Supplementary Information to 'Addressing false discoveries in network inference'

Tobias Petri, Stefan Altmann, Ludwig Geistlinger,
Ralf Zimmer and Robert Küffner

April 15, 2015

Contents

List of Tables

List of Figures

1 A quick overview of the information in this SI

This Supplementary Information (SI) will possibly provide more information than you might need to grasp the key aspects of our work. Therefore, this section will point out the most important sections to support our key messages.

Additional methods

- Section 2.1 describes the **compendia** used throughout this work.
- For our prediction we loosely rely on a size estimate of a **hypothetical complete yeast regulatory network**. Section 2.2 and 3.2 discuss both methodology and outcomes of this procedure.
- Section 2.4 discusses the applied **validation setup**, in particular our notion of true positives and canonical quality measures like **ROC** and **Precision50** (**P50**).
- In the main paper, we show the results of a network **motif analysis**. The associated method is described in Section 2.7.
- We provide a **modularized network** representation in the main paper, see Section 2.9 for a description of the module derivation, and Figure 9 for the complete network.
- As a **baseline reference** of unsupervised prediction we use a correlation approach, see Section 2.6.1.
- We extended and combined existing approaches to obtain a sensible overall network. This network is referred to as '**consensus**' in the main paper. For detailed information see Section 3.9.

Additional results

- We concluded that the tendency to preferentially attach novel predictions to regulators with many known targets resembles **Simpson's Paradox**. In Section 3.1, we provide a more in-depth discussion on the origin and implications of the seemingly paradox situation.
- In Tables 5 the **performances of all applied methods** are shown in a standard setup, while 6 provides the same values for randomized networks (maintaining degree information). Notably, micro and macro evaluation reflect a regulator-wise and network-wide viewpoint and thus reflect the origin of Simpson's Paradox. A graphical version is shown in the main paper (see Figure 3).
- Apart from our recalibration procedure, we suggest to explicitly integrate each targets **topology information as additional features** and provide the respective results in Table 7. The associated setup with randomized networks is shown in Table 8.
- We provide an overview of **related approaches** that could likely profit from score recalibration to recover local model properties that are currently covered by global topology effects in Section 5. We conclude the section by listing important review papers on inference.
- **Supplementary File**, *http://www.bio.ifi.lmu.de/en/networks/LTBC*.
  The supplementary archive (83MB, 188MB unpacked) contains a clickable map that represents the network of the main paper (see Figure 4). All contained genes can be retrieved together with additional information. Furthermore, the archive contains the expression data and gold standard used in this paper as well as the obtained interaction predictions. A README.txt describes the contained material.
- Following our **modular network representation** (Figure 9), we discuss each module individually and discuss displayed relationships among interactions, processes and experimental conditions as well as expression patterns in Section 6.

2 Additional material and methods

2.1 Description of expression compendia

The following section describes briefly the details on the five selected yeast expression compendia used in this work. Preprocessing and the computation of $\log_2$ fold-changes is described in the main paper, Section 2.1.

DREAM5 Network 4. The DREAM5 Network 4 (DN4) expression data set (Marbach *et al.*, 2012) comprises 536 expression measurements of 5950 yeast genes compiled from 59 publications. We computed 369 $\log_2$ fold change vectors from this expression compendium. A wide range of experimental conditions, including gene, drug and environmental perturbations, partially conducted in time courses is covered.

Many Microbe Microarrays Database. The compendium containing 904 chips of 6777 yeast genes was obtained from the Many Microbe Microarray Database (M^{3D}, Faith *et al.* (2008)). The data set was built from 62 experiments. After conversion to fold change values the data set contained 727 vectors of length 6777.

Hu *et al.* (2007). This and the following two compendia focus on steady-state TF deletion and over-expression measurements that we obtained as $\log_2$ fold change values from the GEO database. Hu *et al.* (2007) performed a comprehensive study of TF knockout experiments. The GEO accession is number is *gse4654*. It contains expression measurements of 263 transcription factors knockout strains under different experimental conditions. The data set was transformed into 269 $\log_2$ fold change values each measuring 6429 genes.

Chua *et al.* (2006). In contrast to the previous compendium, the data set published by Chua *et al.* (2006) (GEO accession number *gse5499*) consists of knockout but also over-expression experiments for 55 TFs. The data set contains 270 $\log_2$ fold change values for 6307 yeast genes.

Additional GEO datasets. From various GEO data sets (a complete list of all accession numbers is available as supplemental information), we obtained additional 194 independent gene knockout measurements for 6307 genes.

2.2 Estimating the size of the yeast regulatory network

We aim to estimate what fraction of regulatory interaction are currently known in yeast. In summary, we compiled 29,398 interactions from 356 TF-to-promoter binding studies as well as 21,847 interactions from 536 gene expression studies. In the latter case, interactions are assumed between a regulator and a target if the target expression changes in regulator deletion or overexpression mutants. Since expression studies would introduce potentially indirect interactions we restrict the gold standard to interactions determined by binding studies. However, these expression studies play an important role in the estimation of the yeast network as described in the following.

Each published study would contribute a small fraction of regulations to the complete network. Measurement bias and study overlap likely introduce saturation effects in the discovery of novel interactions. Thus, we like to estimate the completeness of the yeast regulatory network by empirical limit analysis. An important assumption here is that increasing the number of studies would converge towards a hypothetically completed gold standard (CGS).

We repeatedly sample (10000 times) a fraction of x from the set of all studies that make up the gold standard. This subset induces a partial regulatory interaction network. The (average) fraction of regulatory interactions detected for x parts of all studies is denoted by $\Theta(x)$. $\Theta(x)$ is not expected to depend linearly on x, but should follow a saturation curve and be convergent towards the CGS. We thus decided to model the expected dependency in terms of a Hill coefficient Hill (1910):

$$\Theta(x) = \frac{m * x}{k + x} \tag{1}$$

The two parameters of this equation have a direct interpretation in terms of the network completeness. First, the parameter m is the fraction of interactions in the CGS in relative to all currently known regulations, *i.e.* $m = 1.0$ would imply the currently known gold standard is complete.

Secondly, k is the fraction of available studies when half of all completed gold standard interactions are detected. The coefficients have been estimated using the sample mean of the interaction count for a given x such that the root mean squared deviation was minimized.

Assuming that not all possible interactions are yet known, m will be greater than 1. Thus, multiplying scaling the number of currently known interactions by m would approximate the total number of interactions in the CGS.

We use the approach to contrast the convergence of regulations derived from (i) binding studies (ii) the intersection of binding studies and regulator perturbation-based expression profiling. We sampled from all binding studies in both cases, but in (ii) the population of sampled interactions was limited to the intersected set. As a consequence, $m = 1$ corresponds to the number of interactions supported by both promoter binding as well as TF perturbation studies. The tuple (k, m) was estimated separately for both scenarios.

2.3 Overview of regulatory network inference

2.3.1 Prediction schemes

Several approaches exist to derive a confidence value for predicted regulatory interactions. The following section provides an overview of general input and prediction schemes.

It has been suggested to use mutual information and correlation based *association scores* to estimate the confidence mapping s. A matrix $\hat{C} \in \mathbb{R}^{|R| \times |T|}$, $\hat{c}_{rt} = s(r, t)$ of confidences is calculated among the columns of M (Margolin *et al.*, 2006; Faith *et al.*, 2007). Rather than parameterizing a mapping function s the discrete matrix $\hat{C}$ is estimated directly.

A well-known problem affecting these approaches is that the observed confidences cannot be credited to direct and indirect regulatory influences. Several extensions have been introduced to estimate the network of direct effects (de la Fuente *et al.*, 2004; Feizi *et al.*, 2013; Barzel *et al.*, 2013).

Approaches that exploit experiment annotations (*annotation-aware*) for gene knockout or over-expression have shown promising results (Küffner *et al.*, 2012; Greenfield *et al.*, 2013; Haynes *et al.*, 2013). These models assume that experiment perturbations influence the observed data directly. Yet, it is often not known in practice. Furthermore, the annotation of perturbations and conditions must be present for all experiments and uniformly structured. Only few existing resources provide this level of detail.

Target-centric approaches (Shimamura *et al.*, 2009; Gustafsson and Hörnquist, 2010; Greenfield *et al.*, 2013; Haynes *et al.*, 2013) model the observed value or change m_{eg} of a target gene t. The idea is to explain the observed behavior by that of all other genes in an experiment e. The dependency is reflected by a single function d that is parameterized across all experiments and targets:

$$d(M_{e.}) = m_{et}, \quad e \in E,\, t \in T \tag{2}$$

It is non-trivial to extract (r, t) confidence values from d. Usually, d is therefore modeled such that its parameters are interpretable. The most common choice are (penalized) linear regression models that allow model coefficients to be transformed into confidences (see Sections 2.6.6 and 2.6.7).

Similarly, *pattern-centric* methods (Brown *et al.*, 2000; Qian *et al.*, 2003; Mordelet and Vert, 2008; Huynh-Thu *et al.*, 2010; Ambroise *et al.*, 2012) (see Section 2.6.1) estimate the regulator-target confidence $\hat{c}_{rt}$ as a function of both regulator and target expression:

$$c_{rt} = s(M_{.r}, M_{.t}) \tag{3}$$

This resembles the detection of covariance patterns within and across experiments. Equation 3 thus captures the outcome of *global* approaches that build a single model for all regulators.

By contrast, *local* or *regulator-centric* train a single model s_r for each regulator (Figure 1):

$$c_{rt} = s_r(M_{.t}) \tag{4}$$

2.3.2 Structural priors

A natural extension to target-centric methods is the use of *structure priors* where the value of the target is dependent on previously known regulators rather than all observed genes (Greenfield *et al.*, 2013; Haynes

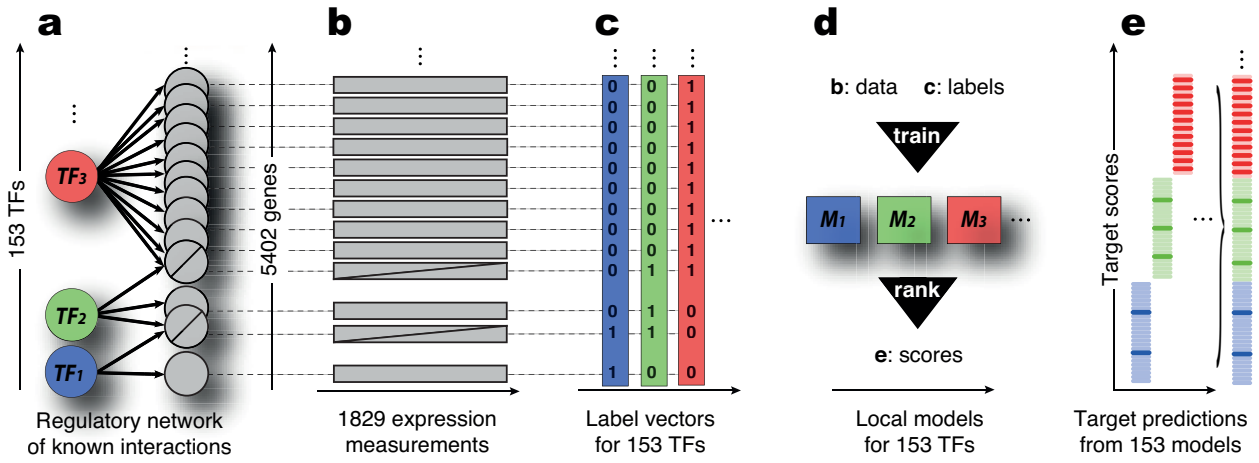

Figure 1: **Supervised regulator specifc inference** Supervised inference methods can utilize **(a)** known interactions as well as **(b)** an expression data matrix. **(c)** Known interactions are transformed into regulator-specific label vectors of length 5042: each gene is labeled 1 if it is targeted by the regulator and 0 otherwise. **(d)** A model M_i is trained for regulator i. Each model consists of n sub-models, where n cross-validation splits are used to avoid overfitting (not shown). The model incorporates the structure prior (a+c) and target expression (b) to distinguish known from non-target genes. **(e)** All potential regulations are predicted by each model and the respective targets are ranked by the predicted confidence scores. A simplified example is shown whereas known targets (saturated) are indistinguishable from non-targets (pastel). Yet, even if all models produce random confidences common evaluation routines would assess the union of all models' predictions as accurate. This effect can be attributed to the fact that large regulators (red, high out-degree) systematically achieve higher confidences than smaller ones (green/blue, low out-degree).

et al., 2013). It has been shown that the combination of *annotation-aware* models and *target-centric* methods could considerably improve predictive performance (Haynes *et al.*, 2013).

