**Algorithm 2** Update Minimal I-MAP (*UMI*)

**Input:** Current permutation $\pi_i$, previous permutation $\pi_{i-1}$, previous minimal I-MAP $G_{\pi_{i-1}}$, significance level $\alpha$, data $D$
**Output:** $G_{\pi_i}$
$k$ = min index of adjacent transposition
**if** $k = 1$ (first and last element swapped) **then**
  Compute $\hat{G}_{\pi_i}$ from $\hat{\mathcal{O}}_n(D, \alpha)$
**else**
  $G_{\pi_i} = G_{\pi_{i-1}}$
  Reverse edge from $X_{\pi_i(k+1)}$ to $X_{\pi_i(k)}$ in $G_{\pi_i}$ if such an edge exists
  **for** $s = 1$ **to** $k - 1$ **do**
    **for** $j = k$ **to** $k + 1$ **do**
      $S = \{\pi(1), \cdots, \pi(j-1)\} \setminus \{\pi(s)\}$
      Let $z = \hat{\mathcal{O}}^{(n)}_{i,j|S}(D, \alpha)$
      Update edge from $X_{\pi(s)}$ to $X_{\pi(j)}$ to $z$ in $G_{\pi_i}$
    **end for**
  **end for**
**end if**

## A. CI Testing for Gaussian Data

In the case of multivariate Gaussian data, one may use the Fisher z-transform (Fisher, 1915) to perform CI testing. The Fisher z-transform is given by

$$Z(i, j \mid S) := \frac{1}{2} \frac{\log(1 + \hat{\rho}_{i,j|S})}{\log(1 - \hat{\rho}_{i,j|S})},$$

where $\hat{\rho}_{i,j|S}$ is the empirical partial correlation between $X_i$ and $X_j$ given $X_S$. To conduct a two-sided hypothesis test at significance level $\alpha$, one may test if

$$\sqrt{n - |S| - 3}|Z(i, j \mid S)| \leq \Phi^{-1}(1 - \alpha/2),$$

where $\Phi^{-1}$ is the inverse CDF of $N(0, 1)$.

## B. Update Algorithm

Algorithm 2 specifies the update procedure used in Algorithm 1 to reduce the number of CI tests needed.

## C. Discussion of the Assumptions

Based on the discussion of Kalisch & Buhlmann (2007), Assumption 3.1(b) is not such a strong assumption and seems more of a regularity condition needed to prove the bounds. Assumption 3.1(d) has an intuitive interpretation; it says that the best prediction of $G$ based on the data and order is captured by the constructed network. Conditioned on the order, the inference problem is not hard; i.e., we just need to recover the skeleton. Since we can recover the skeleton via the empirical CI relations, $\hat{G}_\pi$ is indeed the best prediction of the network given the data and order in many cases, which would imply that $\hat{G}_\pi$ can reasonably be assumed to be a sufficient statistic. Assumption 3.1(e) is a quite weak assumption; it says that the information of $\hat{G}_\pi$ does not help in predicting the probability of a CI error. This makes sense because we want to know if $\hat{G}_\pi$ does not equal $G^*_\pi$. But, without observing $G^*_\pi$, or conditioning on some property of $G^*_\pi$ in addition to $\hat{G}_\pi$, it seems reasonable to assume that our prediction is left unchanged when knowing $\hat{G}_\pi$.

## D. Proofs

### D.1. Proof of Lemma 3.3

The proof relies heavily on the concentration bounds used to prove the high-dimensional consistency of the PC algorithm (Kalisch & Buhlmann, 2007). To start, notice that

$$\begin{aligned}\mathbb{P}(G_\pi \neq \hat{G}_\pi \mid G, \theta) &= \mathbb{P}(\text{CI error(s) constructing } \hat{G}_\pi) \\ &\leq \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \mathbb{P}(E_{i,j}(G^*, \theta^*)), \quad (6)\end{aligned}$$

where $E_{i,j}(G^*, \theta^*)$ is the event that a CI error is made when testing $X_{\pi(i)} \not\perp\!\!\!\perp X_{\pi(j)} \mid X_S$, for $S = \{\pi(1), \cdots, \pi(j-1)\} \setminus \{\pi(i)\}$, conditioned on the Bayesian network $(G^*, \theta^*)$ generating the observed data. Note that these tests are performed at the significance level provided in the statement of the lemma.

