robust orientation of v-structures from finite dataset including latent variables]

Assuming that the underlying graphical model is an ancestral graph  $\mathcal{G}$  on  $\mathbf{V}$ ,

if  $\exists X, Y, Z, \{A_i\} \in V$  s.t.  $I'(X; Y; Z | \{A_i\}) < 0$  then,

- i. if  $X, Y, Z$  form an unshielded triple,  $X * \circ Z \circ * Y$  with  $X \neq Y$ , then it should be oriented as  $X * \rightarrow Z \leftarrow * Y$ , with endmark probabilities at  $\underline{Z}$ ,

$$P_{X * \rightarrow \underline{Z}}^{\circ} = P_{Y \circ \rightarrow \underline{Z}}^{\circ} = \frac{1 + e^{NI'(X; Y; Z | \{A_i\})}}{1 + 3e^{NI'(X; Y; Z | \{A_i\})}} \quad (\text{S24})$$

- ii. similarly, if  $X, Y, Z$  form an unshielded triple, with one already known converging arrow into the middle node,  $X * \rightarrow Z \circ * Y$ , with endmark probability at  $\underline{Z}$ ,  $P_{X * \rightarrow \underline{Z}} > P_{X * \rightarrow \underline{Z}}^{\circ}$ , then the second edge should be oriented to form a v-structure,  $X * \rightarrow Z \leftarrow * Y$ , with endmark probability at  $\underline{Z}$ ,

$$P_{Y * \rightarrow \underline{Z}} = P_{X * \rightarrow \underline{Z}} \left( \frac{1}{1 + e^{NI'(X; Y; Z | \{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \quad (\text{S25})$$

**Proof.** The implications (i.) and (ii.) rely on Eq. S15 to estimate the probability that the two edges form

a v-structure. We start proving (ii.) using the probability decomposition formula:

$$\begin{aligned}
P_{Y^* \rightarrow Z} &= P_{X^* \rightarrow Z} \frac{P_{X^* \rightarrow Z \leftarrow^* Y}}{P_{X^* \rightarrow Z \leftarrow^* Y} + P_{X^* \rightarrow Z \rightarrow Y}} \\
&\quad + (1 - P_{X^* \rightarrow Z}) \frac{P_{X \leftarrow Z \leftarrow^* Y}}{P_{X \leftarrow Z \leftarrow^* Y} + P_{X \leftarrow Z \rightarrow Y}} \\
&= P_{X^* \rightarrow Z} \left( \frac{1}{1 + e^{NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2}
\end{aligned} \tag{S26}$$

which also leads to (i.) if one assumes  $P_{X^* \rightarrow Z} = P_{Y^* \rightarrow Z}$  by symmetry in absence of prior information on these orientations.  $\square$

Following the rationale of constraint-based approaches, it is then possible to ‘propagate’ further the orientations downstream of v-structures, using Eq. S14 for positive (conditional) three-point information. For simplicity and consistency, we only implement the propagation of orientation based on likelihood ratios, which can be quantified for finite datasets as proposed in the following Proposition 2. Hence, we do not apply the complete propagation rules for ancestral graphs [11], which enforce in particular acyclic constraints, that are necessary to have a complete reconstruction of the Markov equivalent class of the underlying ancestral graph model.

**Proposition 2** [Robust propagation of orientations from finite dataset including latent variables]

*Assuming that the underlying graphical model is an ancestral graph  $\mathcal{G}$  on  $\mathbf{V}$ ,*

*$\forall X, Y, Z, \{A_i\} \in V$  s.t.  $I'(X; Y; Z|\{A_i\}) > 0$ , if  $X, Y, Z$  form an unshielded triple with one already known converging orientation,  $X^* \rightarrow Z \circ -^* Y$ , with endmark probability at  $\underline{Z}$ ,  $P_{X^* \rightarrow Z} > 1/2$ , then this orientation should be ‘propagated’ to the second edge as  $X^* \rightarrow Z \rightarrow Y$ , with endmark probability at  $\underline{Z}$  and  $\underline{Y}$ ,*

$$P_{\underline{Z} \rightarrow \underline{Y}} = P_{X^* \rightarrow Z} \left( \frac{1}{1 + e^{-NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \tag{S27}$$

**Proof.** This results is shown using the probability decomposition formula,

$$\begin{aligned}
P_{\underline{Z} \rightarrow \underline{Y}} &= P_{X^* \rightarrow Z} \frac{P_{X^* \rightarrow Z \rightarrow Y}}{P_{X^* \rightarrow Z \leftarrow^* Y} + P_{X^* \rightarrow Z \rightarrow Y}} \\
&\quad + (1 - P_{X^* \rightarrow Z}) \frac{P_{X \leftarrow Z \rightarrow Y}}{P_{X \leftarrow Z \leftarrow^* Y} + P_{X \leftarrow Z \rightarrow Y}} \\
&= P_{X^* \rightarrow Z} \left( \frac{1}{1 + e^{-NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2}
\end{aligned} \tag{S28}$$

$\square$

Proposition 1 and Proposition 2 lead to the following Algorithm 3 for the orientation of unshielded triples of the graph skeleton obtained from Algorithm 1 with possibly additional edge filtering through Algorithm 2.

---

**Algorithm 3:** Probabilistic Orientation / Propagation of edges including latent variables

---

**In:** Graph skeleton from Algorithm 1, possibly filtered through Algorithm 2, and corresponding conditional three-point information  $I'(X; Y; Z|\{A_i\})$ .

**Out:** Partially oriented causal graph  $\mathcal{G}$  with endmark orientation probabilities.

**Probabilistic Orientation / Propagation Step including latent variables**

**sort** list of unshielded triples,  $\mathcal{L}_c = \{\langle X, Z, Y \rangle_{X \neq Y}\}$ , in decreasing order of their endmark orientation/propagation probabilities initialized at 1/2 and computed from:

- (i.) Proposition 1, if  $I'(X; Y; Z|\{A_i\}) < 0$ , or
- (ii.) Proposition 2, if  $I'(X; Y; Z|\{A_i\}) > 0$

**repeat**

Take  $\langle X, Z, Y \rangle_{X \neq Y} \in \mathcal{L}_c$  with highest endmark orient./propa. probability  $> 1/2$ .

**if**  $I'(X; Y; Z|\{A_i\}) < 0$  **then**

**Orient/propagate** edge direction(s) to form a **v-structure**  $X * \rightarrow Z \leftarrow * Y$  with endmark probabilities  $P_{X * \rightarrow Z}$  and  $P_{Y * \rightarrow Z}$  given by **Proposition 1**.

**else**

**Propagate** second edge direction to form a **non-v-structure**  $X * \rightarrow Z \rightarrow Y$  assigning endmark probabilities  $P_{Z \rightarrow Y}$  from **Proposition 2**.

**end**

Apply new orientation(s) and **sort** remaining list of unshielded triples  $\mathcal{L}_c \leftarrow \mathcal{L}_c \setminus \langle X, Z, Y \rangle_{X \neq Y}$  after **updating propagation probabilities**.

**until** no additional endmark orient./propa. probability  $> 1/2$ ;

---

### 3 Algorithmic implementation and tools

We provide the `miic` software in two formats, an R-package to be used in the R environment, and executables to be used directly in a terminal.

#### 3.1 `miic` R-package

`miic` R-package contains the full implementation of the method designed for the R environment.

To install `miic`, enter R and install the R package as,

```
install.packages("<path>/miic.0.1.tar.gz", repos=NULL, source=T)
```

The package also provides functions to plot the predicted networks within the R environment and to export the results in graphml format for cytoscape graphical display. See the package documentation for additional details on `miic` R-package.

The R-package contains the three application datasets presented in the main text (Figures 2-4). An example of network reconstruction is detailed below:

```
library(miic)
# load hematopoiesis data frame
data(hematoData)

# execute MIIC (reconstruct graph)
miic.res = miic(inputData = hematoData, latent = TRUE, confidenceShuffle = 100,
confidenceRatio = 0.001)

# plot graph
miic.plot(miic.res)

# write graph to graphml format. Note that to correctly visualize the network we created the miic style for Cy-
toscape (http://www.cytoscape.org/).
miic.write.cytoscape(g = miic.res, file = "..")
```

### 3.2 `miic` and FCI executables

We provide also `miic` and FCI executables, which were used for all benchmarks included in the paper.

#### Directories and scripts

The main folder contains the scripts and source code for the reconstruction of networks from observational data.

The directories are organized as follows:

```
/
├── common
│   ├── miic.R
│   ├── gmPlot.R
│   └── gmSummary.R
├── sharedLib
├── data
│   └── some data input/output...
├── miic
│   └── all miic scripts and executables
└── fci
    └── all fci scripts
```

## Package requirements

To launch the `miic.R` script, R should be installed, along with some packages available in the CRAN repository.

**Rpackages** `getopt`, `plotrix`, `igraph`, `ppcor`, `bnlearn`, `pcalg`

## Calling the inference methods with `miic`

The inference methods can be called through the `miic.R` script.