Let the subset of interactions $x \in G$ that regulates a target t be $I_{x \to t}$, then the regulator-centric and global methods can be formulated to integrate structure priors by explicitly encoding the topology information I of N as

$$c_{rt} = s^I(M_{\cdot r}, M_{\cdot t}, I_{x \to t,\, x \neq r}) \tag{5}$$

and its regulator-centric equivalent

$$c_{rt} = s^I_r(M_{\cdot t}, I_{x \to t,\, x \neq r}), \tag{6}$$

respectively. Here, $I_{x \to t,\, x \neq t}$ is the set of other known regulators for the target t.

A wide range of methods applied to network inference stems from machine learning (ML). Therefore, some terminology is equivalent. In general, methods that do not train a parameterized models s are referred to as *lazy*. Methods that do not rely on a previously known network structure I are termed *unsupervised*. By contrast, methods using subsets of I as structure priors are usually *supervised* in a ML sense.

It has been argued that the parameterization of models by false negative interactions may mislead supervised methods. Consequently, approaches using only confirmed interactions have been developed to separate real regulations from false positive ones (Geurts, 2011; Cerulo *et al.*, 2010). We refer to this class of approaches as *one-class*, resembling the idea of wrong regulations being outliers to the single true class of regulations (see Section 2.6.5).

Integrative methods use additional data sources explicitly ranging from sequence binding motifs (Ernst *et al.*, 2008) to the derivation of highly specific networks by combination of experimental verification and automated prediction in an iterative manner (Ciofani *et al.* (2012), see Section 5.2).

2.3.3 A network-only model

It has been shown previously that the node degree can have a seriously impact predictive performance estimates (Gillis and Pavlidis, 2011). To estimate the predictive power of a network N's topology alone we define a naïve regulator-specific confidence mapping

$$s_r(t) = \hat{c}_{rt} = |r|_N^{out} \tag{7}$$

This is merely the assignment of a regulator's out-degree to all its targets. Consequently, all targets t of r receive the same score, namely the out-degree of r. Obviously, this confidence function cannot distinguish real from random targets: larger regulators affect more targets and trivially obtain higher scores. Calculation of predictive quality then reflects the baseline expected by random guessing.

It seems valid to assume that the likelihood for any novel target to be regulated by a larger factor is higher as well. Any evaluation that computes a factor-wise performance measure would observe that the predictions are indeed random and no real target can be distinguished from random targets.

Unfortunately, the compilation of all individual predictions into a single list will hide this effect. In fact, ranking regulations among large regulators and their possible targets higher than smaller regulator's interactions is likely superior to any random prediction. Arguably, the result network predicting all possible interactions for say, the 5% largest regulators and nothing else is superior to a complete random solution.

It is important to observe that common measures like ROC and PR curves share the global viewpoint and would score degree-sorting better than random predictions.

2.3.4 Consensus predictions across methods

To compute a consensus across multiple methods we apply a rank merging procedure (Marbach *et al.*, 2012). For regulators $r \in R$ and targets $t \in T$ each method m provides a confidence $\hat{c}_{rt}^m$ (see main paper, Section 2.2). A consensus score for the regulation (r, t) is given by its average rank across all m.

2.4 Summary of Prediction Setup and Validation

The trained models (Figure 1 (d) and main paper, Figure 1 (d), main paper) assign a confidence score to each possible regulation (r, t). Ranking all putative interactions results in a list of $|G| * |T|$ confidence scores $c_{(rt)}$ for each regulator. This list is compared to a gold standard N_{gold} that contains known or experimentally confirmed interactions (see main paper, Section 2.1). For each $r \in R$ we set up a 3-fold cross-validation (3-CV). The set of all network nodes G of N_{gold} is split into n stratified sets. For local models, a scoring model s_r is built on $n - 1$ splits and the n-th set is predicted. For global models we split the set of nodes G into k stratified folds (w.r.t. the number of regulations). The process is repeated n times for each split and repeated k times. A corresponding stratified n-fold split is set up across all regulators to train global models.

To estimate the quality of local or global methods we combine all predictions across all regulators (which is not necessary for global methods) and sort them by their confidences. As previously suggested (Mordelet and Vert, 2008), we apply so-called micro-averaging, *i.e.* the complete list of interactions ranked by their confidences is compared to the corresponding gold-standard annotation. By contrast, macro-averaging would combine regulator-wise performance metrics instead. Macro-averaging is relatively complex to interpret and far less frequently applied. We calculate several quality estimates for each method. For a detailed definition of all applied evaluation metrics see Section 2.5.

To estimate the functional consistency of a prediction we compute compare the expected biological function overlap of novel predicted targets to known targets. A detailed description of this approach is given in the Section 2.8.

2.5 Evaluation metrics for network prediction

In general, we compared predicted interactions to experimentally confirmed interactions, *i.e.* the gold-standard. True positives (TP) are predicted interactions that can be confirmed by the gold standard.

True negatives (TN) are neither predicted nor in the gold standard. False negatives (FN) are not predicted but present in the gold-standard while false positives (FP) are regulatory interactions that are predicted but are not confirmed. Canonical measures are the precision $pr = TP/(TP + FP)$, the sensitivity $sn = TP/(TP + FN)$ as well as specificity $sp = TN/(TN + FP)$.

Each predictive method results in a list of confidence values $\hat{c}_{rt}$ covering all potential regulatory interactions (r, t) among regulator $r \in R$ and target $t \in T$. A ranked list of regulations is obtained by sorting by confidence. All methods below inherently deal with ties present in these lists by averaging results in intervals of equal confidence.

We computed three performance metrics commonly used to estimate the quality of predictive methods:

1. The Precision-50 (**P50**) is the maximal number of predictions that exceed or equal a precision of 50% when lowering a confidence threshold on the predicted scores. The higher the number, the more interactions may be actually predicted with sufficient reliability in practice.

2. The precision recall curve (PR) is the precision pr as a function of sensitivity sn. To vary sensitivity all possible thresholds for interaction predictions within the ranked list are screened. The AUPR is the area under the PR.

3. By contrast, the AUC is the area under the receiver operator characteristics curve (ROC). The ROC is the sensitivity sn as a function of (inverse) specificity $1 - sp$. Similar to the PR all possible confidence thresholds are screened and plotted accordingly.

Random predictions are expected to receive an AUC of 0.5. Vice versa, an AUC of above 0.5 would imply a non-random covariance of the prediction scores and the gold-standard. The best possible AUC value is 1.0 if predictions and gold-standard perfectly agree.

2.6 Summary of applied inference approaches

2.6.1 Predictive Correlation

A simple way to get a predictive supervised local scoring function s_r for an unknown target x is to compare the correlation of $M_{.x}$ (*i.e.* the experiment fold-change values for gene x) to that of known targets of r. We compute an average correlation from individual correlations of known target fold-changes $M_{.t}$, $t \in T_N(r)$ and $M_{.x}$ using Pearson's correlation ρ:

$$s_r(x) := \frac{\sum_{t \in T_N(r)} \rho(M_{.x}, M_{.t})}{|T_N(r)|} \tag{8}$$

This takes into account only existing edges and can be regarded as a prototype of an one-class lazy learning scheme. It served as a baseline comparative approach in Geurts (2011) and led to the results shown in main paper, Figure 3 (main paper) denoted by method number 3.

2.6.2 Decision Trees

Decision trees are decision structures which classify genes with regard to the values of $M_{.x}$. We applied decision trees to train local models. In particular, a TF-specific decision tree imposes an order for experiment examination. Nodes in a tree represent the expression measurements (columns of M) and the corresponding threshold to optimally distinguish between targets and non-targets of the given TF. For each putative target, the prediction procedure starts at the root node and decides for each level which of the possible decision branches is chosen. The choice is based on the node-specific threshold and expression level of the examined target. Leaves assign predictions on whether or not the tested target is regulated by the given TF. Training and prediction using decision trees is performed using C4.5 (Winston, 1992) via probabilistic thresholds. A single decision tree is error-prone thus usually many trees on subsets of data are build and integrated via meta-learning techniques like boosting or bagging. Here, we employ bagging (Quinlan, 1996) to arrive at a numerical scoring function s_r by computing the empirical confidence values for each prediction. In each cross-validation fold (see Section 2.4), we trained 20 trees each using 80% of the positive and 20% of the negative examples in the training fold. Each possible interaction thus received a confidence score averaged from 20 trees.

2.6.3 Random Forests

An extension to decision trees are random forests, which sacrifice the ability of model interpretation in favor of predictive power. This tree learner builds a set of predictive decision trees on experimental subsets and uses a majority voting procedure across all trees to arrive at a decision. Decision values returned are a matrix of class probabilities (one column for each class and one row for each input). Probabilities are calculated from the votes of each generated tree. For random forests (R-package *randomForest*, Liaw and Wiener (2002)) and all approaches described in the following sub-sections, we used default parameters as selected by the corresponding cited software packages.

2.6.4 Two-class SVM classification

Support vector classification (SVC) as originally proposed (Cortes and Vapnik, 1995) provides a robust learning technique based on optimal separation of two-class high-dimensional input vectors.

The use of SVM models for regulator-centric pattern detection has been suggested following global (Brown *et al.*, 2000; Qian *et al.*, 2003) and local (Mordelet and Vert, 2008) prediction schemes, whereas the latter have shown superior performance. In all cases the separation of regulatory from non-regulatory interactions is enforced.

All pairwise similarities of gene measurements $M_{\cdot i}$ and $M_{\cdot j}$ for $i, j \in T$ (the set of targets) are used to derive a hyper-plane intersecting certain training set members, so-called support vectors. The similarity measure is usually a positive semi-definite kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ like a scalar product (linear kernel) or a radial basis function (RBF). The parameter C controls the amount of misclassification allowed during model building. In case of the RBF kernel the bandwidth γ controls how far two instances may be apart to be considered similar. The all-against-all pairwise kernel evaluations is then transformed into a convex optimization problem. The distance of a potential target to the optimized hyperplane is then applied as the local supervised scoring function s_r. Throughout this work, we applied the implementation of libSVM (Chang and Lin, 2011) either directly or via the corresponding R-wrappers (Dimitriadou *et al.*, 2011).

As SVMs solve two- or multi-label classification problems with high accuracy and enforce a regularized solution (Schölkopf and Smola, 2001). In practice, we therefore expect SVMs to generalize well to previously unseen regulation predictions.

2.6.5 Supervised one-class SVM

It has been argued that information on non-targets may be unreliable and thus merely known positive targets should be used to derive regulatory interactions. One-class SVMs build predictive local models based on only positive examples and provide a statistical outlier-detection for targets to be predicted (Chang and Lin, 2011).

2.6.6 Graphical Lasso and Penalized Regression

The graphical LASSO (Least Absolute Shrinkage and Selection Operator) method has been proposed by Tibshirani for the estimation of linear models (Tibshirani, 1994). Lasso fits a generalized linear model via penalized maximum likelihood. This method uses L_1 penalties and hence provides automatic feature selection. The L_1 penalty causes a subset of the solution coefficients to become zero (Hastie *et al.*, 2001). This corresponds to a feature selection and results in a sparse model with regard to gene coefficients. The approach has been adapted using the R package *glmnet* (Friedman *et al.*, 2010; Simon *et al.*, 2011).

2.6.7 Elastic Nets

The Elastic Net combines Lasso and ridge regression by a simultaneous optimization of both L_1 and L_2 penalties. The ridge penalty (L_2) shrinks the coefficients of correlated variables towards each other. The elastic net penalty can be used for regression or classification (Hastie *et al.*, 2001). The elastic net algorithm has first been proposed by Zou and Hastie (2005) for the analysis of microarray data and construction of classification rules. It has been used for various studies with different extensions and settings: the inference of expression values of yeast genes during the DREAM3 challenges where it performed best (Gustafsson and Hörnquist, 2010). Elastic net is used for gene selection in the gene expression analysis framework (Barla

et al., 2008). Shimamura *et al.* (2009) used the elastic net with an extension of a vector autoregressive (VAR) model to infer gene networks from microarray experiments. As in the case of Lasso, the R package *glmnet* is used.

2.6.8 Direct Integration of Network Topology

Methods that integrate prior knowledge of topology usually rely on the data induced by known regulator-target interactions. They do not explicitly integrate the adjacency matrix of the underlying graph.

Say we derive a model for $r \in R$. Then, in order to integrate the knowledge of known targets in a training set, we extend the vector of expression data of each potential target $t \in T$ by information on other known regulators. In particular, the (fold-change) column vector $M_{.t}$ is concatenated to the vector $W_{t.} = (w_{tj})$, $j \in \{R \setminus \{r\}\}$. (see main paper, Section 2.2). For the predictive model s_r, the regulator information for r is excluded to avoid over-fitting. In a cross-validation setting all interactions in the current test set $T \subset I$ are treated as non-existing, *i.e.* $(r,t) \in T \Rightarrow w_{tj} = 0$.