By assumption, $Q^*_{\theta^*, G^*} \leq q^* < 1$ and $0 < r^* \leq R^*_{\theta^*, G^*}$ (without loss of generality, ignore measure zero sets). Picking such $q^*$ and $r^*$ then satisfy the assumptions required in Lemma 4 of Kalisch & Buhlmann (2007). Equations (16) and (17) from Kalisch & Buhlmann (2007) imply that there exist constants $C_1, C_2$ that depend *only* on $q^*$ such that

$$\mathbb{P}(E_{i,j}(G^*, \theta^*)) \leq C_1(n-p)\exp\{-C_2(r^*)^2(n-p)\}$$

for any $i, j$. Hence,

$$\mathbb{P}(G^*_\pi \neq \hat{G}_\pi \mid G, \theta) \leq f(n, p), \quad (7)$$

where $f(n, p) = p^2 C_1(n-p)\exp\{-C_2(r^*)^2(n-p)\}$.

Now,

$$\begin{aligned}&\mathbb{P}(G^*_\pi \neq \hat{G}_\pi) \\ &= \sum_{G \in \mathcal{G}} \int_\theta \mathbb{P}(G^*_\pi \neq \hat{G}_\pi \mid G, \theta)\mathbb{P}(\theta \mid G)\mathbb{P}(G)d\theta \\ &\leq \sum_{G \in \mathcal{G}} \int_\theta f(n, p)\mathbb{P}(\theta \mid G)\mathbb{P}(G)d\theta \\ &= f(n, p),\end{aligned}$$

as desired.

### D.2. Proof of Theorem 3.2

By the tower property,

$$\mathbb{E}_{\mathbb{P}(G|D)} f(G) = \mathbb{E}_{\mathbb{P}(\pi|D)} \mathbb{E}_{\mathbb{P}(G|D,\pi)} f(G).$$

As before, define $A_\pi$ as the event that $\hat{G}_\pi = G^*_\pi$. We may expand $\mathbb{E}_{\mathbb{P}(G|D,\pi)} f(G)$ as

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(G|D,\pi)} f(G) \\
&= \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G \mid D, \pi) \\
&= \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G, A_\pi \mid \hat{G}_\pi) + \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G, A_\pi^C \mid \hat{G}_\pi) \\
&\quad \text{by Assumption 3.1(d) and the law of total probability} \\
&= f(\hat{G}_\pi) + \mathbb{P}(A_\pi^C \mid \hat{G}_\pi) \\
&\quad \cdot \left[ \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G \mid \hat{G}_\pi, A_\pi^C) - f(\hat{G}_\pi) \right] \\
&\quad \text{by the fact that } \mathbb{P}(G \mid \hat{G}_\pi, A_\pi) = I(G = \hat{G}_\pi) \\
&\quad \text{according to the exact reasoning used in Section 4} \\
&= f(\hat{G}_\pi) + \mathbb{P}(A_\pi^C) \\
&\quad \cdot \left[ \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G \mid \hat{G}_\pi, A_\pi^C) - f(\hat{G}_\pi) \right],
\end{aligned}$$

where the final equality uses Assumption 3.1(e).