### Overview

**main** `~/common/miic.R`

**lib** `~/common/lib/...`

### Arguments (*mandatory*: \*)

- i** \* file path of the input dataset<sup>1</sup>
- o** \* directory path for the output of the inference method<sup>2</sup>
- m** inference method (*miic*, *fci*)  
default: *miic*
- d** steps to perform<sup>3</sup> ('1,2,3,4' or '1,2' or '1,3' etc...)  
default: '1,2,3,4'
- p** parameters for the inference method (see the following subsections). The value expected here is of type character: '*param<sub>1</sub>:value<sub>1</sub>,param<sub>2</sub>:value<sub>2</sub> etc...*'
- t** file path to the true edges; used during the *summary* step<sup>4</sup>
- l** file path to the layout of each vertex; used during the *plot* step<sup>5</sup>
- c** if given, edges will be filtered according to their confidence ratio. It needs two parameters, described in the part "Option '-c' for **miic**".
- s** if given, this file provides an ordering of each variable categories for the calculation of the sign of the edges. The signs are calculated using Spearman's partial correlation coefficient. The ordering file is automatically generated if not provided by the user and can be edited to properly reorder the variable categories. See an example of `stateOrder` file in the data directory.

---

<sup>1</sup>The input dataset should be a tab separated table, with column names but no row names. Missing values should be indicated with *NA*. Each column corresponds to a categorized variable and each row to one sample.

<sup>2</sup>To prevent from overwriting existing results, if the output directory already exists, the skeleton inference step returns a message and stops.

<sup>3</sup>(1) skeleton, (2) probabilistic orientation, (3) summary, (4) plot

<sup>4</sup>The true edges file has two space-separated columns. Each line corresponds to one true edge. The orientation is *col1* → *col2*.

<sup>5</sup>The layout file has three tab separated columns, the first column being optional. Each line corresponds to the (*x*, *y*) coordinates of each vertex. The first column can contain the label of the vertex as indicated in the *colnames* of the input dataset table. The order in which the coordinates are given also corresponds to the order of the *colnames* of the input dataset table.

A full example to run in ‘common’ directory:

```
Rscript miic.R -i ../data/alarm1000samples.txt -o ../data/alarmNetwork -m miic -c csh:100,ccr:0.01  
-p cpx:nml,efn:1000,lat:yes,prg:yes -l ../data/alarmLayout.txt -s ../data/alarmStateOrder.tsv
```

When calling the available inference methods with *miic.R*, the ‘p’ option can be used to indicate the chosen parameters. The value expected for this option is of type character: ‘*param<sub>1</sub>:value<sub>1</sub>,param<sub>2</sub>:value<sub>2</sub> etc...*’. The possible *param<sub>i</sub>* and *value<sub>i</sub>* for each method are detailed in the following subsections.

### Option ‘-p’ for **miic**

**cpx** formula used to compute the complexity term [‘mdl’<sup>6</sup> or ‘nml’<sup>7</sup>]  
default: nml (Ex.: -p ‘...,cpx:mdl,...’)

**lat** should the network be reconstructed under the hypothesis that some variables might not be observed?  
[‘yes’ or ‘no’]  
default: no (Ex.: -p ‘...,lat:yes,...’)

**prg** should the network be oriented using the propagation rule? [‘yes’ or ‘no’]  
default: yes (Ex.: -p ‘...,prg:yes,...’)

**efn** number of uncorrelated samples  
default: number of rows of the input dataset (Ex.: -p ‘...,efn:1000,...’)

A ‘-p’ example: -p cpx:mdl,efn:1000

### Option ‘-p’ for **FCI**

**aph**  $\alpha$  significance level for the statistical independence tests<sup>8</sup>  
default: 0.01 (Ex.: -p ‘...,aph:0.001,...’)

**cit** type of statistical independence test [‘gaussCItest’, ‘dsepTest’, ‘disCItest’, ‘binCItest’]  
default: gaussCItest (Ex.: -p ‘...,cit:disCItest,...’)

**skm** choose the original PC or the order-independant version [‘stable’, ‘original’, ‘stable.fast’]  
default: stable (Ex.: -p ‘...,skm:original,...’)

**typ** which type of fci variant should be used<sup>9</sup> [‘normal’, ‘rapid’, ‘adaptive’, ‘anytime’ ]  
default: normal (Ex.: -p ‘...,typ:rapid,...’)

**con** set the state of the conservative rule [‘TRUE’, ‘FALSE’]  
default: FALSE (Ex.: -p ‘...,con:TRUE,...’)

<sup>6</sup>Minimum Description Length or Bayesian Information Criterion (BIC) [5, 6]

<sup>7</sup>Normalized Maximum Likelihood criterion [3, 7–10]

<sup>8</sup>refer to the documentation of the R package ‘pcalg’ for details

<sup>9</sup>RFCI corresponds to ‘rapid’

**maj** set the state of the majority rule [*TRUE*, *FALSE*]  
default: *FALSE* (Ex.: *-p ...,maj:TRUE,...*)

A full '-p' example: *-p aph:0.000001,cit:disCitest,skm:stable,typ:normal,con:FALSE,maj:FALSE*

### Option '-c' for **miic**

**csH** number of random shuffling of the input dataset, in order to get the random mutual information between **miic** inferred edges (Ex.: *-c ...,csH:100,...*)

**ccr** confidence ratio used as a threshold for filtering the edges. (Ex.: *-c ...,ccr:0.01,...*)

A full '-c' example: *-c csH:100,ccr:0.01*

### Viewing inferred networks

The inferred networks can be viewed either in pdf format (automatically generated with igraph <http://igraph.org/>) or with interactive graphml format for better display using cytoscape (<http://www.cytoscape.org/>). The files are located in the following directories:

- Unfiltered network: 'edgesList.miic.summary.plot.confidence.pdf' (only with the -m miic option), 'edgesList.miic.summary.plot\_pCor.pdf' and graphml '[output name].graphml' files are located in the output directory set by the -o entry in the command line
- Filtered networks (using -c option with **miic**): pdf and graphml files are in the subdirectory 'shuffle\_[csHValue]/filtered\_network\_[ccrValue]', which can be found in the output directory set by the -o entry in the command line. The output folder name in this case is 'filtered\_network\_[ccrValue]'.

We recommend the utilization of Cytoscape tool, version 3.1.0 or later is available for Windows, Linux and OsX. Visualizing **miic** networks with Cytoscape requires to go through the following steps:

1. Import the network: File⇒Import⇒Network⇒File, and select the graphml file in the created output directory
2. Import the style: File⇒Import⇒Styles, and select the miic\_style.xml file present at the root of the provided source directory
3. Select the loaded style: under the Style panel present in Control Panel select the miic\_style

### Detailed information on output files

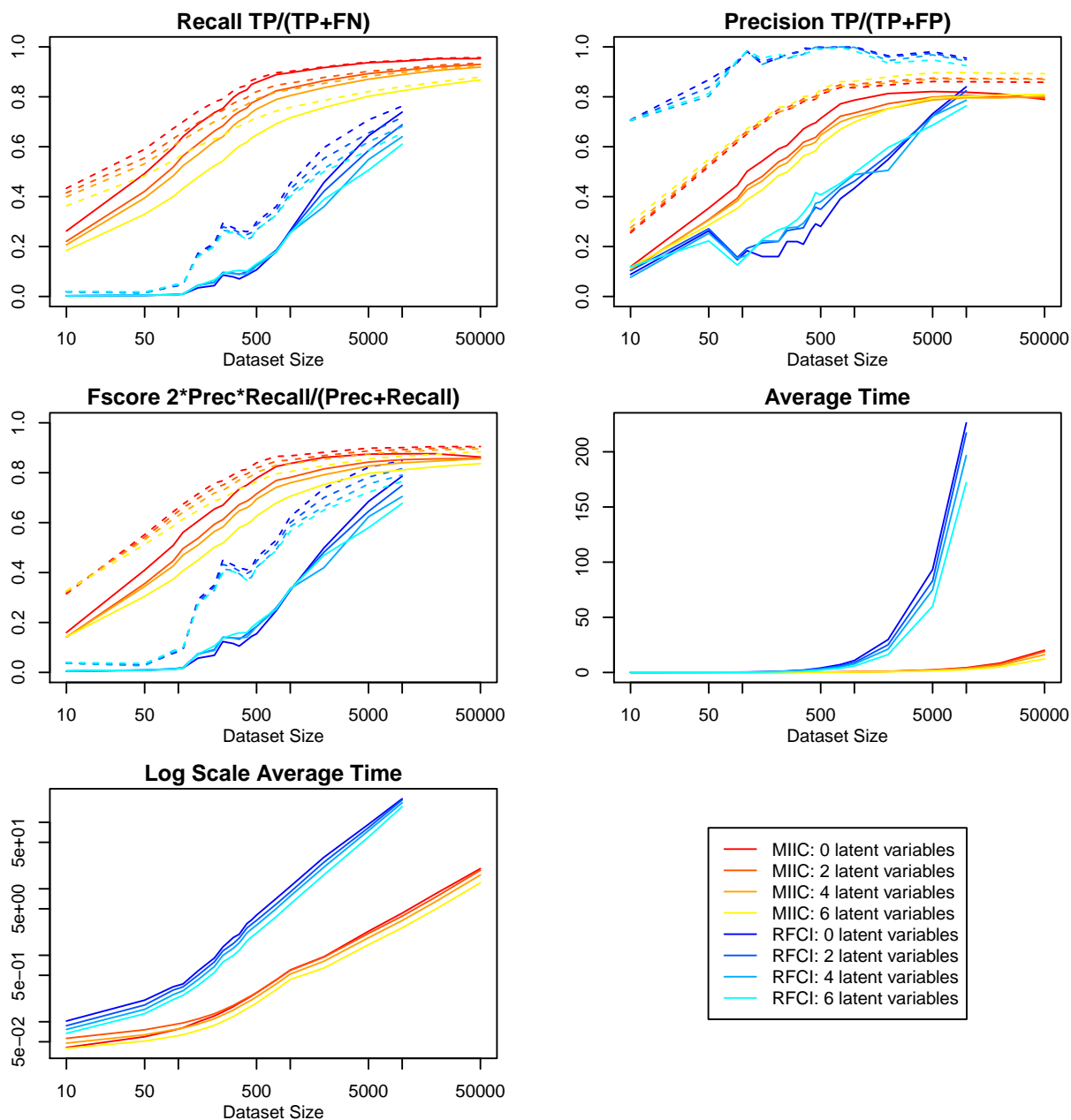
Additional information on output files can be found in the user manual.