2.7 Analysis of interactions in network motifs

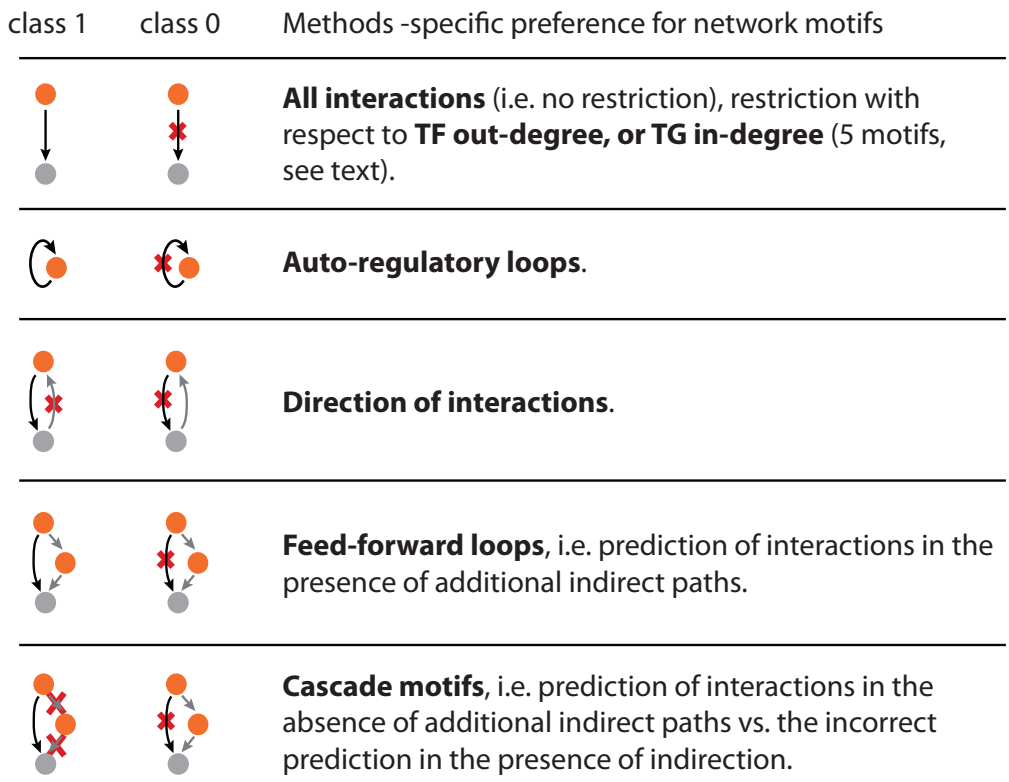

Figure 2: **Motif prediction preferences.** We analyzed method-specific preferences that depend on whether predicted interactions (orange=transcription factor or TF, grey=target gene or TG) take part in 9 different network motifs. Our analysis evaluated, in terms of AUC, how well correct and incorrect predictions (black interaction = class 1 and black crossed-out interaction = class 0, respectively) can be distinguished. The motif context was defined by the presence or absence of further edges in the gold standard (gray interactions). The first row yielded 5 motifs based on additional restrictions on the black interactions: (i) no restriction, (ii) low target in-degree ($\leq$ 2 TFs), (iii) high target in-degree ($>$ 2 TFs), (iv) low TF out-degree ($\leq$ 25 targets) and (v) high TF out-degree ($>$ 25 targets).

It is desirable to estimate the predictive power of an approach in the context of known motif contexts. In the following we describe how we measure motif dependency in the context of these motifs as present in a

Table 1: Assignment of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) in a gold standard motif context. The regulatory interaction (r, t) between a regulator $r \in R$ and its target $t \in T$ is predicted if the regulatory interaction confidence $\hat{c}_{rt}$ exceeds a given cutoff h. The in- and out-degree of gene $g \in G$ in the network N is $|g|_N^{in}$ and $|g|_N^{out}$, respectively. Screening the cutoff allows the computation of ROC and PR curves. Each motif-contrast (*i.e.* the comparison of two distinct motif classes) is separated by a horizontal line and evaluated individually. All interactions that match neither class are discarded for this contrast. Some motifs require the presence or absence of an additional regulator $r' \in R$. This table resembles the classes in Figure 2.

Motif Class	Gold-Standard Context	$\hat{c}_{rt} \leq h$	$\hat{c}_{rt} > h$
regulation	$(r,t) \in N_{gold}$	FN	TP
no regulation	$(r,t) \notin N_{gold}$	TN	FP
regulation, low t in-degree	$(r,t) \in N_{gold}$, $\lvert t\rvert_{gold}^{in} < d_t$	FN	TP
no regulation, low t in-degree	$(r,t) \notin N_{gold}$, $\lvert t\rvert_{gold}^{in} < d_t$	TN	FP
regulation, high t in-degree	$(r,t) \in N_{gold}$, $\lvert t\rvert_{gold}^{in} \geq d_t$	FN	TP
no regulation, high t in-degree	$(r,t) \notin N_{gold}$, $\lvert t\rvert_{gold}^{in} \geq d_t$	TN	FP
regulation, low r out-degree	$(r,t) \in N_{gold}$, $\lvert r\rvert_{gold}^{out} < d_r$	FN	TP
no regulation, low r out-degree	$(r,t) \notin N_{gold}$, $\lvert r\rvert_{gold}^{out} < d_r$	TN	FP
regulation, high r out-degree	$(r,t) \in N_{gold}$, $\lvert r\rvert_{gold}^{out} \geq d_r$	FN	TP
no regulation, high r out-degree	$(r,t) \notin N_{gold}$, $\lvert r\rvert_{gold}^{out} \geq d_r$	TN	FP
auto-regulation	$(r,t) \in N_{gold}$, $(r = t)$	FN	TP
no auto-regulation	$(r,t) \notin N_{gold}$, $(r = t)$	TN	FP
directed regulation	$(r,t) \in N_{gold}$, $(t,r) \notin N_{gold}$	FN	TP
reverse regulation	$(r,t) \notin N_{gold}$, $(t,r) \in N_{gold}$	TN	FP
feed-forward	$(r,t) \in N_{gold}$, $\exists r' \in R : (r',t), (r,r') \in N_{gold}$	FN	TP
cascade	$(r,t) \notin N_{gold}$, $\exists r' \in R : (r',t), (r,r') \in N_{gold}$	TN	FP
direct regulation	$(r,t) \in N_{gold}$, $\nexists r' \in R : (r',t), (r,r') \in N_{gold}$	FN	TP
cascade	$(r,t) \notin N_{gold}$, $\exists r' \in R : (r',t), (r,r') \in N_{gold}$	TN	FP

gold-standard. In particular, we contrast two motif types at a time to obtain sensible positive and negative classes to classify each prediction (a regulation exists or not) as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). Given this definition common performance values like AUROC can be computed.

2.7.1 Simple regulations

In principle, simple regulations are no motifs. Thus, it is straightforward to decide whether a predicted regulation is present in the gold-standard (TP) or not (FP). Similarly, a gold-standard regulation that is missed by the prediction is FN while a TN is reported by neither prediction nor gold-standard. To get a more specific idea of the influence of node degree we restrict the set of regulations that are considered for AUROC analysis (see Figure 2).

2.7.2 Auto-regulation

It is useful to decide how well predictions can resolve auto-regulatory loops. Then two classes do exist in the gold standard: (1) auto-regulation and (2) non-autoregulation. For each regulator-target pair we check whether a predicted regulation exists in the gold-standard (TP) or not (FP). It is also correct to predict no regulation if no regulation is present in the gold-standard (TN), yet would imply a FN otherwise.

2.7.3 Directed interactions

The simplest motif involving two distinct entities of the network is a directed interaction. If no reverse regulation is present in the gold-standard, then a predicted regulation is considered TP and FP if the gold-

standard features a reverse regulation. By contrast, it is considered FN not to predict a regulation if the reverse regulation is present in the gold-standard and TN if is not.

2.7.4 Feed-forward loops and cascades

In case of regulations embedded within feed-forward loops the definition of classes is slightly more complicated for the set of non-feed-forward loops is too general. Instead, we restrict the analysis to feed-forward-loops and cascades in this case. All other motifs are neglected. For each regulator-target pair we thus check whether a regulation is predicted and if that is the case if the gold-standard context of the regulation is a feed-forward loop (TP) or a cascade (FP). The prediction of no regulation is considered a FN if a gold-standard feed-forward context is present. In case of a cascade motif it is correct not to predict any regulation (TN).

2.7.5 Direct regulation and cascades

Similarly, for cascade motifs, the contrasting classes are regulations without existing bypass on the one hand and on the other hand cascades. Thus, the prediction of a direct regulation while only a bypass is actually present in the gold standard is considered FN. Consequently, it is correct not to predict an interaction (TN). For the positive class, the prediction of a direct regulation is correct (TP) since no cascade is present. If we miss the direct regulation despite there is no existing bypass in the data we consider the missing regulation a FP.

In general different types of regulatory interactions according to network patterns surrounding them. For each type two classes Each interaction defined in the gold-standard N_{gold} is assigned to one or more types (see Figure 2) and predicted confidences are evaluated in this context (see Table 1). Given a prediction method we evaluate the specific advantages or disadvantages for each interaction type.

For a given method we analyze the list of confidences for all possible $|R|*|T|$ regulatory interactions. The types are defined by the gold-standard network. The list of confidence values is restricted to include only one type of interaction at a time (see Figure 2). Then, for the remaining interactions, AUC values are computed as guided following the assignments defined in Table 1. The resulting AUC values are motif-specific and may be compared across several methods.

For a given threshold h an interaction (r, t) is predicted if $\hat{c}_{rt} > h$. The interaction is considered correct in the motif context if it is supported by the gold standard. Each type induces a subset of both gold-standard regulations and non-regulations. This is necessary to arrive at sensible contexts *e.g.*, the restriction to high out-degree regulators.

The filtered set of interactions is then relevant for the motif of interest. Regulations that do not match any class are discarded for this type. Motifs of up to three nodes $(r, r', t) \in (R \times R \times T)$ are analyzed. We define degree cutoffs d_r and d_t to distinguish low from high node degrees.

2.8 Functional overlap between known targets and novel predictions

Network inference methods suggest additional interactions that are not yet contained in the gold standard of experimentally supported interactions. We defined a *functional coherence* score to determine whether biological functions (The Gene Ontology Consortium, 2010) – annotated by GO processes to the known, experimentally supported targets of a given TF r – match the functions of newly predicted target genes (see Figure 3).

A functional profile for r was defined based on the known targets t in the gold standard network. The profile is represented by a vector $ont_R(r) \in \mathbb{R}^K$, where K is the number of functional categories, such that functions associated to many targets of the given TF receive a higher weights. The functional coherence of newly predicted targets was then evaluated by comparing the profile vector to according profiles $ont_G(t)$ of each predicted target. The d-th component of $ont_G(t)$ is 1 if t is associated to the d-th functional category, and 0 otherwise. It reflects how well novel target predictions correspond to the functional annotations of targets in the gold standard.