We claim that

$$\mathbb{E}_{\mathbb{P}(\pi|D)} f(\hat{G}_\pi) = \mathbb{E}_{\hat{\mathbb{P}}(G|D)} f(G). \tag{8}$$

To prove Equation (8), notice that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(\pi|D)} f(\hat{G}_\pi) \\
&= \sum_{\pi \in S_p} f(\hat{G}_\pi) \mathbb{P}(\pi \mid D) \\
&= \sum_{G \in \mathcal{G}} f(G) \sum_{\pi \in S_p} \mathbb{1}\{G \in \hat{\mathcal{G}}\} \mathbb{1}\{G = \hat{G}_\pi\} \mathbb{P}(\pi \mid D) \\
&= \sum_{G \in \mathcal{G}} f(G) \hat{\mathbb{P}}(G \mid D) \\
&= \mathbb{E}_{\hat{\mathbb{P}}(G|D)}[f(G)].
\end{aligned}$$

Finally,

$$\left| \sum_{G \in \mathcal{G}} f(G) \mathbb{P}(G \mid \hat{G}_\pi, A_\pi^C) - f(\hat{G}_\pi) \right| \leq 2M$$

and

$$\mathbb{P}(A_\pi^c) \leq C_1 (n-p) \exp\{-C_2 (r^*)^2 (n-p)\}$$

by Lemma 3.3. The result now follows by taking expectations and using the above bounds.

### D.3. Proof of Proposition 5.1

Ergodicity follows from the fact that any permutation can be reached from adjacent transpositions, and aperiodicity follows from our constraint that $s \in (0, 1)$. Since adjacent transpositions trivially satisfy the detailed balance equations, the Markov chain has stationary distribution $\mathbb{P}(\hat{G}_\pi \mid D)$.

### D.4. Proof of Proposition D.1

**Proposition D.1.** *If $\pi_t$ and $\pi_{t+1}$ differ by an adjacent transposition, Algorithm 2 correctly calculates $\hat{G}_{\pi_{t+1}}$ from $\hat{G}_{\pi_t}$.*

This update rule was also used by Solus et al. (2017). We here provide the proof for completeness. The result trivially follows if $\pi_{t+1}$ is obtained by swapping the first and last element of $\pi_t$ since all CI tests are recomputed in this case. Hence, we may assume $\pi_t$ and $\pi_{t+1}$ differ by an adjacent transposition not at the border. Suppose $\pi_t = (n_1 \cdots n_i n_{i+1} \cdots n_p)$ and $\pi_{t+1} = (n_1 \cdots n_{i+1} n_i \cdots n_p)$, where the permutations differ at an adjacent permutation at position $i$. Then, the only edges that can be different in $\hat{G}_{\pi_t}$ and $\hat{G}_{\pi_{t+1}}$ are those edges connected nodes $n_i$ / $n_{i+1}$ with nodes $n_k, 1 \leq k < i$. Correcting the edges $(n_i, n_k)$ and $(n_{i+1}, n_k)$ corresponds to recomputing the conditional independence statements $X_{n_i} \not\perp\!\!\!\perp X_{n_k} \mid X_{S_i}$ and $X_{n_{i+1}} \not\perp\!\!\!\perp X_{n_k} \mid X_{S_{i+1}}$, for $X_{S_i} = \{n_1, \cdots, n_{i+1}\} \setminus \{n_k\}$ and $X_{S_{i+1}} = \{n_1, \cdots, n_{i-1}\} \setminus \{n_k\}$ and updating the corresponding edges. The for loop in Algorithm 2 carries out the CI tests specified in the previous sentence. Finally, we need to reverse the edge between nodes $X_{n_i}$ and $X_{n_{i+1}}$ if there was an edge between them in the old DAG $\hat{G}_{\pi_t}$; this reversal is accomplished at the very start of Algorithm 2.

### D.5. Proof of Proposition 5.2

The memory complexity follows trivially from the fact that it takes $\mathcal{O}(p^2)$ memory to store $\hat{G}_\pi$ in an adjacency matrix. Computing partial correlations takes at most $\mathcal{O}(p^3)$ time using the well-known partial correlation recursive formula (Viertl, 2011). Instantiating $\hat{G}_{\pi_0}$ requires $\mathcal{O}(p^2)$ CI tests and hence takes at most $\mathcal{O}(p^5)$ time to compute. The subsequent $\hat{G}_{\pi_i}$ are computed using Algorithm 2. The correctness of Algorithm 2 was shown in Appendix D.4. We claim Algorithm 2 takes average case $\mathcal{O}(p^4)$ time.

