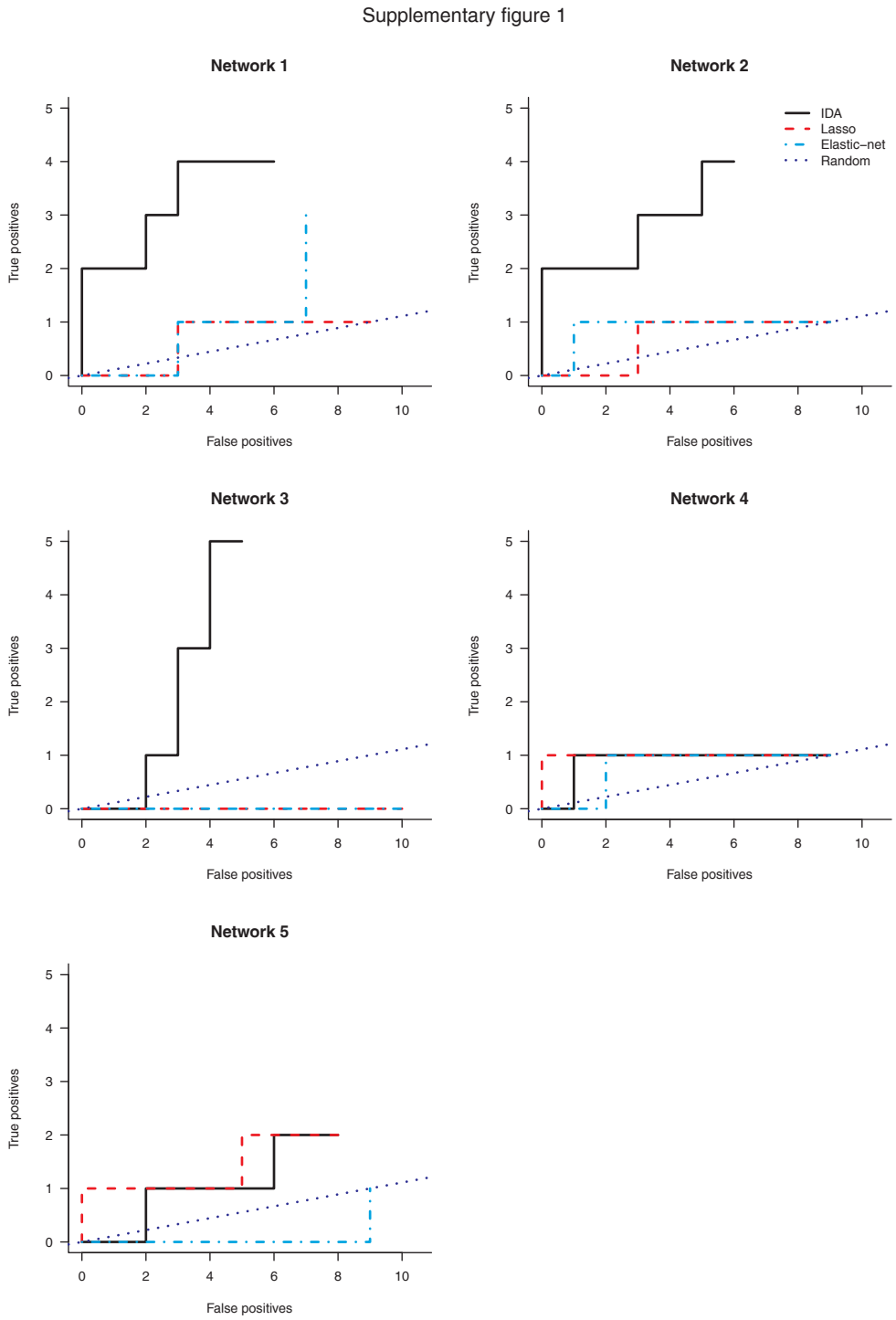
nature | methods

Predicting causal effects in large-scale systems from observational data

Marloes H Maathuis[1], Diego Colombo[1], Markus Kalisch[1] & Peter Bühlmann[1,2]

Supplementary figures and text:

Supplementary Figure 1	Comparing IDA, Lasso and Elastic-net on the five DREAM4 networks of size 10 with multifactorial data.
Supplementary Table 1	Comparing IDA, Lasso and Elastic-net to random guessing on the Hughes et al. data.
Supplementary Table 2	Comparing IDA, Lasso and Elastic-net to random guessing on the five DREAM4 networks of size 10, using the multifactorial data as observational data.
Supplementary Methods	



Supplementary Figure 1. Comparing IDA, Lasso and Elastic-net on the five DREAM4 networks[1] of size 10 with multifactorial data. For each network, the number of true positives are plotted versus the number of false positives, for the top 10 predicted effects from the observational data. The target set is the top 10% of the effects as computed from the interventional data.