## 4 References for Supplementary Text

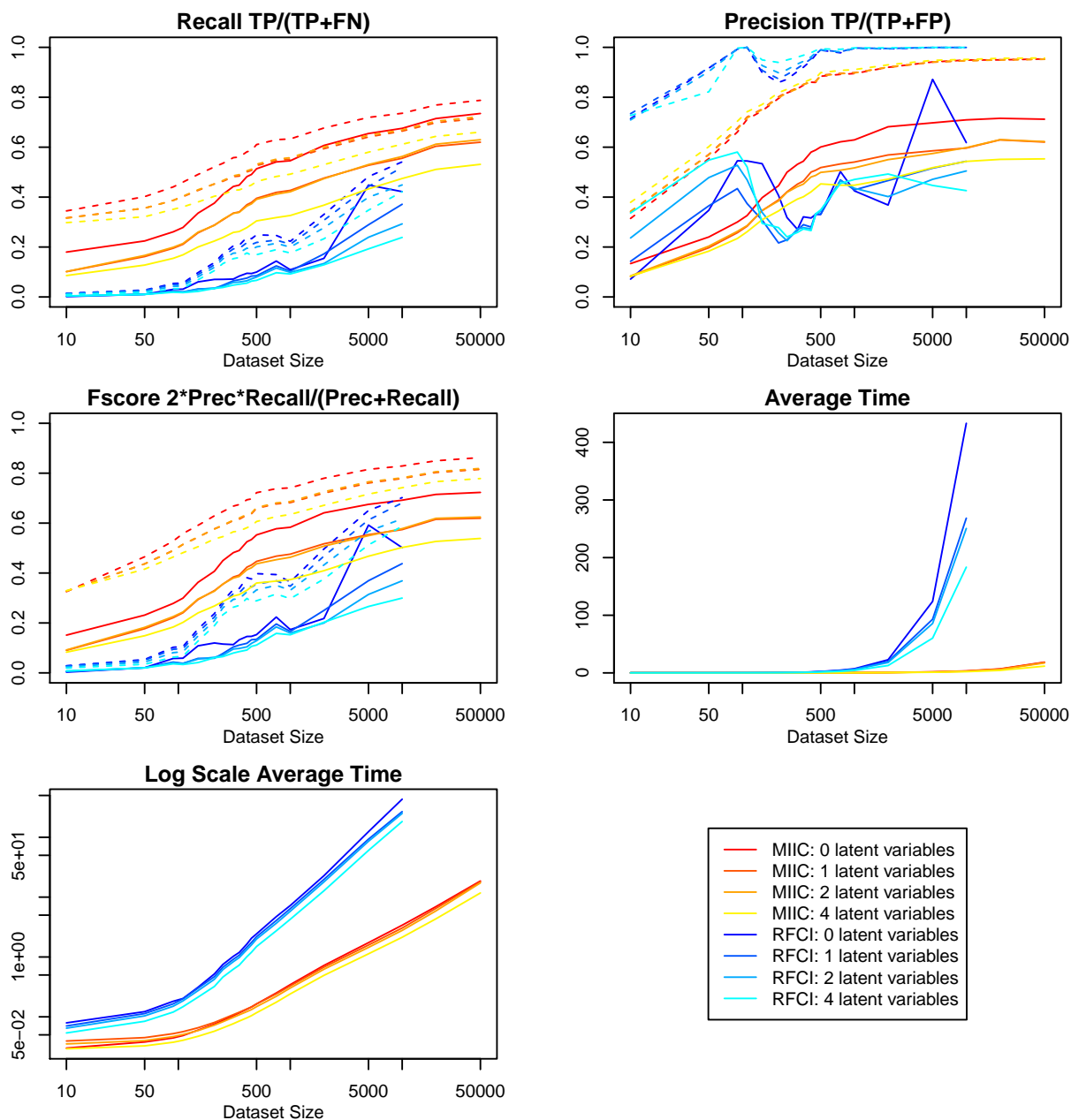
- [1] Richardson T, Spirtes P. Ancestral graph Markov models. *Ann Statist.* 2002;30(4):962–1030. doi:10.1214/aos/1031689015.
- [2] Affeldt S, Isambert H. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015*; 2015. p. 42–51.
- [3] Affeldt S, Verny L, Isambert H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics.* 2016;17(S2).
- [4] Sanov IN. On the probability of large deviations of random variables. *Mat Sbornik.* 1957;42:11–44.
- [5] Rissanen J. Modeling by shortest data description. *Automatica.* 1978;vol. 14:465–471.
- [6] Hansen MH, Yu B. Model Selection and the Principle of Minimum Description Length. *J Am Stat Assoc.* 2001;96:746–774.
- [7] Shtarkov YM. Universal sequential coding of single messages. *Problems of Information Transmission.* 1987;23(3):3–17.
- [8] Rissanen J, Tabus I. Kolmogorovs Structure function in MDL theory and lossy data compression. In: *Adv. Min. Descrip. Length Theory Appl.* MIT Press; 2005. p. 245–264.
- [9] Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf Process Lett.* 2007;103(6):227–233.
- [10] Roos T, Silander T, Kontkanen P, Myllymäki P. Bayesian network structure learning using factorized NML universal models. In: *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press; 2008.
- [11] Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif Intell.* 2008;172(16-17):1873–1896. doi:10.1016/j.artint.2008.08.001.
- [12] Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann Statist.* 2012;40(1):294–321. doi:10.1214/11-aos940.
- [13] Spirtes P, Meek C, Richardson T. An Algorithm for causal inference in the presence of latent variables and selection bias. In: *Computation, Causation, and Discovery*. Menlo Park, CA: AAAI Press; 1999. p. 211–252.
- [14] Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review.* 1991;9:62–72. doi:10.1177/089443939100900106.
- [15] Pearl J, Verma T. A theory of inferred causation. In: *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.*; 1991. p. 441–452.
- [16] Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. *J Stat Softw.* 2012;47(11):1–26. doi:10.18637/jss.v047.i11.

- [17] Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol.* 2015;33(3):269–276. doi:10.1038/nbt.3154.
- [18] Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 2015;43(D1):D805–D811. doi:10.1093/nar/gku1075.