First, we show that the first and last elements of $\pi_i$ are swapped with probability less than $\frac{1}{p}$ when moving from $\pi_i$ to $\pi_{i+1}$. Notice from our definition of the adjacent transposition distribution $q$ that the probability of either the first or last element undergoing an adjacent transposition is $\frac{2(1-s)}{p}$. Conditioned on either the first or last element being chosen to be swapped, there is probability $\frac{1}{2}$ that the first (last) element will be swapped with the last (first) element. Hence, the probability of the first and last element being swapped

equals $\frac{(1-s)}{p}$ which is less than $\frac{1}{p}$. When the first and last element are swapped, all $p^2$ CI tests need to be recomputed. All the remaining adjacent transpositions require at most $2p$ additional CI tests to be performed in the for loop of Algorithm 2. Hence, on average, the number of additional CI tests is $\mathcal{O}(p)$ which implies the average running time of Algorithm 2 is $\mathcal{O}(p^4)$.

## E. Justification for restricting the prior space

Following nearly the same reasoning used to motivate our likelihood approximation in Section 4, here we justify

$$\mathbb{P}(\pi \mid \hat{\mathcal{O}}_n) \approx \mathbb{P}(G = \hat{G}_\pi).$$

Notice that

$$\begin{aligned}
&\mathbb{P}(\pi \mid \hat{\mathcal{O}}_n) \\
&= \mathbb{P}(\pi \mid \hat{\mathcal{O}}_n, A_\pi)\mathbb{P}(A_\pi|\hat{\mathcal{O}}_n) + \mathbb{P}(\pi \mid \hat{\mathcal{O}}_n, A_\pi^C)\mathbb{P}(A_\pi^C|\hat{\mathcal{O}}_n) \\
&= \mathbb{P}(\pi \mid \hat{\mathcal{O}}_n, A_\pi)\mathbb{P}(A_\pi) + \mathbb{P}(\pi \mid \hat{\mathcal{O}}_n, A_\pi^C)\mathbb{P}(A_\pi^C),
\end{aligned}$$

where the final equality follows from Assumption 3.1(e). We claim that

$$\mathbb{P}(\pi \mid \hat{\mathcal{O}}_n, A_\pi) = \mathbb{P}(G = \hat{G}_\pi). \tag{9}$$

Given $\mathcal{O}_n$, we can construct $\hat{G}_\pi$, and conditioned on $A_\pi$, $\hat{G}_\pi = G_\pi^*$. Each permutation $\pi$ may therefore be associated with its true corresponding DAG $G_\pi^*$ which equals $\hat{G}_\pi$. Hence, the conditional probability $\mathbb{P}(\pi|\hat{\mathcal{O}}_n, A_\pi)$ equals the prior probability of $\hat{G}_\pi$, namely $\mathbb{P}(G = \hat{G}_\pi)$.

Finally, since $\mathbb{P}(A_\pi)$ goes to zero exponentially fast by Lemma 3.3, $\mathbb{P}(\pi|\hat{\mathcal{O}}_n)$ is well approximated by $\mathbb{P}(G = \hat{G}_\pi)$.