Supplementary Table 1. Comparing IDA, Lasso, and Elastic-net (Enet) to random guessing on the Hughes *et al.* data[2]. The target set is the top *m*% (*m* = 5 or 10) of the effects as computed from the interventional data. Columns 3-5 show the number of true positives in the top *q* (*q* = 50, 250, 1,000 or 5,000) estimated effects from the observational data. Column 6 contains the mean and standard deviation of the number of true positives for random guessing. Columns 7-12 contain *P* values for statistical tests that assess if the rankings obtained by the three methods are better than random guessing: columns 7-9 contain *P* values based on the hypergeometric distribution, and columns 10-12 contain *P* values computed with respect to the partial area under the receiver operating characteristic curve (pAUC), based on a simulated null distribution for random guessing. Significance of the *P* values is indicated by stars: $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***).

		Nr. of true positives				*P* values (hypergeometric)			*P* values (pAUC, simulated)		
	q	IDA	Lasso	Enet	Random (SD)	IDA	Lasso	Enet	IDA	Lasso	Enet
Top 5%	50	22	5	6	2.50 (1.54)	5.55E-16 (***)	1.03E-01	3.77E-02 (*)	0.00E+00 (***)	1.40E-02 (*)	2.50E-02 (*)
	250	112	19	22	12.5 (3.45)	0.00E+00 (***)	4.73E-02 (*)	7.76E-03 (**)	0.00E+00 (***)	9.00E-03 (**)	5.00E-03 (**)
	1,000	294	67	62	50.0 (6.89)	0.00E+00 (***)	1.06E-02 (*)	5.10E-02	0.00E+00 (***)	8.00E-03 (**)	1.40E-02 (*)
	5,000	635	333	297	250 (15.4)	0.00E+00 (***)	1.46E-07 (***)	1.58E-03 (**)	0.00E+00 (***)	0.00E+00 (***)	1.40E-02 (*)
Top 10%	50	33	10	8	5.00 (2.12)	0.00E+00 (***)	2.45E-02 (*)	1.22E-01	0.00E+00 (***)	7.00E-03 (**)	2.40E-02 (*)
	250	161	41	43	25.0 (4.74)	0.00E+00 (***)	1.13E-03 (**)	3.17E-04 (***)	0.00E+00 (***)	3.00E-03 (**)	9.00E-03 (**)
	1,000	434	142	132	100 (9.48)	0.00E+00 (***)	1.59E-05 (***)	6.92E-04 (***)	0.00E+00 (***)	0.00E+00 (***)	0.00E+00 (***)
	5,000	1,044	629	594	500 (21.2)	0.00E+00 (***)	2.21E-09 (***)	8.18E-06 (***)	0.00E+00 (***)	0.00E+00 (***)	0.00E+00 (***)

Supplementary Table 2. Comparing IDA, Lasso and Elastic-net (Enet) to random guessing on the five DREAM4 networks[1] of size 10, using the multifactorial data as observational data. For each network, the target set is the top *m*% (*m* = 5 or 10) of the effects as computed from the interventional data. Columns 3-5 show the number of true positives in the top 10 estimated effects from the observational data. The mean ± standard deviation of the number of true positives for random guessing is 0.56 ± 0.69 for *m* = 5 and 1 ± 0.90 for *m* = 10. Columns 6-11 contain *P* values for tests that assess if the rankings obtained by the three methods are better than random guessing: columns 6-8 contain *P* values based on the hypergeometric distribution, and columns 9-11 contain *P* values computed with respect to the partial area under the receiver operating characteristic curve (pAUC), based on a simulated null distribution for random guessing. Significance of the *P* values is indicated by stars: $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***).