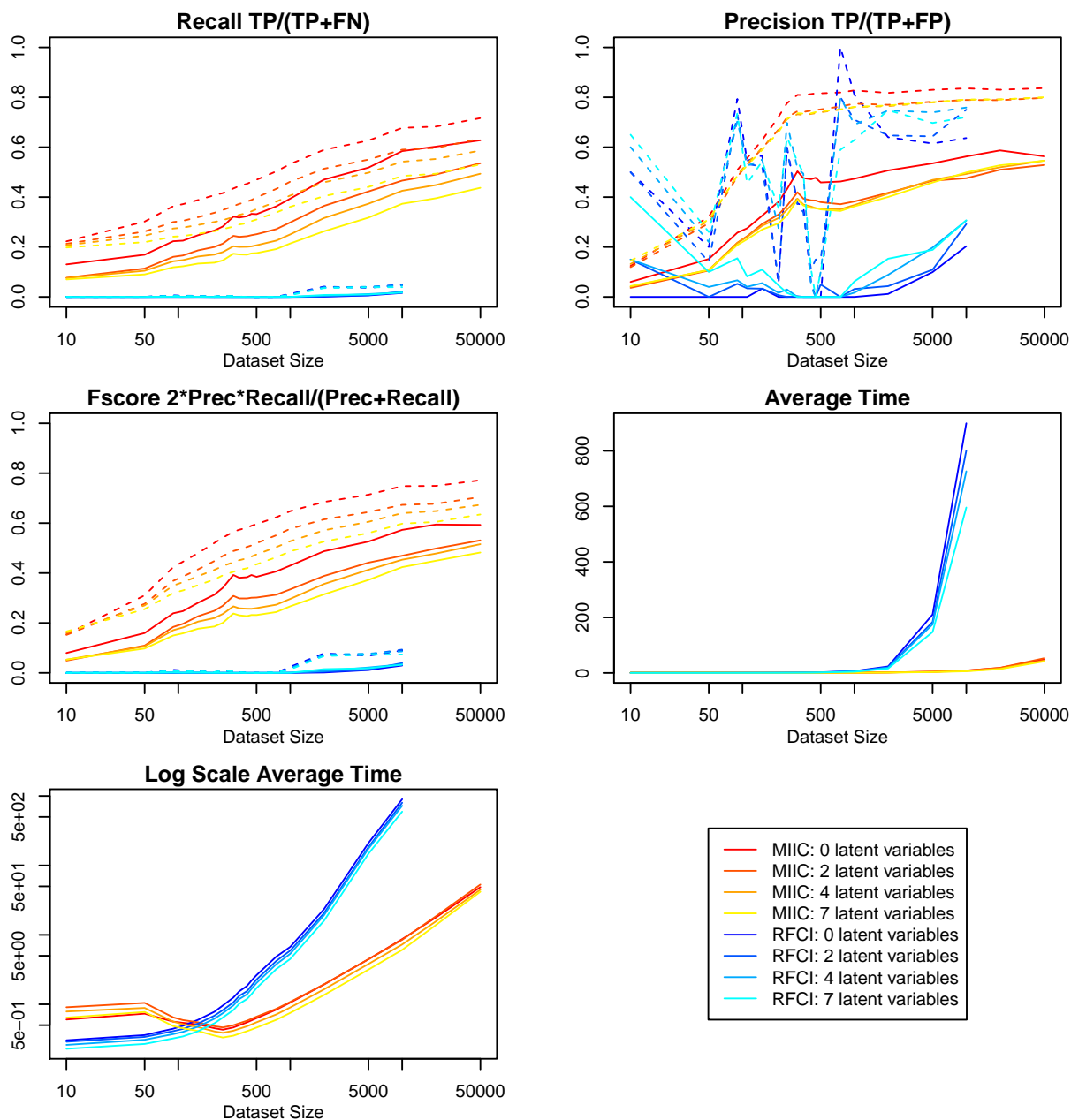
## 5 Supplementary Figures



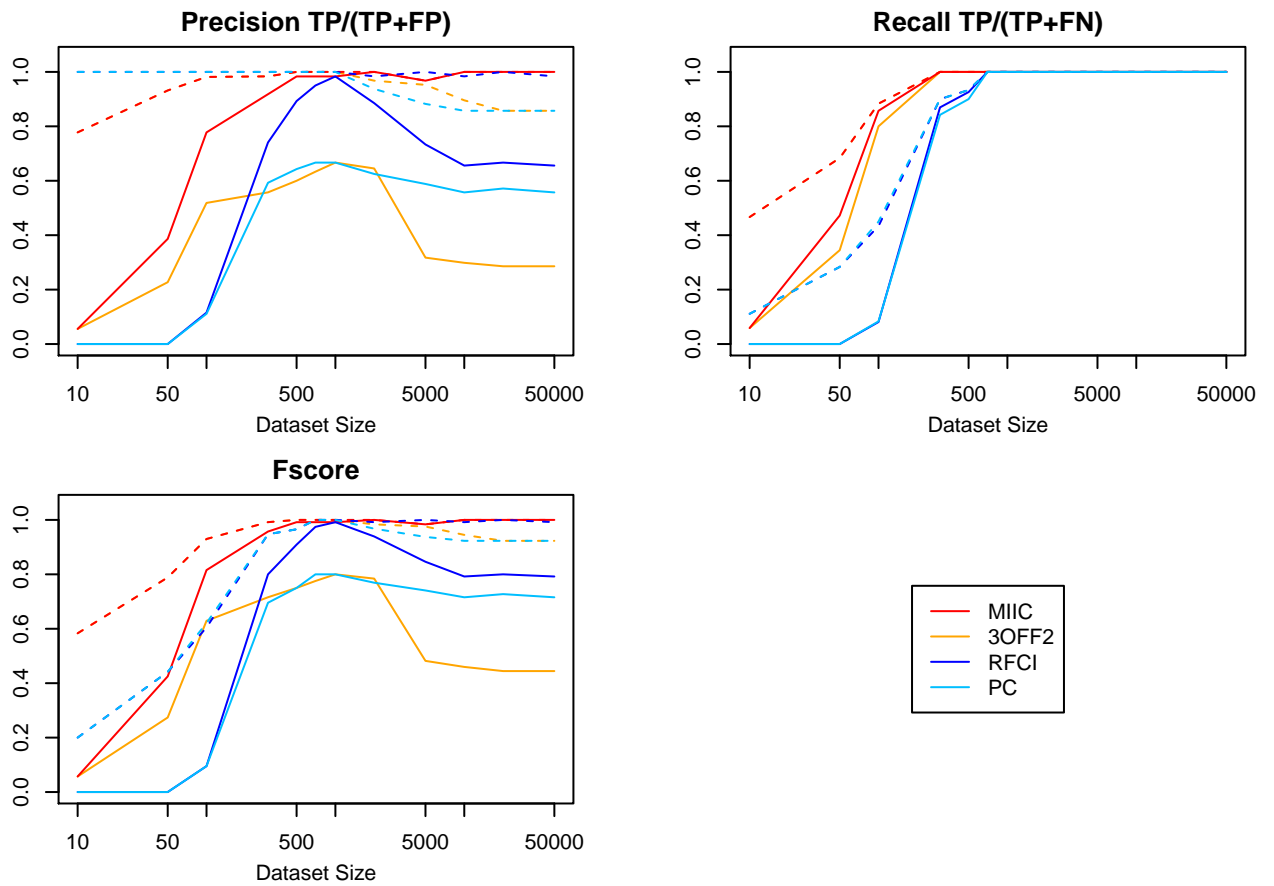
**S1 Fig. Real-life Alarm network with hidden latent variables** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [12] (cold colors) for 0, 2, 4 and 6 latent variables out of the 37 nodes. Computation times in log scale show a linear scaling in the limit of large datasets,  $\tau_{\text{cpu}} \sim N^{0.9}$ , for the `miic` algorithm, and a stronger nonlinear increase,  $\tau_{\text{cpu}} \sim N^{1.5}$ , with the `RFCI` algorithm.



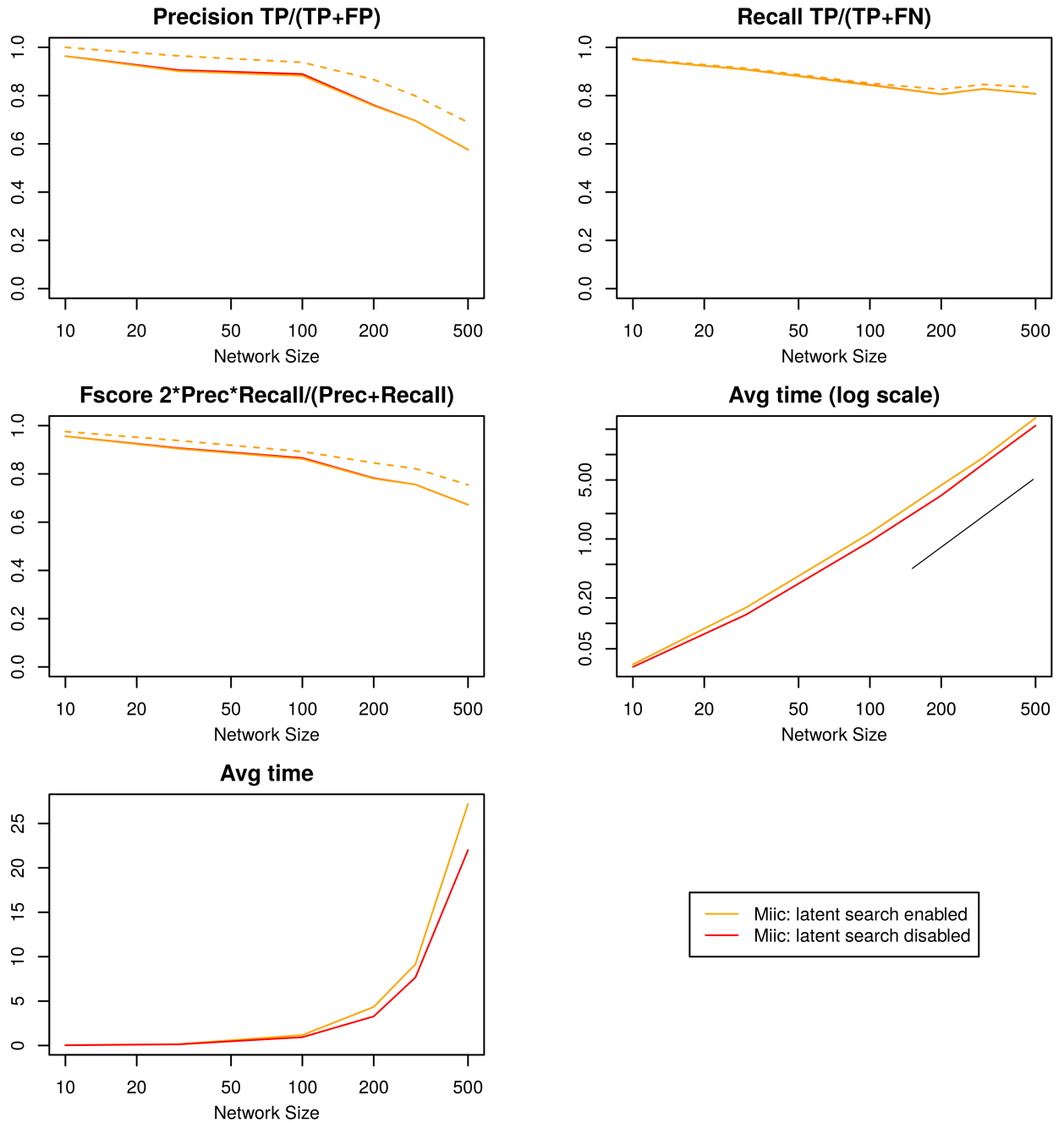
**S2 Fig. Real-life Insurance network with hidden latent variables** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [12] (cold colors) for 0, 1, 2, and 4 latent variables out of the 27 nodes. Computation times in log scale show a linear scaling in the limit of large datasets,  $\tau_{\text{cpu}} \sim N^{1.0}$ , for the `miic` algorithm, and a stronger nonlinear increase,  $\tau_{\text{cpu}} \sim N^{1.7}$ , with the `RFCI` algorithm.