## F. Prior Specification on Topological Orderings

Here we illustrate the computational difficulty of specifying a posterior $\mathbb{P}(\pi|D)$ that agrees with our original prior $\mathbb{P}^*(G)$ and likelihood $\mathbb{P}(G|D)$ on the space of DAGs. Notice that

$$\mathbb{P}(D \mid \pi) = \sum_G \mathbb{P}(D \mid G)\mathbb{P}(G \mid \pi). \tag{10}$$

Equation (10) implies that we must specify a conditional distribution $\mathbb{P}(G|\pi)$ to calculate the likelihood term for $\mathbb{P}(\pi|D)$. To understand what this conditional distribution should be, notice that the induced prior over DAGs equals

$$\mathbb{P}(G) = \sum_{\pi \in S_p} \mathbb{P}(G \mid \pi)\mathbb{P}(\pi). \tag{11}$$

In order MCMC, the assumed prior $\mathbb{P}(\pi)$ is equal to $\frac{1}{p!}$ (Friedman & Koller, 2003). A natural distribution one may specify for $\mathbb{P}(G|\pi)$, and the one assumed in (Friedman & Koller, 2003), is

$$\mathbb{P}(G \mid \pi) = I(G \precsim \pi)\mathbb{P}^*(G). \tag{12}$$

However, it is trivial to check that Equation (12) implies Equation (11) equals $|\#\mathsf{linext}(G)|\mathbb{P}^*(G)$, where $|\#\mathsf{linext}(G)|$ denotes the number of linear extensions of $G$ (Ellis & Wong, 2008). Therefore, we instead need

$$\mathbb{P}(G \mid \pi) = \frac{1}{|\#\mathsf{linext}(G)|}(G \precsim \pi)\mathbb{P}^*(G)$$

to construct a model that agrees with our desired prior $\mathbb{P}^*(G)$ on DAGs. The difficulty of defining a prior on $\mathbb{P}(\pi|\mathcal{O}_n)$ is calculating $|\#\mathsf{linext}(G)|$, which is $\#P$ in general. We should note that we avoid these issues by instead defining a prior on $\mathbb{P}(\pi|\mathcal{O}_n)$. $\mathbb{P}(\pi|\mathcal{O}_n)$ allows us to define a distribution that *approximately* induces the correct DAG prior; see the discussion in Section 4.

## G. Path and Order Priors

Here we provide the specific form of the order and path priors used in the experiment in Section 6.5. Let $L$, $R$, and $C$ denote the set of ligands, receptors, and cytosolic proteins, respectively, in the network in Figure 5. For the order prior, $\mathbb{P}(\pi)$, we set

$$\mathbb{P}(\pi) := \exp\left(\sum_L f_L(l) + \sum_R f_R(r)\right),$$

where $f_L(l)$ indicates if ligand node $l$ came before all nodes in $R \cup C$ and $f_R(r)$ indicates if receptor $r$ came before all nodes in $C$ and after $L$ in $\pi$. For our method, the order prior is incorporated into our prior on DAGs. Specifically, we replace the DAG prior of $\mathbb{P}(G) = \exp\big(-\gamma\|G\|\big)$ used in our other experiments with,

$$\mathbb{P}(G) := \exp\big(-\gamma\|G\|\big)\exp\left(\sum_L f_L(l) + \sum_R f_R(r)\right).$$

We refer to the prior above as *minIMAP w/ path prior* in Table H. To incorporate path information, we take a prior of the form,

$$\exp\left(\sum_L h_L(l) + \sum_R h_R(r)\right),$$

where $h_L(l)$ indicates if ligand node $l$ had a path to at least one node in $R$ and $h_R(r)$ indicates if receptor $r$ had a path to at least one node in $C$. Combined with the order prior, the prior *minIMAP w/ path and order* in Table H is given by,

*Table 3.* Average correlation of directed features between runs seeded with the true network and runs seeded with MMHC from two hundred randomly generated DAGs with $p = 30$ nodes. Higher is better.

| METHOD | AVG. CORRELATION | STD. ERROR |
|---|---|---|
| MINIMAP | **.977** | .004 |
| ORDER | .928 | .007 |
| PARTITION | .784 | .006 |

$$\mathbb{P}(G) := \exp\big(-\gamma\|G\|\big)\exp\left(\sum_L f_L(l) + \sum_R f_R(r)\right) \exp\left(\sum_L h_L(l) + \sum_R h_R(r)\right).$$

## H. Additional Experiments and Plots

To further analyze the mixing behavior of the different methods, we compute the correlation between different seeded runs for estimating marginal directed edge probabilities. Table 3 shows the average correlations and standard errors based on two hundred synthetic datasets with $n = 1000$ observations and $p = 30$ nodes. Note: Each method was run with $1 \times 10^5$ iterations and a burn-in of $2 \times 10^4$ iterations.