		Nr. of true positives			*P* values (hypergeometric)			*P* values (pAUC, simulated)		
	Network	IDA	Lasso	Enet	IDA	Lasso	Enet	IDA	Lasso	Enet
Top 5%	1	3	1	3	9.02E-03 (**)	4.53E-01	9.02E-03 (**)	1.00E-03 (**)	2.21E-01	2.21E-01
	2	3	1	1	9.02E-03 (**)	4.53E-01	4.53E-01	1.00E-03 (**)	2.14E-01	1.23E-01
	3	4	0	0	3.88E-04 (***)	1.00E+00	1.00E+00	4.00E-03 (**)	1.00E+00	1.00E+00
	4	1	1	0	4.53E-01	4.53E-01	1.00E+00	1.32E-01	9.80E-02	1.00E+00
	5	1	1	1	4.53E-01	4.53E-01	4.53E-01	1.87E-01	3.46E-01	1.00E+00
Top 10%	1	4	1	3	7.74E-03 (**)	6.72E-01	5.88E-02	0.00E+00 (***)	3.69E-01	3.69E-01
	2	4	1	1	7.74E-03 (**)	6.72E-01	6.72E-01	2.00E-03 (**)	3.85E-01	2.46E-01
	3	5	0	0	5.89E-04 (***)	1.00E+00	1.00E+00	2.70E-02 (*)	1.00E+00	1.00E+00
	4	1	1	1	6.72E-01	6.72E-01	6.72E-01	2.92E-01	2.21E-01	3.51E-01
	5	2	2	1	2.61E-01	2.61E-01	6.72E-01	2.11E-01	1.09E-01	1.00E+00

Supplementary methods

Description and pre-processing of the Hughes *et al.* data[2]. We considered the compendium of expression profiles of *S. cerevisiae* provided by Hughes *et al.*[2] (available at http://www.rii.com/publications/2000/cell_hughes.html). As interventional data, we used the files data_expts1-75.xls, data_expts76-150.xls, data_expts151-225.xls, and data_expts226-300.xls. When combined, these files contained expression data on 6,325 genes for 300 chemically treated or mutant yeast strains. We only used the column named 10log(ratio), which contained the log ratio of the transcript levels of the mutant strains compared to that of wild-type strains. As observational data, we used the file control_expts1-63_geneerr.txt, which contained gene expression measurements of 6,330 genes for 63 wild-type yeast cultures. Again, we only used the column named 10log(ratio), which contained the log ratio of the transcript levels of the wild-type cultures compared to that of other wild-type cultures.

In both datasets, we removed all genes with missing values. Subsequently, we removed genes that were present in only one of the datasets. In the interventional data, we removed strains 11, 21, 33, 41, 42, 67, 89, 127, 142, 225, 278-300 because they were not single-gene deletion mutants. Moreover, we removed all mutant strains for which the deleted gene was no longer present in the observational dataset. The resulting observational dataset contained gene expression measurements on 5,361 genes for 63 wild-type cultures. The resulting interventional dataset contained gene expression measurements on the same 5,361 genes for 234 single-gene deletion strains. We standardized both datasets, such that the measurements for each gene had mean 0 and standard deviation 1.

Description and pre-processing of the DREAM4 data[1]. The DREAM4 In Silico Network Challenge[1] (http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project) is a competition in reverse engineering of gene regulation networks, involving five networks of size 10 and five networks of size 100. For each of the simulated networks, several types of data were provided. As interventional data we used the files *knockouts.tsv, containing steady state levels of known single-gene knockouts for each of the genes in the network. As observational data we considered the following two possibilities: (i) the files *multifactorial.tsv, containing steady state levels of variations of the networks obtained by applying unknown multifactorial perturbations, and (ii) the files *timeseries.tsv, containing

time course data of the response and recovery of the networks to unknown external perturbations (omitting the time stamps).

The networks and data were simulated by version 2.0 of GeneNetWeaver (http://sourceforge.net/projects/gnw/), using methods of Marbach *et al.*[2]. We provide a brief summary for completeness. All simulated data corresponded to mRNA concentration levels. The network topologies consisted of sub-networks from transcriptional regulatory networks of *E. coli* and *S. cerevisiae*. The dynamics of the networks were simulated using a kinetic model of gene regulation, incorporating independent and synergistic gene regulation, transcription, and translation. Internal noise was simulated via stochastic differential equations, and measurement noise was added based on a model of noise observed in microarrays.

We standardized all datasets, such that the measurements for each gene had mean 0 and standard deviation 1.