The functional coherence measure depends on the functional representation of r as a vector of K GO biological processes $ont_R(r) \in \mathbb{R}^K$. Each dimension $d = 1 \ldots K$ is the statistical significance of an intersection set, *i.e.* of genes that are both known targets of a given regulator r as well as associated with the d-th biological process. The significance of the overlap was calculated as functional enrichment score of the

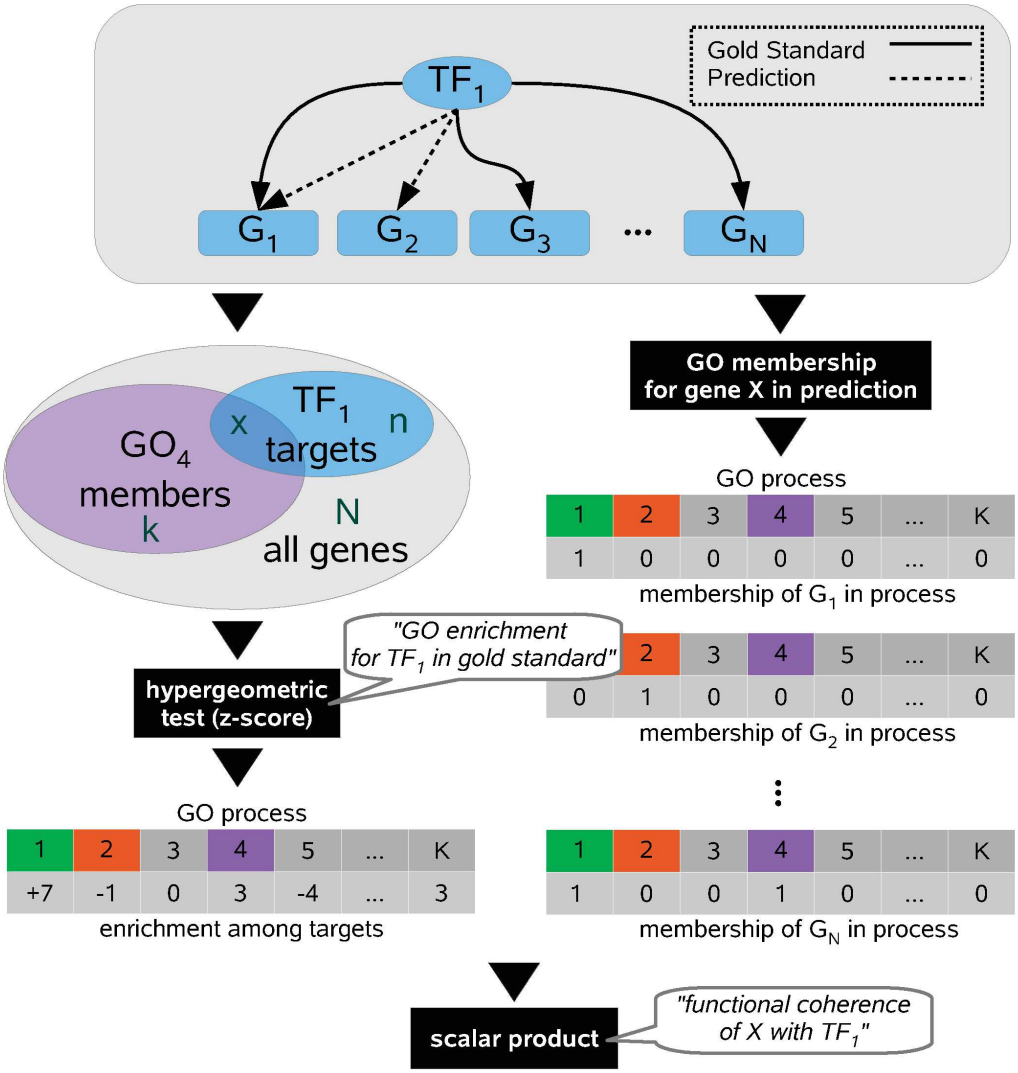

Figure 3: **Functional Coherence Measure.** For each TF (top), a measure of functional coherence is derived by assessing the overlap of functional annotations of (1) its experimentally supported targets and (2) newly predicted putative targets among G. In a first step (left side), we apply the hypergeometric test to analyze the enrichment of functional annotations among the experimentally supported targets. The enrichment score is computed with respect to observing an overlap of x or more genes among targets of TF_1 and those genes annotated with the hypothetical GO category GO_4. The table 'enrichment among targets' denotes this enrichment as z-scores for all $1..K$ GO processes in the second row. Positive or negative z-scores denote process annotations that are enriched or depleted, respectively, among the targets of TF_1. Each table of newly predicted targets of TF_1 (right side) refers to a single gene, which might either be part of GO_4 or not, hence assigning 0 or 1, respectively (column 4 of the second row). Finally, functional coherence is computed as a scalar product between an enrichment vector (left table) and a gene-specific vector (right table). Note that if the coherence for an known TF_1-target such as G_N is calculated, it is removed from the calculation of the enrichment vector in a leave-one-out setup.

targets $T_{N_{gold}}(r)$. For a given functional category, it was computed as a hypergeometric z-score $h_z(x, N, n, k)$ given the number of genes k in the category, the number of genes n known to be regulated by r, the number N of all possible targets in the gold standard and the number of genes x in the intersection (see Figure 3). Similarly, each $t \in T_{N_{pred}}(r)$ was then assigned to a vector $ont_G(t) \in \{0,1\}^K$ encoding the membership of t in each process. For a regulatory interaction $e = (r, t)$, we then computed the functional coherence as the normalized scalar product $cons_{rt} := \langle ont_R(r), ont_G(t) \rangle$.

We then selected a set of regulatory interactions $\{(r,t) \,|\, c_{low} \leq s_r(r,t) < c_{high}\}$ for each interval of prediction scores $c = \langle c_{low}, c_{high} \rangle$. Each interval is associated with a row in a two-dimensional density map that displays a histogram across equally sized bins of coherence scores.

2.9 Derivation of modules from the predicted network

We applied a k-means clustering approach using an euclidean distance metric on the predicted network N_{pred}. We represent each TF as a binary vector of all targets $t \in G$ (the set of all genes, see above). An interaction was encoded as 1, non-interactions as 0. The representation resulted in a matrix M_N with 153 rows (TFs) and 3747 columns (targets). Clustering was performed in two dimensions: (1) clustering of TFs and (2) clustering of targets. For both clusterings, k is screened randomly 100 times in the range of 8 to 15. Overall, 10000 biclusterings were thus prepared. We filtered the result to retain only biclusters with a minimum density of 40% predicted interactions. Subsequently, biclusterings were ranked based on the retained biclusters using

- the number of biclusters in the biclustering n_b
- the number of interactions n_i covered by the biclustering
- the number of TF clusters n_t and
- the number of target clusters n_g

by the empirical ranking criterion

$$n_i - (n_b * n_t * n_g). \tag{9}$$

The criterion is designed to cover as many interactions as possible within a minimal number of clusters. The key result of this procedure, the set of highest scoring biclustering, is represented in main paper, Figure 4. Here, TF clusters are connected to target clusters they regulate. Interaction clusters then represent the biclusters derived by this procedure. A detailed discussion of each TF and target cluster is given in Section 6.

3 Additional results

3.1 Simpson's paradox in network inference

We discussed that the regulator-wise and network-wide viewpoints for the evaluation of inference approaches resemble Simpson's Paradox. In this section we will discuss this in more detail.

3.1.1 A working example

Let's start with an example of Simpson's Paradox to clarify the conditions that lead to the (seemingly) paradox situation. For the evaluation of network inference we aim to compare two methods A and B. Each methods provides us with a confidence for each potential regulatory interaction. There are two common approaches to evaluate the result network. (1) We sort all predicted regulations based on their assigned confidence values and compute some canonical network-wide quality measure (like an AUC). This is known as micro-evaluation. (2) We sort the predicted regulations *per regulator* and compute local quality measures. This is often referred to as macro-evaluation. In practice, a situation may occur where micro-evaluation suggests that B is superior to or on par with A and, simultaneously, most or all macro-evaluations would prefer A. This seems to be paradox, because we intuitively think that a method that is better for all sub-problems (or subsets) should perform better for the complete set as well. This reversal given two points of view (complete and subsets) is often referred to as Simpson's Paradox (Simpson (1951); Pearl (2009)).

3.1.2 Observations on real-world inference

Mapped to our setting, method A is a regulator-specific machine learning model that can be used to predict novel targets from known regulator target patterns, *e.g.*, a random forest approach. Method B randomly re-assigns the known regulations to random targets and then uses method A on the shuffled network and data. Method C is a baseline method that works free of data would predict the number of known targets for each regulator as a confidence value for all its targets (see Section 2.3.3).

For a network-wide estimate of quality (like an AUROC) both A and B seem to be on par. For example a random forest model achieves a micro-evaluation AUC of 79.6 (see Table 6). The same model being trained on a shuffled topology achieves 72.9 (see Table 6).

In practice, both models would be considered to yield useful results given their overall performance. Yet, for an averaged macro-evaluation we observe 63.1 for standard random forests and 49.4 using randomized topologies. Notably, a model that provides random predictions for almost all regulators obtains a global quality of more than 70 percent. The model quality is also evident in the Precision50 (P50): for shuffled random forest predictions the P50 is plain 0, whereas 6996 regulation can be predicted at 50% precision otherwise.

While both networks are of similar overall quality with respect to the micro-evaluation AUC, the regulator performance is crucial. It seems inconsistent that the AUC fails to recognize this shortcoming as it provides a network-wide point-of-view.

3.1.3 Simpson's Paradox motivates confidence recalibration

The Simpson's Paradox refers to the counter-intuitive interpretation of observed results. In fact, both the micro-evaluation AUC and the average macro-evaluation are correct. The common perception is that a network cannot be correct globally, but random for each regulator. This view neglects an important aspect: both methods A and B have access to the degree of a regulator. This prior information seems to somehow override the predictions that individual, regulator-specific models provide. In fact, we observed a strong degree-dependency for predicted confidences in all models, and the micro-evaluation AUC would benefit from ranking larger regulators first, while macro-evaluations do not rely on this ranking. We refer to the preference to predict targets for larger regulators as *High Degree Preference* (*HDP*).

We can by now tell that the Simpson's Paradox is induced by the integration of topology information. Strikingly, method C yields an AUC of 79.8, a score that is superior to methods that integrate data. Since the AUC itself is a reasonable quality measure one may argue to choose this globally best model. This

argument is easily disproved: The regulator-wise quality is essential for almost any kind of application, and method C cannot rank the predictions for individual regulators – neither can method B.

To resolve the Simpson's Paradox would then mean to select a network-wide set of regulations with reasonable performance whereas individual regulators should maintain the quality that state-of-the-art predictive methods can provide. To tackle this problem, and thus bridge the gap that leads to Simpson's Paradox, we suggest to capture the regulator-specific nature of B as a random background and use it to contrast the results of the corresponding method A. We refer to this as confidence calibration (*CoRe*). This is the motivation behind the κ-transformation procedure (see main paper, Section 2.3) as key element of *CoRe*.

Obviously, while we aim to uncover regulator-wise information, the topology information should not be cancelled out completely: it is implicitly reflected by an increased κ-value, *i.e.* the degree-specific contrast among random and non-random confidence values.

As expected the Simpson's Paradox and thus the *HDP* disappears upon recalibration. While the maro-evaluation AUC stays the same, the semi-global P50 estimate for these networks slightly drops. Yet, by design, the estimated network-wide false discovery rate is drastically reduced.

3.2 An estimate for complete size of the yeast regulatory network

The Yeastract database (Abdulrehman *et al.*, 2011) compiles yeast interactions from two types of experiments. The first type detects physical binding of transcription factors (TFs) to promoter regions of target genes. The second one tests whether a perturbation (*e.g.*, knockout) of a TF leads to changes in the expression of putative targets. We speak of *active* interactions if they are observed in both types of studies, *i.e.* if the TF both binds to and effects transcriptional changes in a corresponding target gene. Just 9% of all interactions detected by binding studies are thus confirmed (Figure 4a). It is important to note that interactions are unevenly distributed among the TFs (TF out-degrees = number of known targets per TF, Figure 4b).

To demonstrate that network inference is necessary, we estimated the size of the complete yeast regulatory network (see Section 2.2). We treated the number of interactions as a function of the available binding studies. We found that a hypothetically complete network would contain 3.5 time the number of interactions in Yeastract ($3.5 * 29,398 \approx 105,000$ interactions) given an infinite number of binding studies (Figure 4c).

Based on this estimation, we further reason that 2.5 times the number of currently available binding studies would be required to obtain half of the completed network ($2.5 * 356 \approx 900$ studies). Furthermore, our results suggest that 50% of all "active" interactions are currently known. However, the low confirmation rate of 9% impedes their identification and separation from the inactive ones. Inference methods are potentially able to close that gap.

3.3 Performance of network predictions without expression data

Expression data is the principal source of information exploited to infer interactions. However, by disregarding expression data in a network-only approach, basic issues of regulator-specific methods can be illustrated (see Section 2). An analogous approach was suggested previously for function prediction (Gillis and Pavlidis, 2011). For the network-only approach, we assigned confidence scores based on the out-degree of regulators such that scores for targets of a regulator A are always higher than scores for targets of a regulator B if A has the higher out-degree. In contrast, scores among the candidate targets of a single regulator are distributed uniformly so that true and false targets of a given regulator are indistinguishable (Figure 1e and Section 2.3.3).

Accordingly, we calculated a cross-validated AUC for a single network combining all regulator-specific confidences as suggested (Mordelet and Vert, 2008). In addition, we determined the AUC for all regulators separately. The latter indeed resulted for each regulator in an AUC of 0.5 expected for random predictions. However, the integration of the same predictions across regulators into a joint confidence score distribution resulted in an AUC of 0.798, seemingly indicating a substantial performance. Thus, despite the fact that individual predictions were random, an integrated network can exhibit a substantial enrichment of true TF targets at higher scores (Figure 1e).