The ROC plots for the $n = 100$, $n = 1000$, and Dream4 datasets are shown in Figure 4; see Section 6.3 for a discussion of these plots. The network in (Mukherjee & Speed, 2008) used for the experiments in Section 6.5 is given in Figure 5.

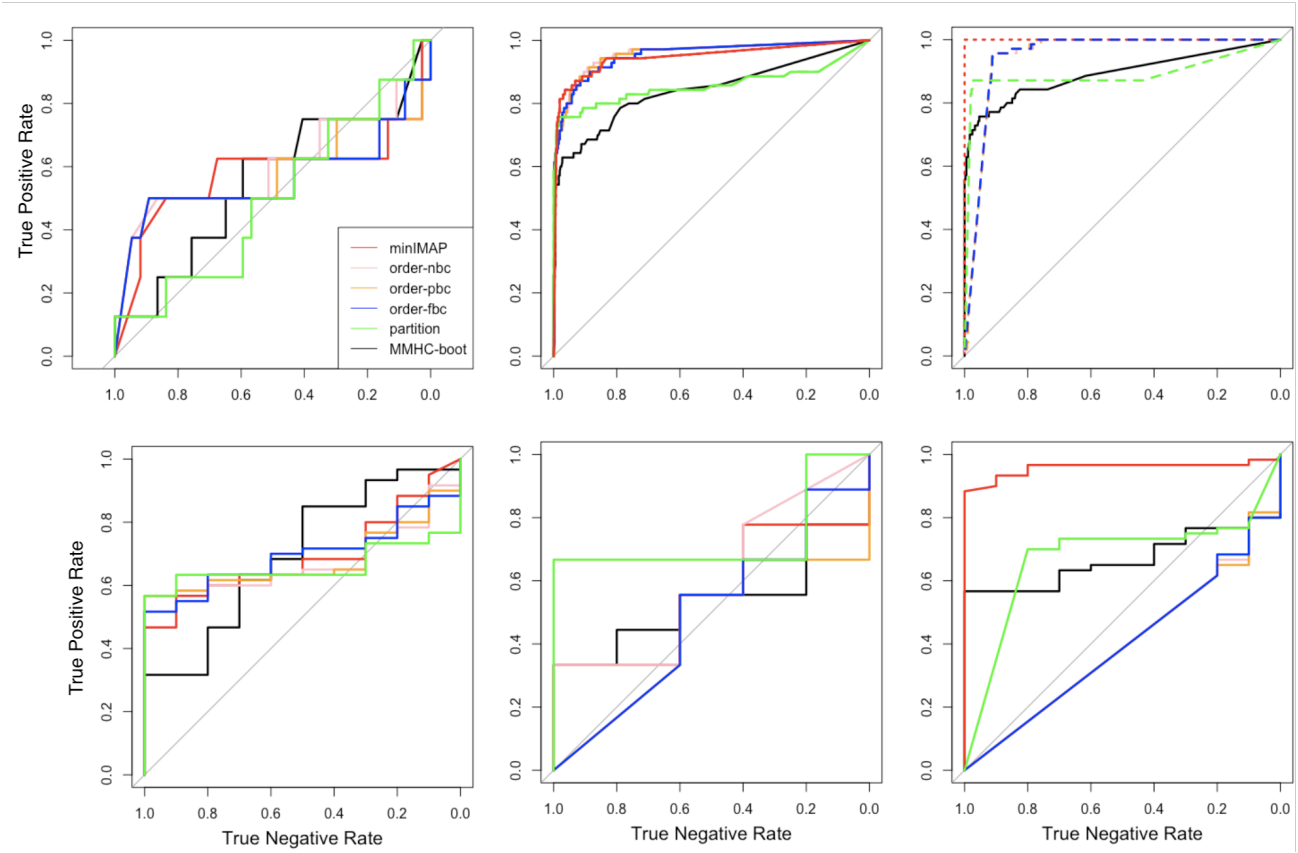


*Figure 4.* The top ROC curves represent recovery of undirected features and the bottom for compelled features. From left to right, the plots correspond to the Dream4, n=100, and n=1000 datasets.