Definition of the target set of causal effects. We denote the matrix containing the interventional data by A. Each row in A corresponds to a single-gene deletion strain and each column corresponds to a gene. Thus, for the Hughes *et al.* data[2] the dimension of the matrix is 234 x 5,361, and for the DREAM4 networks[1] the dimensions are 10 x 10 or 100 x 100. We denote the entries of A by $a_{i,j}$, and we let $c(i)$ denote the column index of the gene that was deleted in the strain of row i. Then the size of the causal effect of gene i on gene j, $i \neq j$, can be quantified by $|a_{i,j} - mean(a_{-i,j})| \,/\, |a_{i,c(i)} - mean(a_{-i,c(i)})|$, where $mean(a_{-i,j})$ denotes the mean of the jth column when the ith entry is omitted. Note that we divided by $|a_{i,c(i)} - mean(a_{-i,c(i)})|$ to measure the effect in terms of unit changes in the expression profile of gene i. For each network, we defined the top m% of these effects as the target set of causal effects.

Applying IDA to the observational data to estimate bounds on total causal effects. When data are generated from a given directed acyclic graph (DAG), the total causal effect of a unit increase in one variable on another variable can be computed via established methods, such as Pearl's "do-calculus" (or "intervention-calculus")[3]. In our applications, however, the data generating DAG was unknown. It is well-known that it is generally impossible to estimate a DAG from observational data, even with an infinite amount of data[4]. Under some assumptions, however, it is possible to estimate bounds on total causal effects, using the two-step approach proposed by Maathuis *et al.*[5]. First, one estimates the equivalence class of

DAGs that are consistent with the data, using for example greedy equivalence search[4] or the PC-algorithm[6]. Next, one applies Pearl's do-calculus to each of the k DAGs in the equivalence class, yielding a set of k possible causal effects for each (ordered) pair of variables (e.g., genes). The minimum absolute value of such a set estimates a lower bound on the size of the true total causal effect. In large problems, it is computationally infeasible to evaluate all DAGs in the estimated equivalence class, and Maathuis *et al.*[5] proposed a localized algorithm for such situations. This localized approach also has advantages from an estimation point of view, since it only requires a local neighborhood of the graph to be estimated correctly, rather than the entire graph. We refer to the method of Maathuis *et al*.[5] as IDA, which is short for "Intervention-calculus when the DAG is Absent".

We applied IDA as follows. First, we estimated the equivalence class of DAGs that were consistent with the observational data, using the PC-algorithm[6] as implemented in the R-package pcalg (http://cran.r-project.org/web/packages/pcalg/index.html) with tuning parameter $\alpha_{PC} = 0.01$. We then applied the localized algorithm as implemented in the function ida in the R-package pcalg. For each pair of genes, we summarized the resulting set of estimated possible total causal effects by its minimum absolute value, and we ranked the effects according to this summary measure.

The main assumptions underlying IDA are that the joint distribution of the observational variables is (i) multivariate Gaussian and (ii) faithful to the true (unknown) causal DAG. The main reason for imposing Gaussianity is that it implies linearity and allows conditional independence tests via partial correlations. For both the Hughes *et al.* data[2] and the DREAM4 data[1], we verified that the marginal distributions of the variables were indeed approximately Gaussian. We did not confirm multivariate Gaussianity, since that is almost impossible to check. The faithfulness assumption is satisfied (with probability one) if the data are generated by an underlying DAG without hidden confounders. This is a strong requirement, since DAGs do not allow for feedback loops which are often present in biological systems, and many systems are only partially observed. In particular, feedback loops and hidden variables were used to simulate the DREAM4 data[1]. The Hughes *et al.* data[2] contained gene expression profiles of about 85% of the *S. cerevisiae* genome. Hence, we expect that these data captured most, but not all, of the important variables.

In general, great care should be exercised in the interpretation of results from IDA when the underlying assumptions are violated (in particular the assumption of no hidden confounding variables). In such cases, IDA is best viewed as a new method of determining variable importance, based on causal rather than associational concepts, but not directly

representing causal effects. Nevertheless, we demonstrated in this paper that our approach based on causal concepts performed better than association type techniques, even in scenarios where some of the underlying assumptions are likely to be violated.