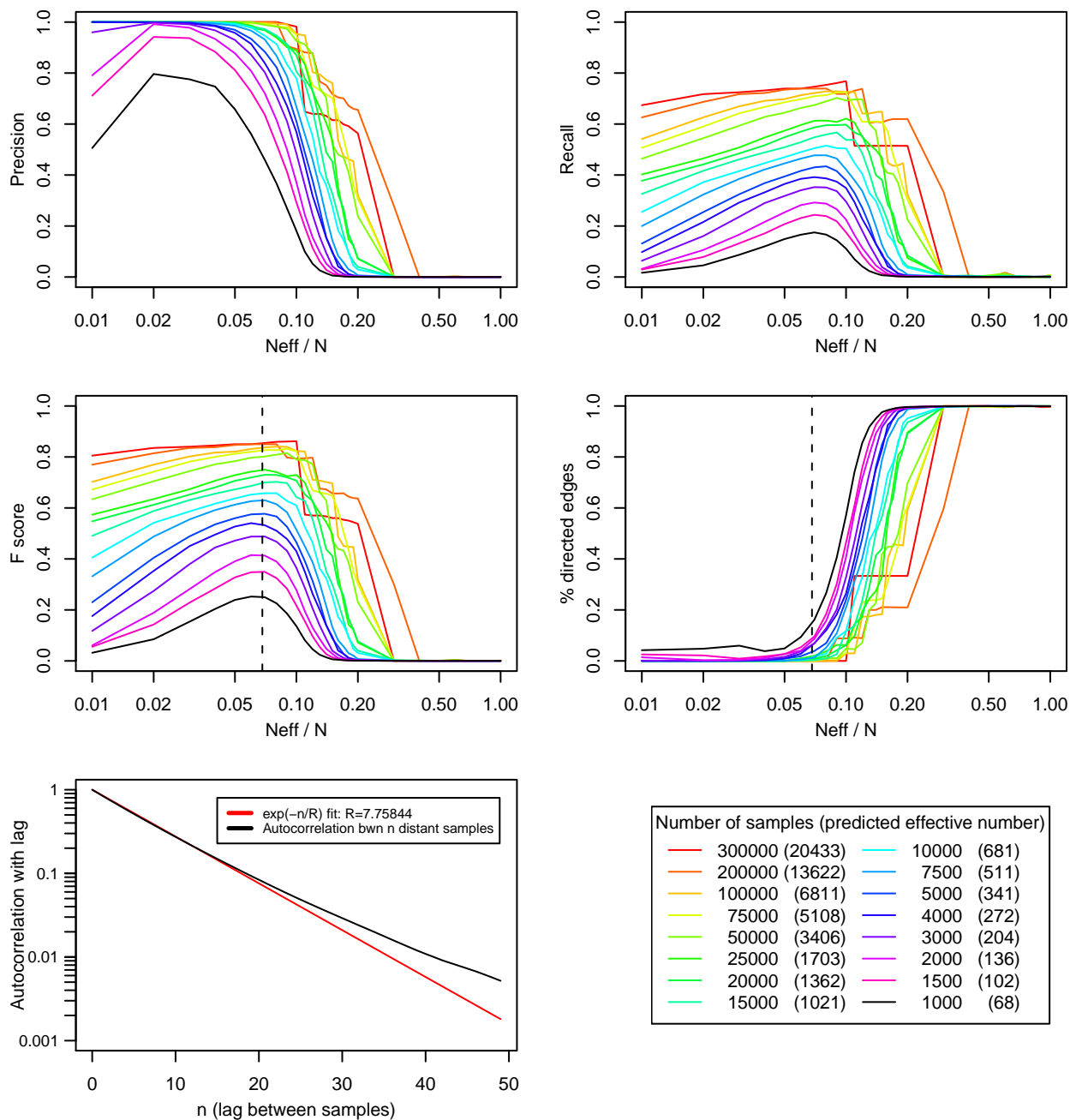
**S3 Fig. Real-life Barley network with hidden latent variables** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the *miic* algorithm (warm colors) compared to the RFCI algorithm [12] (cold colors) for 0, 2, 4 and 7 latent variables out of the 48 nodes. Computation times in log scale show a nearly linear scaling in the limit of large datasets,  $\tau_{\text{cpu}} \sim N^{1.1}$ , for the *miic* algorithm, and a stronger nonlinear increase,  $\tau_{\text{cpu}} \sim N^{2.3}$ , with the RFCI algorithm.



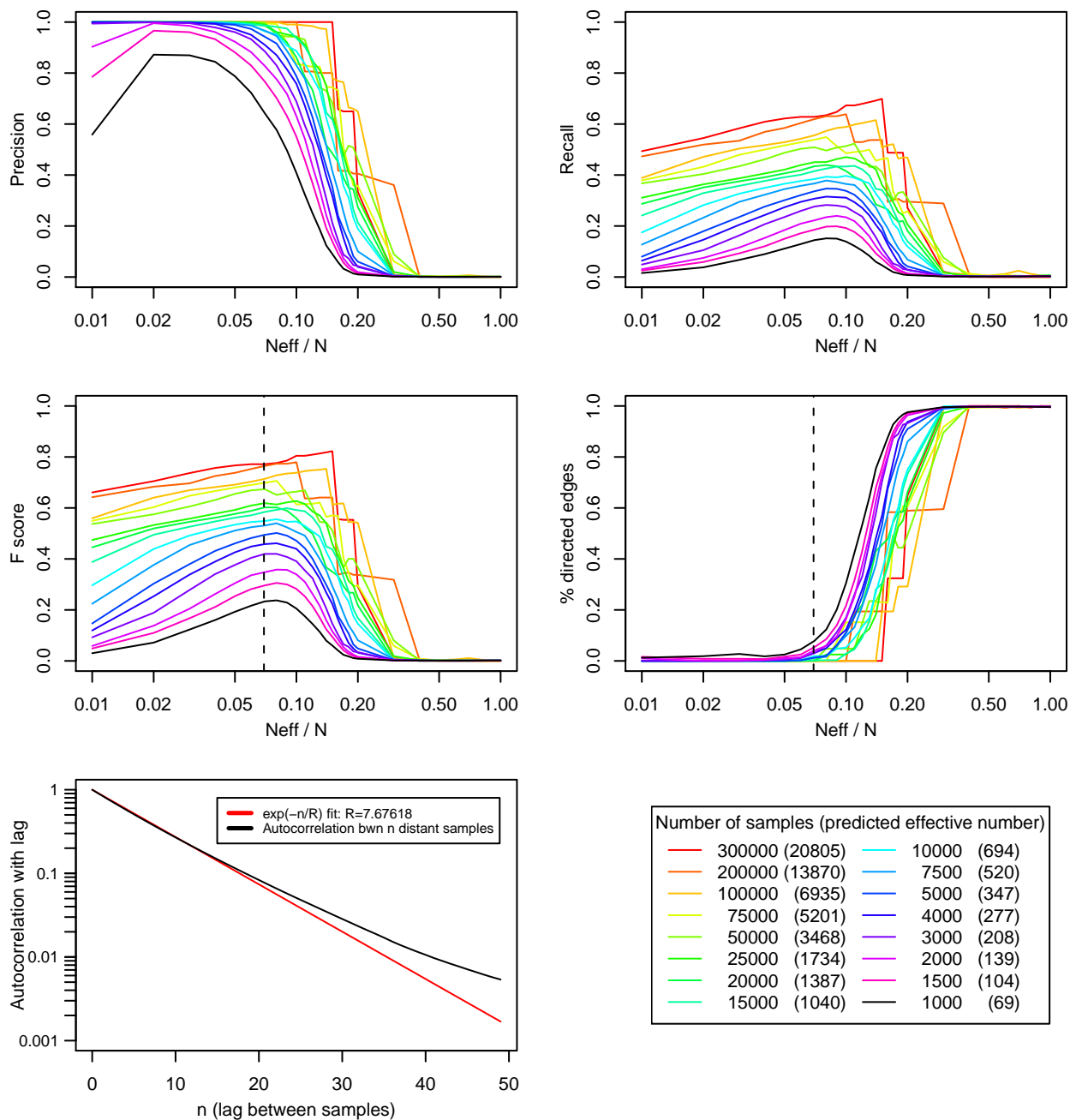
**S4 Fig. Reconstruction of Fig. 1C network from simulated data.** *miic* and RFCI [12, 13] versus *3off2* [3] and PC [14–16] reconstructions of Fig. 1C network are performed from simulated data generated with Tetrad V,  $N = 10 - 50,000$  samples. Precision, Recall and Fscore are given for skeleton (dashed lines) and PAG including orientations (solid lines).