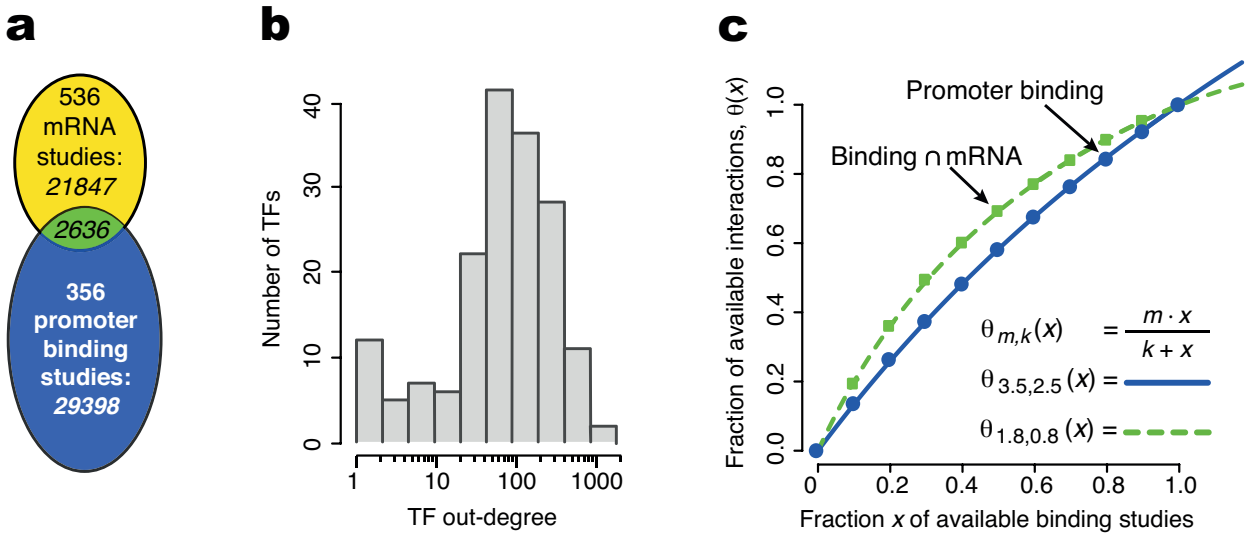

Figure 4: **Properties of yeast interactions.** **(a)** The Venn diagram depicts the number of interactions (italics) in Yeastract obtained from 536 mRNA expression studies (yellow), 356 promoter binding studies (blue), or both (green). **(b)** shows the distribution of TF out-degrees in Yeastract binding studies. **(c)** plots the fraction of interactions contained in random subsets of binding studies as a function of subset size ($x = 1.0 \mathrel{\widehat{=}} 356$ studies). Fractions are plotted for interactions from binding studies (blue circles, ordinate: $1.0 \mathrel{\widehat{=}} 29,398$ interactions) and for interactions detected in both study types (green squares, $1.0 \mathrel{\widehat{=}} 2,636$ interactions). We fit first order Hill functions $\theta(x)$, shown as lines, to estimate the ratio of expected to known interactions m. Thus, an infinite number of promoter binding studies (blue line) would detect $m = 3.5$ times the currently known 29,398 interactions. The second parameter k indicates that $k = 2.5$ times the currently available 356 studies are required for detecting half the expected interactions.

3.4 Performance of network prediction based on TF binding sites

Table 2: Performance (AUC) of SEREND across TFs

Data	Micro	Macro	Corrected
Motif	79.6	56.9	59.7
Expression	79.3	66.2	61.1
Combined	80.4	61.2	65.5

SEREND trains classifiers for each TF individually and is thus based on local models for prediction. For the application to yeast, we used positional weight matrices (PWMs) obtained from the JASPAR database (Bryne *et al.*, 2008) and derived PWM promoter matching scores via CUREOS (Rahmann *et al.*, 2003). As detailed in Section 5.2, SEREND separately trains two logistic regression classifiers to predict GRIs from expression data and TF promoter binding sites, respectively. A third classifier is employed to combine the predictions from the other two classifiers.

SEREND's confidence scores for putative GRIs are reported for each of the three classifiers, which enabled us to separately evaluate the performance. Large difference in performance between regulator-wise and network-wide quality measures (Table 5) suggest that SEREND would preferentially attach novel

regulations to larger regulators. Table 2 indicates that each of the individual scores is susceptible, as shown by inflated micro-averaged AUC values.

We next analyzed the TF-specific performance achieved using only the information on binding sites. Table 3 demonstrates the strong shift towards new targets for high-degree TFs. The two TFs (*ste*12, *rap*1) with the highest out-degrees exhibit the lowest AUC performance but account for 80% of the predictions. This shows that the networks estimated by SEREND may profit from a reduction in False Discoveries by score recalibration. Table 4 depicts the results after recalibrating SEREND's sequence binding scores using *CoRe*. After the recalibration, the predictions are balanced with respect to TF out-degree (compare Figure 7). No significant predictions were obtained for *ste*12, indicating that predictions achieved before were independent of the regulator-specific model and entirely due to *HDP*. However, even after recalibration, suitable numbers of targets were predicted (empirically, we required that the number of targets predicted for a given TF should be $> 10\%$ of its out-degree) for only 10 out of 153 (6.5%) TFs while no or very few (as in case of *fkh*1) targets were predicted for the majority of TFs.

Table 3: Examles for some regulator-specific performance using only promoter binding information

TF orf	Gene	Outdegree	Predicted	AUC
YHR084w	*ste*12	1770	1609	53.4
YNL216w	*rap*1	1159	736	67.7
YJR060w	*cbf*1	313	157	86.7
YBR049c	*reb*1	502	151	87.7
YKL112w	*abf*1	459	94	86.0
YDR207c	*ume*6	166	70	83.3
YEL009c	*gcn*4	284	46	79.9
YOL028c	*yap*7	174	18	84.8

Table 4: TF-specific performance of promoter binding after recalibration

TF orf	Gene	Outdegree	Predicted	AUC
YKL112w	*abf*1	459	532	86.0
YJR060w	*cbf*1	313	361	86.7
YEL009c	*gcn*4	284	237	79.9
YBR049c	*reb*1	502	198	87.7
YOL028c	*yap*7	174	129	84.8
YGL131c	*snt*2	23	79	93.2
YDR207c	*ume*6	166	75	88.3
YMR043w	*mcm*1	238	60	69.8
YBL005w	*pdr*3	107	24	75.5
YKL109w	*hap*4	159	15	65.8
YIL131c	*fkh*1	207	13	71.8

3.5 Robustness of *Confidence Recalibration* (*CoRe*)

As described in the main paper, Section 2.3, the proposed recalibration, *CoRe*, is based on random transcriptional networks. Figure 5 shows how the number of used random networks influences the resulting evaluation metrics. Shown are the results obtained from all possible subsets of 1..10 random networks based on the 10 random networks used in this study. Using a larger number of random networks boosts the scores and decreases their variance. The differences in averaged performance estimates decrease as more random networks are used, indicating ten networks enable a sufficiently accurate recalibration. In addition, the results from evaluation metrics without recalibration are shown, demonstrating the substantial over-estimation of performance.

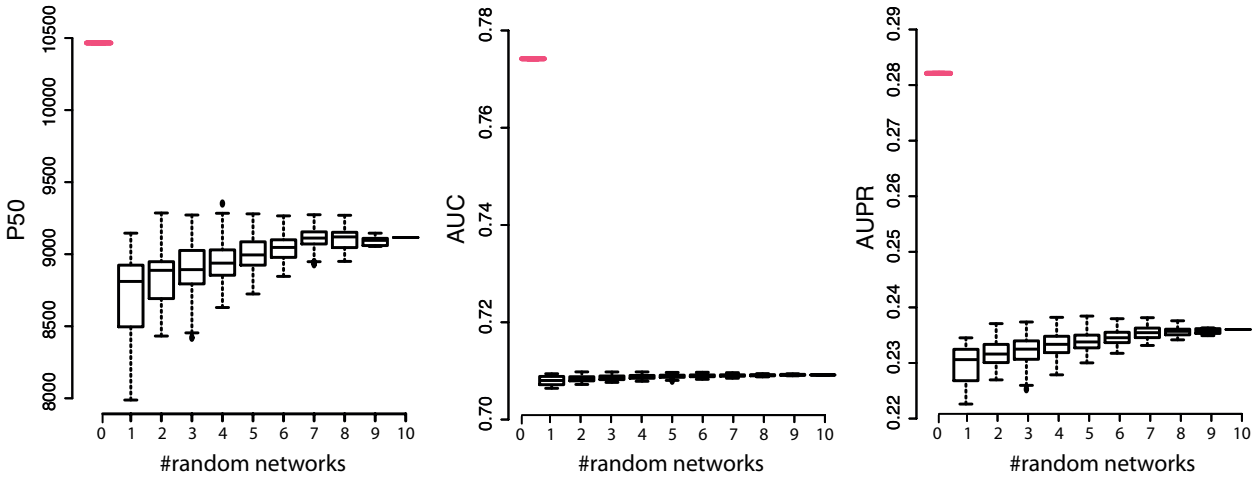

Figure 5: **Robustness of score recalibration** The boxplots depict how the number of random transcriptional networks (abscissa) used for the recalibration influences the results from various evaluation metrics (ordinate) including the AUC, the AUPR and the P50 measures. For comparison, we also show the results obtained without score recalibration (red bar, based on no, *i.e.* 0, random networks).

3.6 Dependency of distribution parameters on TF out-degree

In this section we further examine the properties of score distributions and their dependence on the TF out-degree. Figure 6 depicts the dependency of distribution parameters on the TF out-degree. Both median and maximum of the score distributions exhibit a strong positive correlation with respect to the TF out-degree. Across the range of TF out-degrees, the maximum shows a higher slope than the median. This indicates that the score distribution is not only shifted but also scaled in dependence on the out-degree. As shown in the next section, a threshold on the non-recalibrated scores will therefore select more targets for TFs for which many targets are assigned by the gold standard.

3.7 Relationship between TF out-degree and number of predicted targets

GRIs are typically selected by applying a precision-based threshold (*e.g.*, P50 for a precision of 50%) on a global list of predictions ranked by confidence score (Mordelet and Vert, 2008). In case of non-recalibrated scores, Figure 7 shows that thereby, an excessive number of predictions are selected for high-degree TFs while all predictions may be rejected in case of low-degree TFs. The P50 threshold can also be applied to each TF individually (corresponding to a macro-evaluation), but this leads to similar results. In contrast, a global P50 criterion applied to calibrated confidence scores results in a balanced ratio of predicted to known TF targets, *i.e.* data points in Figure 7 are parallel to the abscissa.

3.8 Score distributions based on probability estimates

As an alternative to the raw confidence scores employed by methods such as Sirene (Mordelet and Vert, 2008), Platt scores have been proposed by Holloway *et al.* (2008). Platt scores transform the raw confidence scores into probability estimates that scale between 0 and 1 (compare Section 5.4). As shown in Figure 8, Platt scores derived from randomized gold standards exhibit similar degree dependencies as the raw confidence scores depicted in Figure 2a. Thus, the transformation into Platt scores alone is not sufficient to correct for the *HDP* effect.

3.9 Improving regulator-specific predictions

In order to increase the number of correctly predicted interactions, we implemented three improvements.

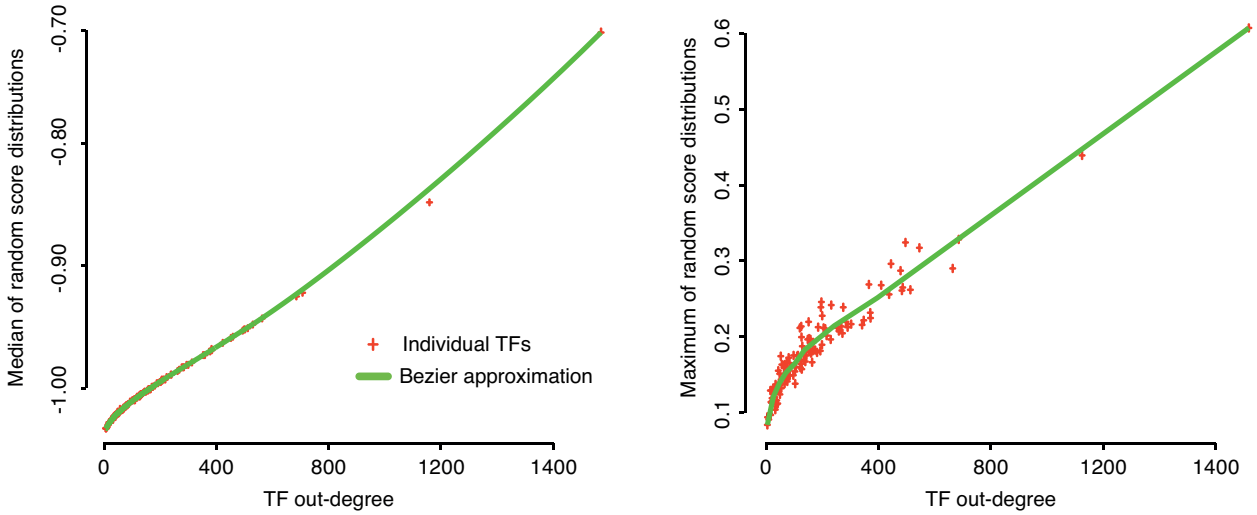

Figure 6: **Degree dependency of location parameters.** For each TF, simple location parameters are estimated such as the median (left panel) and maximum (right panel) from score distributions derived from random gold standards. The plot depicts the dependencies of these parameters (ordinate) on the TF out-degree (abscissa). Shown as red crosses are the parameters as estimated from individual TFs and their approximation via Bezier curves (green line). Score distributions were derived from local SVM models obtained from the Sirene approach.