Applying Lasso to the observational data. Lasso[7] is a ℓ_1-regularized high-dimensional regression technique. As such, it infers associations rather than causal effects. It is commonly used to determine variable importance (e.g., references 8-11). We applied Lasso to the observational data in the following way. For all genes i, we performed the ℓ_1-regularized regression of gene i on the remaining genes, using the R-package lars (http://cran.r-project.org/web/packages/lars/index.html), where we chose the regularization parameter in a prediction optimal way via 10-fold cross validation. The coefficients $\beta_{j,i}, j \neq i$, in the ith regression model can be interpreted as the expected "association effect" on gene i of a unit increase in the expression profile of gene j, when controlling for all the other genes. For all regression models, we stored the coefficients for all j that corresponded to genes that were deleted in the interventional data. We used the absolute values of these coefficients to rank the effects.

Applying Elastic-net to the observational data. Elastic-net[12] is a high-dimensional regression technique with both ℓ_1- and ℓ_2-regularization. As for Lasso, it infers associations rather than causal effects. We applied Elastic-net to the observational data in a similar fashion as we did for Lasso. Thus, for all genes i, we performed the regularized regression of gene i on the remaining genes, using the R-package elasticnet (http://cran.r-project.org/web/packages/elasticnet/index.html). We chose the regularization parameters in a prediction optimal way via 10-fold cross validation, using the default values of λ_2 suggested by Zou and Hastie[12]: 0, 0.01, 0.1, 1, 10, 100. The coefficients $\beta_{j,i}, j \neq i$, in the ith regression model can be interpreted as for Lasso, and we used the absolute values of these coefficients to rank the effects.

Evaluation of the predictions. Recall that we refer to the top m% of the effects as computed from the interventional data as the target set. Considering the top q predicted effects based on the observational data, using the three prediction methods (IDA, Lasso, and Elastic-net), we

computed the number of false positives (top q predicted effects that were not in the target set) and the number of true positives (top q predicted effects that were in the target set).

We used two different tests to determine if the top q estimated effects of a prediction method contained more effects in the target set than can be expected by random guessing (q = 50, 250, 1,000 and 5,000 for the Hughes *et al*. data[2], q = 10 for the DREAM4 networks[1] of size 10, and q = 25 for the DREAM4 networks[1] of size 100). First, we computed one-sided P values based on the hypergeometric distribution. Second, we compared the partial area under the receiver operating characteristic curve (pAUC) (computed up to the false positive rate determined by the top q effects from IDA) to a null-distribution obtained by constructing 1,000 random orderings and their corresponding pAUCs. For each prediction method, the P value was computed as the fraction of random orderings with pAUCs that were at least as large as the one obtained by the given method (**Supplementary Table 1** and **2**). For the DREAM4 data[1], we only showed the results for the multifactorial data on the networks of size 10, since the difference between the methods was largest in this setting. Considering all four possible combinations of observational data (multifactorial or time series) and the size of the networks (10 or 100), IDA was always at least as good as Lasso and Elastic-net when counting the number of networks in which the pAUC of each method was significantly better than random guessing at significance level $\alpha = 0.01$ for both $m = 5$ and $m = 10$.

We note that our evaluation method focuses on the top estimated effects, and that IDA might lose predictive power for estimated effects further down in the ranking. This does not pose a problem for the main application that we envision for IDA, namely to use it as a tool for the design of experiments, since the top estimated effects are most relevant for this purpose.

References

1. Marbach, D., Schaffter, T., Mattiussi, C. & Floreano, D. *J. Comp. Biol.* **16** 229-239 (2009).
2. Hughes, T.R. *et al*. *Cell* **102**, 109-126 (2000).
3. Pearl, J. *Causality. Models, Reasoning, and Inference.* (Cambridge Univ. Press, Cambridge, UK, 2000).
4. Chickering, M. D., *J. Mach. Learn. Res.* **3** 507-554 (2002).
5. Maathuis, M.H., Kalisch, M. & Bühlmann, P. *Ann. Stat.* **37**, 3133-3164 (2009).

6. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search.* (MIT Press, Cambridge, MA, ed. 2*,* 2000).

7. Tibshirani, R. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288 (1996).

8. Bonneau, R. *et al. Genome Biol.* **7**, R36 (2006).

9. Devlin, B., Roeder, K. & Wasserman, L. *Genet. Epidemiol.* **25**, 36-47 (2003).

10. Tibshirani, R. *Stat. Med.* **16**, 385-395 (1997).

11. Steyerberg, E.W., Eijkemans, M.J.C., Harrell, F.E. & Habbema, J.D.F. *Stat. Med.* **19**, 1059-1079 (2000).

12. Zou, H. & Hastie, T. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320 (2005).