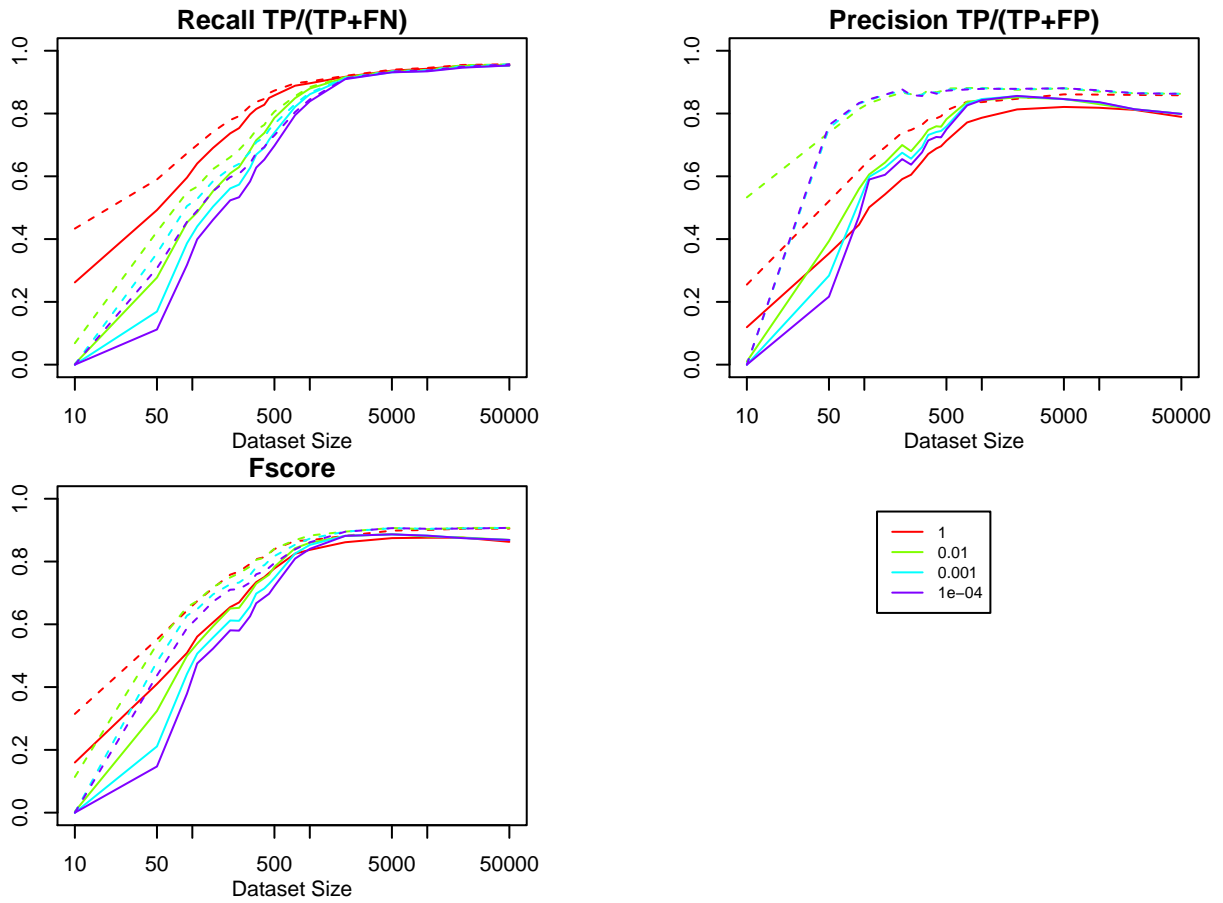
**S5 Fig. Random benchmark networks of increasing size.** `miic` reconstruction of random networks of increasing size ( $P = 10 - 500$  nodes) and fixed average degree 3 from  $N = 1,000$  samples generated with Tetrad V. The average CPU time exhibits an optimal quadratic complexity in terms of network size,  $\tau_{\text{cpu}} \sim P^2$  (solid bar), with only a small time increase when considering latent variables (orange) as compared to excluding them (red).



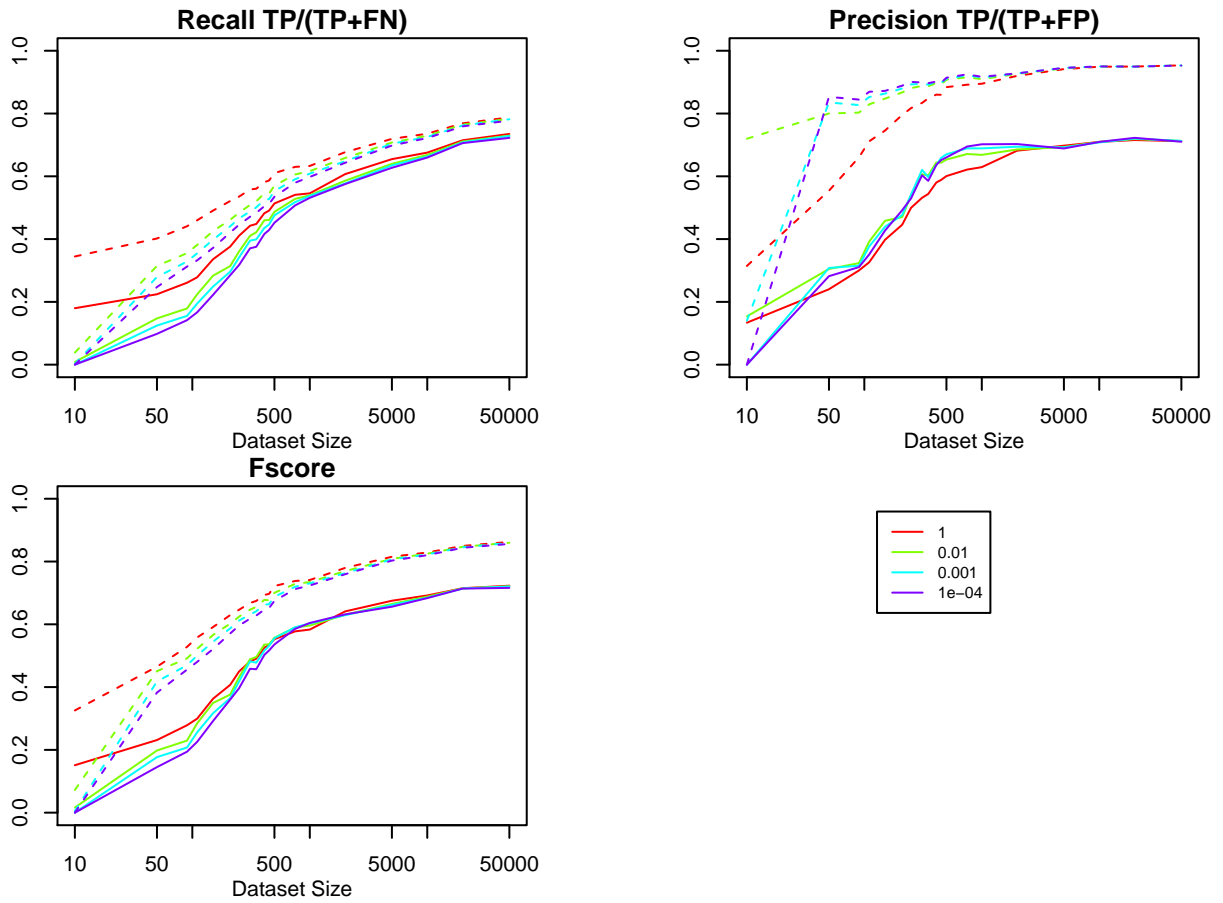
**S6 Fig. Alarm-like undirected network.** Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for  $N = 1,000$  to  $300,000$  consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples  $N_{\text{eff}}^*/N \simeq 0.068$ , see Materials and Methods.