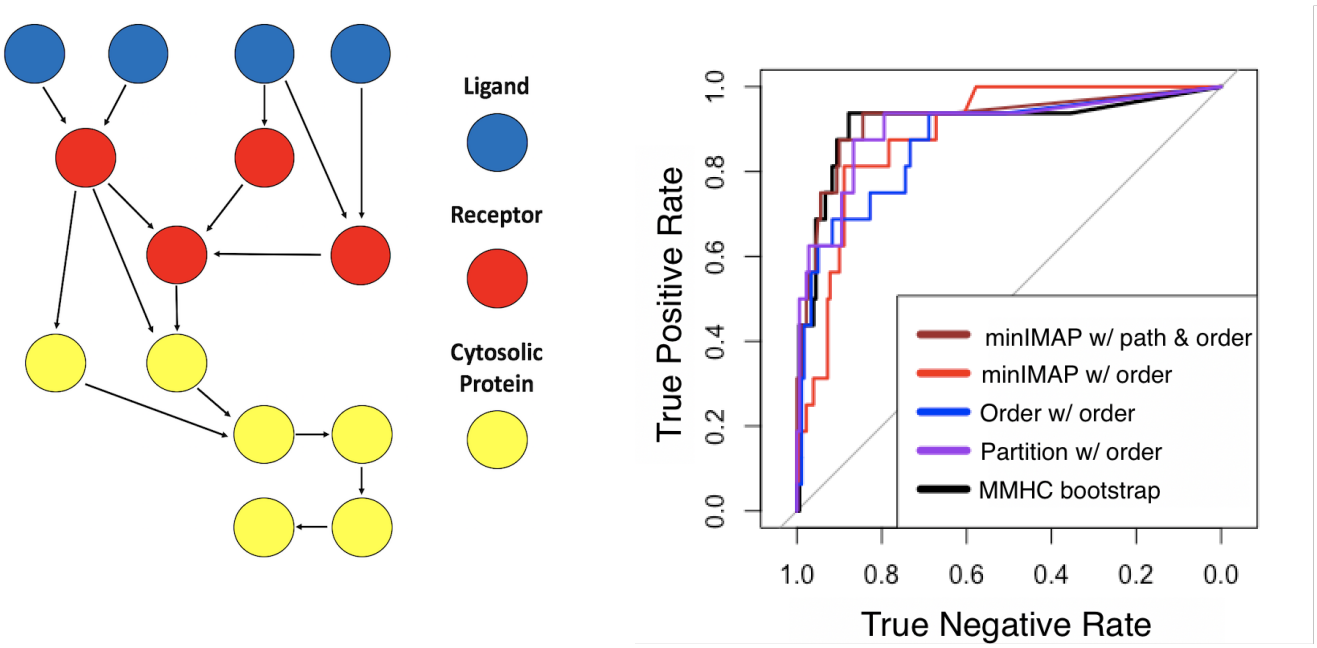


*Figure 5.* The network on the left is taken from (Mukherjee & Speed, 2008). The ROC plot on the right corresponds to the recovery of directed edges. Path and order refers to a prior that takes both path and order information into account as specified in Section 6.5. For order and partition MCMC, only order information can be used in the prior as discussed in Section 6.5.