First, we integrated the five methods selected in the previous section into a consensus to obtain a single network (see Section 2.3.4). This integration is potentially beneficial to exploit complementary advantages of different methods (Marbach *et al.*, 2012). We re-ranked interactions according to the average calibrated score across all methods and selected the top-ranking interactions with a precision of 50% or better. This consensus spanned a network of 8,726 predicted interactions (see main paper, Figure 3c). To examine compendia-specific effects, we built consensus networks from predictions derived from subsets of expression compendia and subsets of methods (see main paper, Figure 3d). The integration of further methods or further compendia generally led to an increased performance.

The second improvement is motivated by the fact that genes are frequently regulated by more than one TF and that several (often functionally related) TFs regulate overlapping sets of targets (Reményi *et al.*, 2004). Local methods predict targets for a single TF at a time and cannot take such combinatorial regulation into account. We therefore encoded the set of known regulators of a gene as additional training data (see Section 2.6.8). The true targets of each modeled regulator are excluded from the training to avoid over-fitting. We observed that the explicit encoding of known regulations roughly doubled the number of P50 interactions, yielding 18,724 interactions (see main paper, Figure 3c). This corresponded to a threshold on the κ confidences of 0.92.

Finally, we aimed to include gold-standard interactions predicted with moderate confidence. We therefore extended the predicted network by gold-standard regulations that met a relaxed confidence threshold of 0.46 (compare P50=0.92, see main paper, Figure 2c). The fact that gold standard interactions have been determined experimentally provides an increased confidence, justifying the relaxation of the threshold. This further increased the size of our final yeast network to 22,231 interactions containing 153 TFs and 3,747 target genes. Of all predicted regulations, 12,869 are contained in the gold standard while 9,362 are novel predictions. Among the 29,398 gold standard interactions (Figure 4a), even by the already relaxed threshold, more than half (56.2%) were not confirmed by our approach. These 'quiet' interactions apparently have no regulatory effect visible in our data.

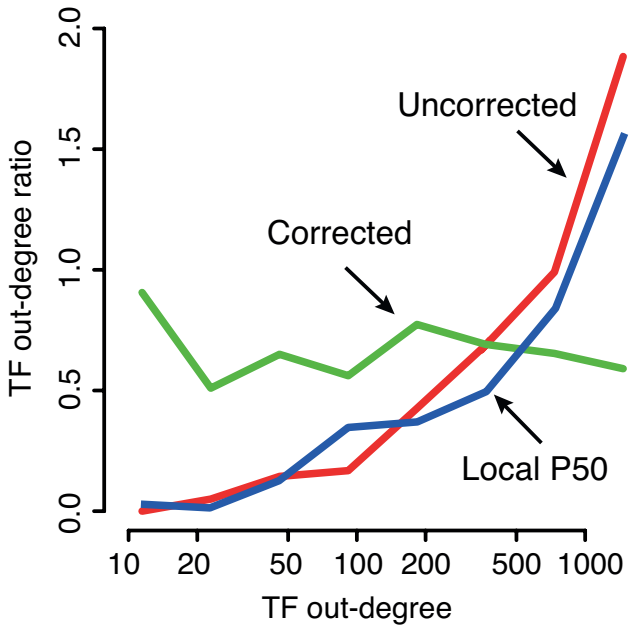

Figure 7: **Influence of TF out-degree on the number of predictions.** The ratio of predicted to gold-standard targets (ordinate) is depicted across the range of TF out-degrees (=number of gold-standard targets, abscissa) before (red) and after (green) recalibration with *CoRe*. Using raw confidence scores, TFs with many targets in the gold standard would receive an overly large number of newly predicted targets. Here, we select the highest scoring targets across all TFs such that a precision of 50% (P50 criterion) is obtained. As an alternative that corresponds to macro-evaluation, the P50 criterion is applied to each TF individually (local P50, blue).

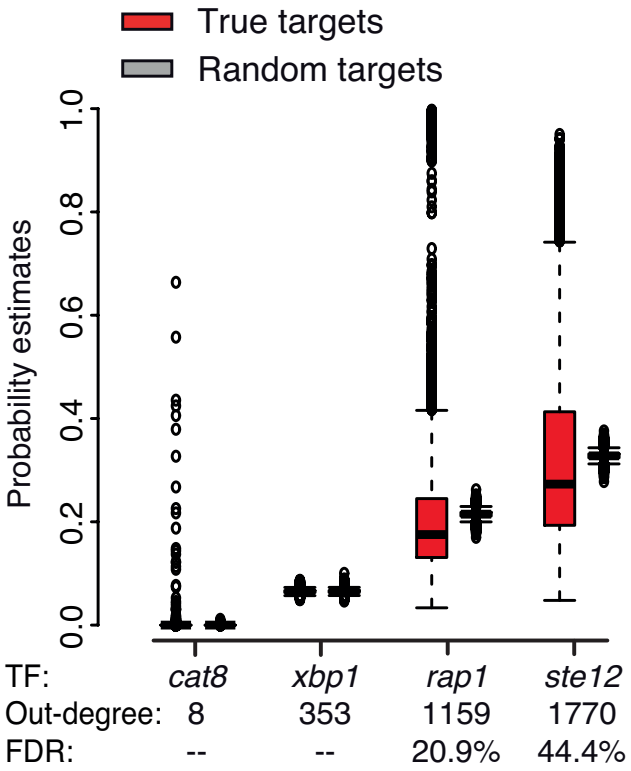

Figure 8: **Degree dependency of probability estimates.** Support vector machines were trained and applied as described in the main document, but resulting confidence scores were transformed to probability estimates (also referred to as Platt scores) via the SVM library libsvm. The probability estimates exhibit increased means and variances in case of confidence score distributions derived for high out-degree TFs. For each TF, the distribution of confidence scores is displayed in a left boxplot for true targets (red) and in a right boxplot for random targets (grey, not visible due to small variance).

3.10 Numerical values of performance estimates

As a reference, the Tables 5 (gold standard, recalibrated) and 6 (randomized networks, recalibrated) provide exact values of the recalibrated network performance measures as depicted by the green bars in Figure 3c of the main paper. When topology features are included explicitly we obtain the values shown in Tables 7 and 8, respectively. The associated values shown are:

- auc := 'area under the receiver operator characteristics curve'
- aupr := 'area under the precision recall curve'
- fmb := 'optimal f-measure for variable threshold'
- p50 := 'number of predictions for a precision of 50%'

In addition, these tables summarize various evaluation approaches and contain further performance estimates such as the F-measure. We compare several evaluation setups, including micro- *vs.* macro-averaging, the influence of additional training features encoding known regulators, raw *vs.* recalibrated confidence scores as well as experimentally derived gold standard *vs.* random networks. See Section 2.5 for details on the scores and their computation.

Due to the long run-time, the random networks were not processed via CLR and, thus, calibrated scores were not computed. Note that the consensus is constructed from the five approaches employing local models, namely Random forest, Decision tree, Lasso, Elastic net and local SVM, which corresponds to the SIRENE approach.

Even without recalibration, macro-evaluation takes a regulator-wise viewpoint and thus enables sensible local performance estimation that may complement the network-wide point of view. Detailed results are shown in Tables 5 and 7. However, macro-evaluation does not provide a mechanism to select a interactions from a wide range of degrees. Due to the degree dependency of confidence scores the resulting networks will preferentially consist of larger regulators (compare Section 3.7).

Table 5: Evaluation for fold-change expression features (measure definition, see 3.10).

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
CLR	47.9	2.8	5.7	14	52.0	3.7	7.8	-	-	-	-	-	-	-
One-class SVM	50.3	6.5	12.5	0	50.9	4.0	8.3	51.4	3.6	6.4	0	50.9	4.0	8.3
Correlation	54.5	5.1	7.0	514	54.9	5.6	10.4	54.5	5.1	7.1	456	54.9	5.6	10.4
SEREND	81.5	22.4	28.6	3684	60.8	12.5	18.6	70.4	21.1	28.5	3332	61.2	12.9	19.2
Global SVM	81.8	21.2	29.1	2692	67.6	11.4	17.3	71.6	15.8	21.3	3054	67.6	11.4	17.3
Random forest	79.6	24.0	28.2	6996	63.1	11.4	17.8	62.9	17.9	23.1	5554	63.1	11.4	17.8
Decision tree	79.3	19.5	23.8	4110	66.1	13.3	20.1	65.3	17.0	23.1	4426	66.1	13.3	20.1
Lasso	81.4	25.7	29.9	8000	67.5	13.5	20.0	65.0	19.7	26.5	6490	67.4	13.5	20.1
Elastic net	81.6	25.9	29.9	8090	67.8	13.7	20.3	65.5	20.1	26.7	6576	67.8	13.7	20.3
Local SVM	78.4	28.2	33.5	10466	68.3	16.2	23.2	70.9	23.6	29.8	9130	68.3	16.2	23.2
Consensus	82.2	28.7	32.3	9918	68.5	15.2	21.8	69.3	23.5	29.3	8726	68.5	15.8	22.8

Table 6: Evaluation for random networks on fold-change features.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
Random forest	72.9	9.9	16.1	0	49.4	3.1	6.4	47.8	3.5	7.5	0	49.4	3.1	6.4
Decision tree	69.9	5.2	10.1	0	49.6	3.1	6.3	48.4	3.2	6.4	0	49.6	3.1	6.3
Lasso	71.4	7.8	14.2	0	49.1	3.1	6.4	47.7	3.9	8.9	0	49.1	3.1	6.4
Elastic net	71.6	7.9	14.5	0	49.0	3.0	6.2	48.0	4.1	9.4	0	49.0	3.0	6.2
Local SVM	67.2	9.4	16.0	20	49.8	3.1	6.4	50.2	3.1	5.8	0	49.8	3.1	6.4

Table 7: Evaluation results for features extended by topology information.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
CLR	46.9	2.8	5.7	6	38.0	3.2	6.7	-	-	-	-	-	-	-
One-class SVM	51.0	6.8	12.9	14	52.2	4.4	9.0	52.5	3.8	6.8	8	52.2	4.4	9.0
Correlation	56.7	5.5	7.7	504	59.4	9.3	14.6	56.7	5.5	7.7	480	59.4	9.3	10.4
SEREND	81.5	22.4	28.6	3684	60.8	12.5	18.6	78.5	24.3	31.6	5364	67.3	18.5	25.6
Global SVM	87.7	27.4	36.3	2630	79.1	17.8	23.3	82.7	20.5	26.2	2692	79.1	17.8	23.3
Random forest	85.4	34.3	37.8	13374	74.2	20.3	27.4	74.2	30.5	36.2	12990	74.2	20.3	27.4
Decision tree	86.1	36.4	39.3	15188	76.2	26.4	33.0	75.7	34.3	39.3	15586	76.2	26.4	33.0
Lasso	86.8	37.1	39.5	15030	77.6	23.2	30.5	73.7	30.6	37.1	13382	77.6	23.2	30.5
Elastic net	86.8	37.0	39.4	15046	77.8	23.5	30.9	73.8	30.6	37.1	13160	77.8	23.5	30.9
Local SVM	85.0	38.8	42.4	17520	77.9	28.3	35.5	80.3	36.8	41.6	17194	77.9	28.3	35.5
Consensus	88.1	42.0	43.3	18472	79.5	27.6	34.1	79.3	39.9	43.5	18724	79.5	29.8	36.5

Table 8: Evaluation results for features on random networks extended by topology information.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmf	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
Random forest	73.0	9.9	16.1	0	49.6	3.1	6.4	48.2	3.6	7.7	0	49.6	3.1	6.4
Decision tree	70.4	5.3	10.3	0	49.9	3.1	6.3	48.7	3.3	6.6	0	49.9	3.1	6.3
Lasso	71.6	8.0	14.5	2	49.1	3.0	6.2	47.9	4.0	9.0	0	49.1	3.0	6.2
Elastic net	71.8	8.1	14.8	0	49.1	3.1	6.2	48.1	4.1	9.3	0	49.1	3.1	6.2
Local SVM	67.8	9.5	16.2	26	50.0	3.1	6.4	49.8	3.1	5.8	0	50.0	3.1	6.4

4 Inference approaches, tasks, and issues

The present study is focused on gene regulatory network (GRN) inference methods that integrate prior knowledge of previously known regulations. Similar approaches have been developed for many fields of application. We have discussed critical issues in terms of prediction and evaluation at the example of GRN inference. Yet, the prediction of protein-protein interactions, drug-target interactions as well as gene function often resemble the core functionality. We thoroughly review a wide spectrum of methods in Section 5. In particular, we discuss common properties that could eventually explain the prevalence of certain observations in these fields as well.