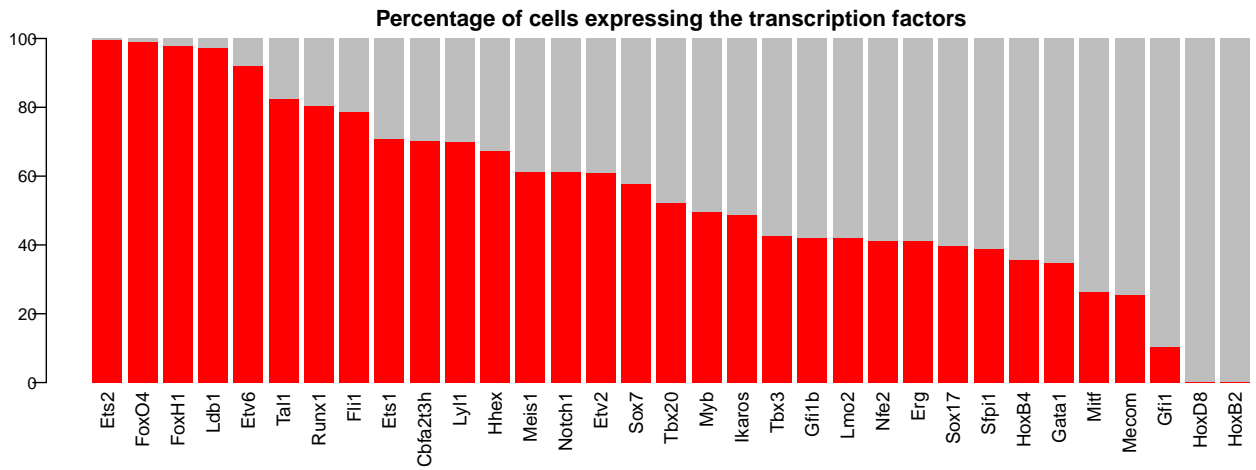
**S7 Fig. Insurance-like undirected network.** Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for  $N = 1,000$  to  $300,000$  consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples  $N_{\text{eff}}^*/N \simeq 0.069$ , see Materials and Methods.



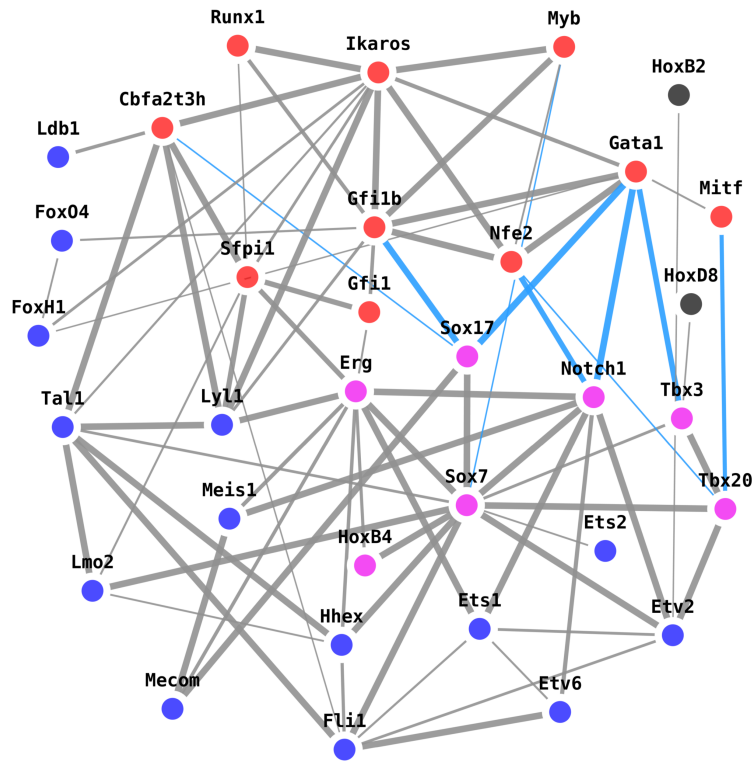
**S8 Fig. Edge confidence filtering on real-life Alarm network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing time for network skeleton (dashed lines) and oriented network CPDAG (solid lines) for a decreasing edge-specific confidence filtering,  $C_{XY} = 1$  (no filtering) 0.01, 0.001 and 0.0001. For sample size  $> 100$ , confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with  $C_{XY} < 10^{-3} - 10^{-2}$ , tends to yield equivalent F-scores as unfiltered benchmark reconstructions.



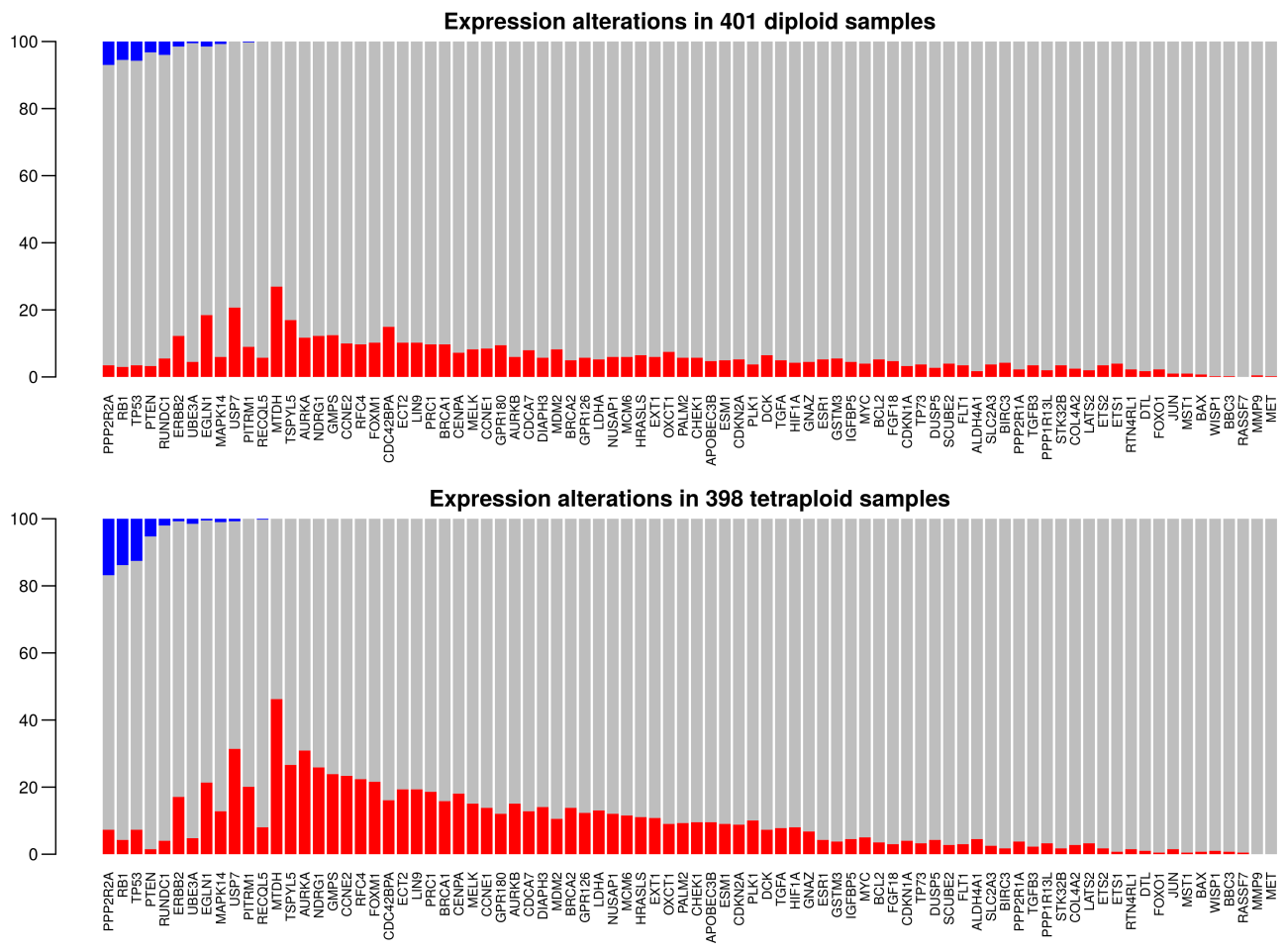
**S9 Fig. Edge confidence filtering on real-life Insurance network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for network skeleton (dashed lines) and oriented network CPDAG (solid lines) for a decreasing edge-specific confidence filtering,  $C_{XY} = 1$  (no filtering) 0.01, 0.001 and 0.0001. For sample size  $> 100$ , confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with  $C_{XY} < 10^{-3} - 10^{-2}$ , tends to yield equivalent F-scores as unfiltered benchmark reconstructions.