4.1 Shared features among inference approaches

Certainly, categories for existing inference methods cannot be assigned in general. Still, focusing on shared features provides the means for a more detailed discussion. Below, we describe some key features to distinguish existing approaches. Selected methods and methodological details are given in Section 5.

1. Mode of network utilization
    - Unsupervised. Only expression data is used for training (Section 5.4), *i.e.* an interaction is usually predicted if the TF and a putative target gene exhibit mutual dependency in their expression profiles. In general, no prior knowledge is integrated.
    - Supervised. Structural priors or experimentally supported interactions are used (*i.e.* gold standard interactions, see Section 5).
2. Prior Integration
    - One-class. Interactions not contained in the gold standard are treated as unknown rather than non-existent (Section 5.1)
    - Two-class. Interactions not contained in the gold standard are treated a non-existent (Section 5.1).
3. Modeling Strategy
    - Lazy. Putative targets of a regulator are scored by comparing their gene expression profiles to the expression profiles of known positive or negative targets. No model is built. (Section 2.3).
    - Eager. A parameterized model is derived using expression profiles of known targets (Section 5). The model is then used to predict novel interactions.
4. Data handling and integration
    - Integrative. Other data besides gene expression or the known interactions is incorporated, *e.g.*, TF binding site sequences (Section 5.2).
    - Non-integrative. Only expression data and part of the known interactions are used for training (remaining methods in Section 5).
5. Model training strategy
    - Global. A single model is trained (Section 5.3) integrating expression profiles as well as the known interactions across all TFs and all putative target genes.
    - Local. TF-wise models are trained (Section 5.1) each using expression data to predict putative targets of a single TF at a time. Each of these models is then called a local model.

4.2 Types of inference tasks

Typically, most inference methods (see main paper, Section 2.2) take prior knowledge into account. A powerful way to represent such knowledge is by networks. In case of gene regulatory inference they would summarize knowledge on known targets of regulators or transcription factors (TFs). kinds of relationships including gene regulatory interactions are subsumed under the generic term *functional associations*. Different kinds of functional associations between proteins or other biological entities further encompass protein-protein interactions, drug-target interactions as well as protein annotations. The latter are associations that link proteins to biological processes or protein functions. A related important concept is that of a *gene set*. In case of interactions, a TF is associated with all genes contained in a corresponding gene set that comprises all target genes of this TF. A biological process is assumed to be described by a gene set containing all genes known to be relevant for that process. Essentially, a gene set is specific to a given entity such as a TF or a biological process and covers a given type of functional association. Supervised inference can thus be applied as shown in Figure 2 of the main paper to infer functional associations of interest if (i) suitable datasets are available and can be structured as a data matrix and (ii) prior knowledge can be provided as a (partial) set of genes in the form of a label vector. Again, in case of interaction inference, (i) could be large scale gene expression data and (ii) would be (part of) the genes regulated by a given TF.

In this context, the prediction of functional associations or simply function prediction (FP) and the prediction of regulatory interactions may be regarded as special cases of the same concept where function is a common property among a set of genes such as the targets of a single transcription factor. Function prediction is of interest in many biological use cases as predictive models complement prior knowledge and thereby enable a deeper understanding of both novel and existing associations. In the process of prediction, associations across different TFs or biological processes are prioritized based on prediction confidence to enable the selection, evaluation or experimental follow-up of promising candidates.

4.3 Shared issues

The prediction of different kinds of functional associations shares important properties and issues. For instance, Myers *et al.* (2006) focus on predicting gene functions, *i.e.* pathway-gene associations that are predicted separately for each pathway of interest and thus, on predictions derived via local models. They argue that the evaluation of functional annotation may be influenced by the uneven size and different properties of certain biological processes. Inclusion or exclusion of the ribosome pathway (among 98 other KEGG pathways) makes the difference between co-expression data being the most or least, respectively, informative dataset. Myers *et al.* thus conclude that each process should be evaluated in isolation to overcome *HDP*. However, this might not be a practical solution if functional associations must be obtained across biological processes or if transcriptional networks must be obtained across TFs.

It is further important to note that, *vice versa*, proteins spanning broad range of functions are much more likely to be confirmed as correct members of an arbitrary functional class in comparison to specific proteins with a narrow range of functions. For protein-protein interaction networks, Gillis and Pavlidis (2011) discuss that the number of functions a protein exposes is coupled to its node degree, *i.e.* the number of interaction partners, and show that for an arbitrary functional category being predicted a ranking based on the network node degree will perform better than expected by chance and can thus unintentionally skew quality estimates. This is referred to as *multi-functionality bias*. Gillis and Pavlidis conclude that there are no suitable techniques available that substantially reduce it without undesired side-effects. In particular, Pavlidis and Gillis (2013) argue that an entirely different problem structure may arise, such that it may often be unclear whether a fix is preferred or not.

We argue that for network inference a recalibration like *CoRe* is preferred. The regulator-wise recalibration using an empirical FDR yields sensible overall networks that maintain the regulator-specific quality of their underlying models.

In contrast to these effects, De Smet and Marchal (2010) and Ambroise *et al.* (2012) observed related preferences in the selection of predicted interactions. They describe that TFs with many known targets receive disproportionately many predictions while hardly any predictions are assigned to TFs with few known targets. De Smet and Marchal (2010) conclude that this is due to the fact that less information is available for the TFs with few known targets and that, based on this observation, supervised approaches should not be applied to infer interactions for TFs with few known interactions. In our present paper, we provided

evidence that independent confidence distribution lead to a skewed overall integration. We demonstrated that *HDP*can be tackled by an appropriate recalibration. While for larger regulators some sensitivity in detecting true novel regulations may thereby be lost, the amount of false predictions is drastically reduced. For most smaller regulators *CoRe*boosts sensitivity and the true positive rate and enables the prediction of novel targets, even in case of low-degree TFs.

Common protocols for supervised prediction using local models have been established by SIRENE (Mordelet and Vert, 2008) and SEREND (Ernst *et al.*, 2008). Approaches that are prone to *HDP*use similar or derived protocols unless denoted otherwise.

5 Related work

This section reviews some twenty methods devised for functional association prediction, focusing on approaches that take prior knowledge into account (usually by training supervised machine learners) in contrast to expression-based (unsupervised) approaches that do not. Methods are categorized into different types as described in the previous section, which is reflected by corresponding subsections in the following. Finally, the last part of this section lists review articles or reports on comparative assessments of inference approaches.

This multitude of inference methods, tasks and corresponding publications demonstrates their wide application. We further highlight how the majority of current supervised inference methods traces back to the approaches we examined in the main paper, indicating that many of their basic properties and issues will be shared.

5.1 Local models

In their seminal paper, **Bleakley *et al.* (2007)** introduce local models for the reconstruction of biological networks. The inference of two network types is addressed at the example of yeast, that of metabolic gene networks and protein-protein interaction (PPI) networks. For the prediction of the metabolic gene network, expression data from (Eisen *et al.*, 1998) and (Spellman *et al.*, 1998) with 157 experiments, a localization vector containing information about appearance of an enzyme in 23 locations (Huh *et al.*, 2003) and a phylogenetic profile about presence or absence of an enzyme in 145 fully sequenced genomes (Kanehisa *et al.*, 2004) are used. Yeast two-hybrid data from (Ito *et al.*, 2001) and (Uetz *et al.*, 2000) are additionally used for the prediction of the PPI network. The gold standard for the metabolic gene network proposed by (Yamanishi *et al.*, 2005) consists of 668 enzymes and 2782 functional relationships. The gold standard for the PPI network of (von Mering *et al.*, 2002) was reduced to 2438 so-called high confidence interactions among 984 proteins.

For each of the involved proteins, an individual SVM is trained to predict its interaction partners, *i.e.* one local model per enzyme or protein is built. The predicted decision values are combined to rank possible interactions between enzymes and proteins and to evaluate the performance of the method via the AUC.

Mordelet and Vert (2008), adopt the local models of Bleakley *et al.* (2007), for the **SIRENE** approach to infer gene regulatory interactions from *E. coli* expression data. The gold standard and the expression data are the same as in **Faith *et al.* (2007)**. The expression data contains 445 microarray measurements of 4345 genes under different experimental conditions. The gold standard contains 3293 regulatory interactions between 154 TFs and 1211 genes. For each TF, a local SVM model is trained. Separate predictions for the individual TFs are combined to obtain a global ranking of previously known and possibly novel interactions and to subsequently evaluate inference performance via the AUC and the AUPR. The highest scoring interactions are selected as long as a defined precision level is reached.

Bleakley *et al.* (2009) extend the concept of local models used in (Mordelet and Vert, 2008; Bleakley *et al.*, 2007) to the bipartite network problem of drug-target interactions. In this setting, one drug can interact with many proteins and one protein can interact with many drugs, but drugs and proteins are not considered to interact among themselves. The interaction data are derived from several databases containing between 45..445 known drugs as well as 26..664 target genes connected by 90..2926 interactions, respectively. The drugs are represented by a similarity matrix calculated from their chemical structure. The proteins are represented in a similarity matrix computed from their sequence similarity. Local models are used to predict drug-target interactions. Therefore, with the use of SVMs, a local classifier is trained for each drug to predict target proteins and for each protein to predict interacting drugs. For each possible drug-target interaction both predictions are combined for a final prediction. All interactions are integrated and ranked by their integrated SVM decision values.

In **Mordelet and Vert (2010)** the influence of bootstrap aggregation with SVM on learning local models from positive and unlabeled data is examined. In this supervised classification approach, positive and negative examples are required to learn a classifier, which can distinguish between the two classes. As there is a lack of negative examples, unlabeled examples are often treated as negative. In particular, different strategies of using unlabeled data as negative examples are determined. The performance is evaluated with different bagging strategies on various data sets including *E. coli* data from (Mordelet and Vert, 2008). For

this data set, the SIRENE approach (local models) is used with bagging and compared to an SVM without recalibration, a 1-class SVM (outlier detection) and a baseline method. The bagging strategy showed an improvement in the performance.

Yip *et al.* (2009) apply the local modeling approach for the prediction of protein, domain and residue interactions, referred to as levels. To improve the prediction of interactions, Yip *et al.*, 2009, use support vector regression combined with a training set expansion approach to learn local models at each interaction level. The models of the three levels are trained in turn and the best predictions of each model are added to the gold standard of the next level (unidirectional flow) or all levels (bidirectional flow), which thereby get additional training examples. This multi-learning approach is implemented by multiple kernels and tested by predicting protein, domain and residue interactions in yeast. For each level appropriate data features were collected: data of the protein level comprise phylogenetic profiles, sub-cellular localization, gene expression and two networks from yeast two-hybrid and TAP-MS data. The final kernel is a sum of all of these kernels. The protein interaction gold standard was constructed of known protein interaction data from MIPS, DIP and iPfam containing 1681 proteins and 3201 interactions. The data features of the domain level contain co-evolution and statistics related to parent proteins. The domain interaction gold standard was created from iPfam data. The domains are defined as interacting, if they are close enough and predicted to form a bond. The gold standard comprises 422 domain interactions of 317 domains. For the residue level data from PSI-BLAST profiles, secondary structure and surface areas of the residues and its neighbors are used. The gold standard taken from iPfam determined 3053 residues and 2000 interactions based on their proximity in known crystal structures of interacting proteins. The results in bidirectional flow show an improvement in the predictions.