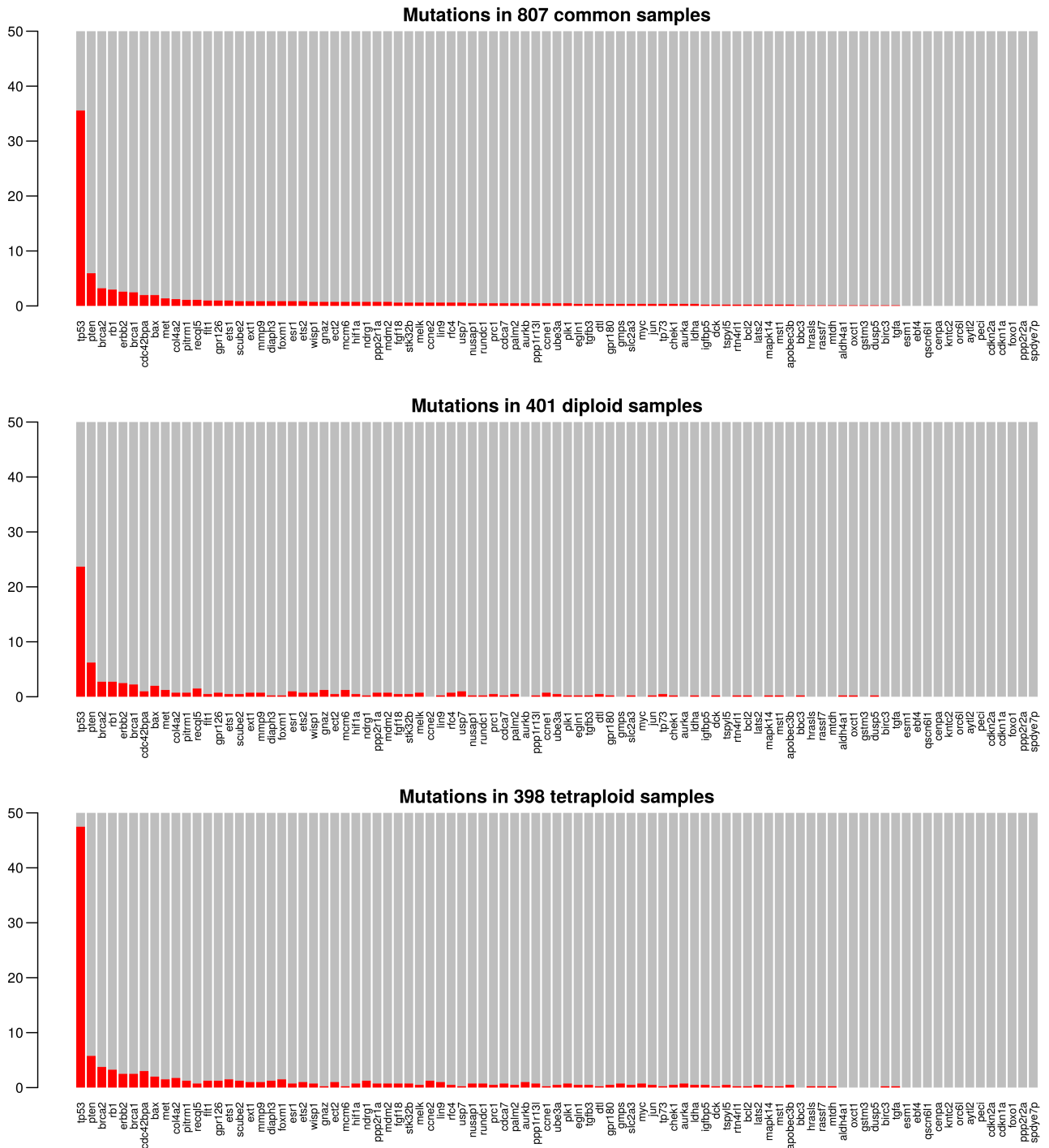
**S10 Fig. Gene expression distribution in 3,934 single cells from mouse embryos.** Expression data on the 33 TFs are obtained from [17]. Percentage of samples with expressed genes (red) and non-expressed genes (gray).



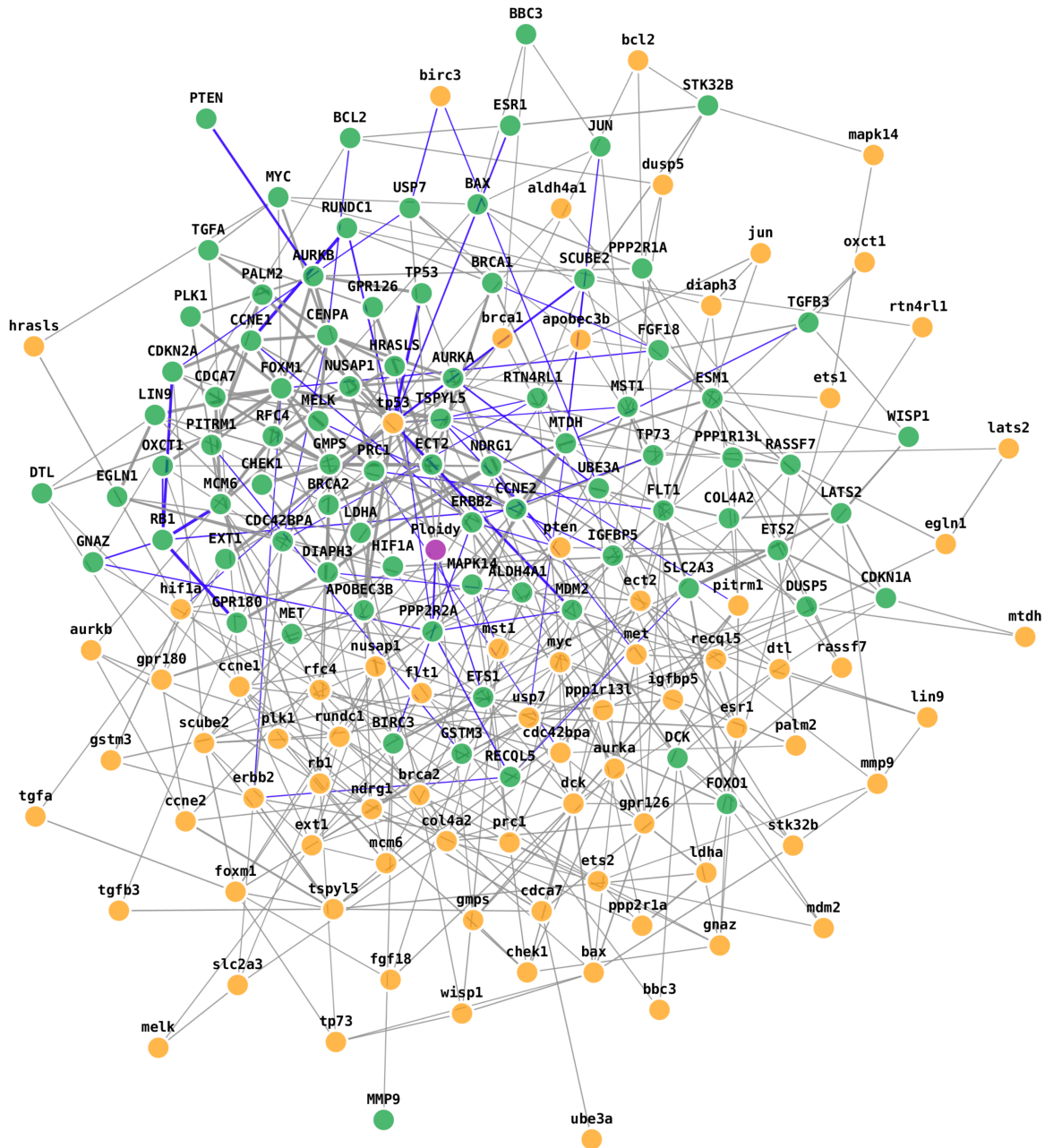
**S11 Fig. Unfiltered network skeleton for hematopoiesis differentiation data.** Hematopoietic / endothelial gene expression data in 3,934 single cells from mouse embryos [17]. 7 out of 82 edges (8.5%) with  $C_{XY} > 10^{-3}$  have been filtered in Fig. 2D (blue edges correspond to anti-correlations).



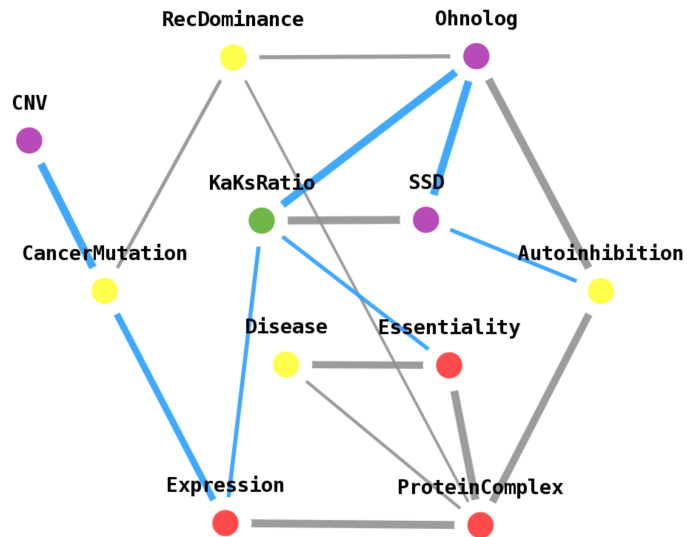
**S12 Fig. Expression alterations in 807 samples of breast tumor data from COSMIC database [18].** Percentage of samples with normalized over-expression (red), normalized under-expression (blue) and unchanged normalized expression (gray) for each gene based on COSMIC.



**S13 Fig. Mutations in 807 samples of breast tumor data from COSMIC database [18]. Percentage of mutated samples (red) for each gene.**



**S14 Fig. Unfiltered network skeleton for breast tumor ploidy-mutation- expression data from COSMIC database [18].** Due to the limited numbers of samples ( $N=807$ ) and recurrent gene mutants (Figure 3-figure supplement 2), most gene mutations (yellow) are not confidently linked to any altered expression levels (green) and have been filtered in the high confidence network Fig. 3D ( $C_{XY} < 10^{-3}$ ), with the notable exceptions of *TP53* and *RB1* mutations, which have a significant impact on gene expressions, Fig. 3D, see main text (blue edges correspond to anti-correlations).



**S15 Fig. Unfiltered network skeleton for ohnolog retention data in human.** Genomic data for the 20,415 human coding genes is provided in S1 Data. The only edge with confidence ratio  $C_{XY} > 10^{-3}$  is RecDominance — ProteinComplex with  $C_{XY} = 0.25$  (blue edges correspond to anti-correlations).