Geurts (2011) examine the problem that only a small subset in a large set of objects is labeled as positive with respect to the class of interest and it is unknown to which degree the unlabeled set contains unknown positive examples. Geurts and colleagues formalize this as "the problem of learning a feature based score function that minimizes the p-value of a non parametric statistical hypothesis test". A solution for this problem is shown for a linear scoring function using local one-class SVM models applied on sampled subsets of the complete set of objects (called PU method). This approach is tested on yeast and *E. coli* expression data (6178 genes, 157 features per gene and 4345 genes, 445 features per gene, respectively). The gold standard of yeast contains 80 TFs, 606 targets and 1164 interactions. For each operon a representative gene was selected and other genes from these operons were removed from the data set (reduced to 2925 genes) and gold standard (63 TFs, 554 targets, 1446 interactions). The PU method was compared to two-class SVM, one-class SVM, CML and CORR. CORR is a baseline method, which ranks genes according to their average correlation with genes in the positive set. This method was also used in our approach as baseline method. The AUC performance of the method shows that the PU method performs best on yeast and second on *E. coli* using two-class SVM.

Function prediction is also usually performed via local models (Lanckriet *et al.*, 2004; Mostafavi *et al.*, 2008). Here, gene sets represent biological processes or protein functions, *e.g.*, derived from the GeneOntology (The Gene Ontology Consortium, 2010). For each process, a dedicated binary classifier is trained to predict putative members. The training is based on a data matrix of several heterogeneous fused data sets representing information on protein domains, protein interactions, gene expression measurements or pairwise protein sequence alignment scores. In particular, **Lanckriet *et al.* (2004)** used the combination of five data sets to train SVM models and predict the functional categories of yeast proteins. This approach was tested in two different settings. For the first setting, the following data was used: (1) the domain structure of each protein (Pfam domains, 4950 bit vector for each protein), (2) protein-protein interactions (from CYGD), (3) genetic interactions, (4) co-participation in a protein complex (determined by tandem affinity purification) and (5) 77 cell cycle gene expression measurements per gene converted to a binary square matrix. For the second setting the matrix of the domain structure contained 5724 domains (Pfam 9.0) and log E-values instead of binary values. The expression matrix is used without conversion to binary values and additionally a matrix containing Smith-Waterman scores for each protein is calculated. The prediction associates yeast proteins with 13 functional categories. For each functional category and data set a SVM is trained and the kernels are linearly combined across data sets for category-specific prediction. For each functional category a 5-fold CV is conducted 3 times. The results are compared against the MRF method of Deng *et al.* (2004). The SVM approach reaches a mean AUC of 0.870 using the second setting. Both settings outperform the MRF method. The size of the categories are imbalanced, ranging between 81

and 1048 members, suggesting the susceptibility of these methods to Simpson's paradox.

Mostafavi *et al.* (2008) have developed GeneMANIA (Multiple Association Network Integration Algorithm), a fast heuristic algorithm derived from ridge regression for gene function prediction. The function prediction is treated as a binary classification problem. In the first step functional association networks from heterogeneous input data sources are created. Linear regression is used to integrate multiple functional association networks into a single composite functional association network. The single composite association network contains the weighted average of the individual functional association networks. For final gene function prediction, an adopted Gaussian field label propagation algorithm for unbalanced classification problems is used. GeneMANIA is tested on several function prediction benchmarks. In the prediction of twelve evaluation categories (GO categories from biological processes, cellular components and molecular function) GeneMania reaches equal or improved performance over previous approaches considering this benchmark. On the yeast benchmark, GeneMANIA is compared against two algorithms in the prediction of 400 GO functional classes. GeneMANIA performs as well as or better than the compared methods. Furthermore, it is much faster than other approaches and function predictions can be calculated on-the-fly.

Brown *et al.* (2000) use SVMs to predict the functional class of yeast genes using gene expression data. The performance of the SVMs in five functional classes is compared against four competing machine learning methods: Fisher's linear discriminant, Parzen windows, and two decision tree learners (C4.5 and MOC1). Yeast expression data containing 79 features of 2467 genes is used for the prediction. For each functional class, a local model is learned and the performance is evaluated per class using the cost savings measure. The radial basis function SVM outperforms the other methods tested.

Text classification or **text categorization** is a frequent task in machine learning and information retrieval. It addresses the automatic classification of documents into categories. It is applied for spam filtering, identification of document genre or indexing of scientific articles. Given that manually classified documents are available, they can be used for training machine learning methods like SVMs. Precision, recall and F-measure are calculated for performance evaluation of these methods. For performance evaluation precision and recall is calculated using micro- and macro-averaging, *i.e.* across single predictions across categories or category-wise. Similar to interaction prediction, text mining features highly imbalanced classes based on the underlying categories.

In particular, a defined class would cover only a fraction of all possible annotated entities Sebastiani (2005). GRN inference and text classification thus share several problems. In particular, an evaluation setup may suggest reasonable performance for some compendium spanning multiple categories. Yet, large and well-defined categories with many training examples may govern the evaluation. A category-wise evaluation would then uncover near-random performance for most categories except some of the largest ones. Following Simpson's Paradox it then unclear which of two methods sharing the same overall quality should be preferred. Either, compendia-wise quality is considered, or a recalibration is done that allows for the selection of the most confident predictions (or those at a specific FDR rate) in the complete compendium.

Özgür *et al.* (2005) propose an approach for text categorization using SVM and keyword selection for all classes (corpus-based) or for each class separately (class-based). For this study, a standardized data set consisting of 21578 documents from 135 categories was used. The maximum number of categories assigned to a document is 14, the average number of categories is 1.24. Training and testing are based on 9603 and 3299 documents, respectively. After removing categories only present in test or training set, 90 categories remain in the gold standard. For performance evaluation, the macro- and micro-averaged F-measure is calculated. The micro-averaged F-measure shows that class-based keyword selection performs better than corpus-based keyword selection with up to 1200 keywords. Keyword selection performs equally compared to the use of all words for classification. The macro-averaged F-measure is lower than the micro-averaged F-measure and increases with increasing number of keywords.

5.2 Integrative approaches

Integrative methods incorporate data from different sources for prediction. An integrative state-of-the-art method for GRN prediction proposed by Ernst *et al.* (2008) is **SEREND**, that utilizes TF binding site information and expression data. Three logistic regression classifiers are trained: (i) using expression data, (ii) using binding sites and (iii) using the two initial predictions to via a meta classifier. Classifiers are trained separately for each TF. Thus, SEREND is also based on local models. The three classifiers generate

ranked predictions of targets for each TF. The authors used SEREND to predict GRIs of *E. coli* using a gold standard of 123 TFs, 974 genes and 1760 interactions. See Section 3 for a discussion of *HDP*in the SEREND approach.

Yip and Gerstein (2009) adopt local models and propose two approaches to improve the reconstruction of biological networks: the training set expansion using prediction propagation and kernel initialization. These two approaches are tested on eight different data sets (phylogenetic profiles, sub-cellular localization, gene expression from environmental response, cell cycle gene expression, 2 different yeast two- hybrid data sets and 2 different tandem affinity data sets) and on an integration of the eight data sets for the prediction of protein-protein interactions. Known protein-protein interactions of yeast are derived from BioGRID, DIP, MIPS and iPfam to create three differently-sized gold standards. Both approaches are compared against a range of other methods (direct, kCCA, kML, em and Pkernel) and tested with two different cross-validation strategies (10-fold cross-validation and random sampling of negative training set). Prediction propagation and kernel initialization are combined with local models. The local model approach in conjunction with the training set expansion approaches show a higher AUC for most data sets than other methods (on using BioGRID-10 gold standard with 2880 interactions).

5.3 Global models

Qian *et al.* (2003) propose a supervised approach based on SVMs to predict TF:target interactions in yeast from expression data. In contrast to local models, a single global SVM model is trained to predict gene regulatory interactions across TFs (see main paper, Section 2.2). The expression data is obtained from two yeast gene expression studies (Spellman *et al.*, 1998; Gasch *et al.*, 2000) and contains 79 measurements from different time points during the diauxic shift, the mitotic cell cycle, sporulation and heat shock for each gene. The expression vector of each target is concatenated to the expression vector of each TF, so that each putative TF:target interaction is characterized by a $2 * 79 = 158$-element gene expression vector. The gold standard contains 36 TFs with 175 interactions obtained from TRANSFAC (Wingender *et al.*, 2001) and SCPD (Zhu and Zhang, 1999). On this data, a global SVM model predicts 46059 putative TF:target interactions. The performance evaluation shows that the inference of TF:target interactions using global models is possible. It is important to note, however, that this work analyzes a very small gold standard and a very small expression compendium. It has been reported in Yip *et al.* (2009) that methods based on global models suffer from both time and space complexity and are thus difficult to apply to more comprehensive data sets and gold standards.

In contrast to most of the previously described inference approaches, **Cai *et al.* (2007)** focus on exploiting functional annotations rather than expression data. The approach is based on nearest neighbor selection and is applied to the global prediction of TF:target interactions. For each gene the associated GO_compress (GO_c) entries were extracted and a boolean vector of size 3860 (overall number of GO_c entries) was created. This vector contains *true*, if the gene belongs to the respective GO_c entry and *false* otherwise. If no GO_c entries were available for a gene, a gene expression vector is used instead. 3543 genes were defined by GO_c entries and 88 genes by expression vectors. For similarity calculation the TF-vector and target-vector was combined, as described in Qian *et al.* (2003). The gold standard was obtained from Qian *et al.* (2003) (see above) and contained 175 interactions of 36 yeast TFs. For similarity calculation the GO_c vectors were used if available, otherwise the expression vectors. The procedure resembles that of Qian *et al.*, 2003 (Qian *et al.*, 2003), and shows equal performance, suggesting that both methods may neglect regulator specific quality.

Seok *et al.* (2010) propose a variant of the approach of Qian *et al.*, 2003 (Qian *et al.*, 2003) for GRN inference in yeast. They use a data set consisting of 643 microarrays of 5940 genes and a gold standard consisting of 3043 interactions, 523 TFs and 919 targets. The examination of the yeast expression data by Seok *et al.* reveals that the correlation of the expression data of TFs and their targets is similar to the correlation of randomly selected gene pairs. A so-called centroid representation of TFs is calculated and used for prediction instead of real expression data (naïve representation). The global model SVM approach of Qian *et al.* is tested using the naïve representation and the centroid representation. It is compared to a correlation cutoff method, CLR and SEREND, which were also tested with both representations. The experiments show that the SVM benefits from a centroid representation and that it can improve the prediction compared to a naïve representation. Notably, the centroid approach of Seok *et al.*is based on Qian *et al.*, 2003 and may

thus inherits the problem of confidence integration in the presence of *HDP*.

5.4 Learning from derived features

TNIFSED by **Ambroise *et al.* (2012)** does not directly exploit expression data for inference but processes the input data to derive new features. The authors state two limitations of SIRENE: (i) the performance of SIRENE decreases proportionally to the number of known targets and (ii) it cannot be used for the prediction of novel TFs or those without known targets. TNIFSED instead builds a global logistic regression model to infer the probability of observing an interaction among TF and target. As input data, correlation scores obtained from the expression data and functional similarity scores are used. The functional scores determine the similarity between TFs and targets based on gene ontology categories. The TFs are not explicitly represented in the TNIFSED approach. As it seems not possible for TNIFSED to capture information on the number of TF targets (*i.e.* its out-degree) it does not seem to be susceptible to the effects of Simpson's paradox. Yet, it shows a substantially lower performance in case of *E. coli* and *S. cerevisiae* if compared to SIRENE.

The approach of **Holloway *et al.* (2007)** builds on derived features from 26 different data sets (*e.g.*, motif hits, phylogenetic profiles, expression correlation, GO term profiles, K-mers) for yeast TF binding site identification. The gold standard contains 104 TFs and 9104 interactions extracted from three different sources: ChIP-chip experiments (Harbison *et al.*, 2004; Lee *et al.*, 2002), Transfac 6.0 Public (Matys *et al.*, 2006). For each TF and each data set, four different SVM classifiers are built (linear, RBF, Gaussian and polynomial) and the function with the highest F1 score is selected as best for that particular TFdataset combination. For each TF, a weighted composite classifier (*i.e.* a local model) is constructed, which is a weighted combination of the 26 selected classifiers. For the selection of positive targets from the predictions, a global threshold is determined. In contrast to the previous approach (Ambroise *et al.*, 2012), Holloway *et al.* (2008) essentially employ a local model approach inheriting its properties and issues.

Holloway *et al.* (2008) published a variant of their former approach (Holloway *et al.*, 2007) adapting method, gold standard and feature data-sets to revisit construction and analysis of the gene regulatory network in yeast. Eight different types of features are used to describe genes: k-mers, k-mers with mismatch, melting temperature profile, homologous conservation, k-mer median positions, expression data, k-mer over-representation and conserved k-mers. The updated gold standard features 9983 interactions among 163 TFs and 3482 targets. For the prediction of interactions, SVMs combined with feature reduction approaches is used. 50 models are trained for each TF for the induced 1500 features using probability predicting Platt's SVM formulation. The procedure is repeated 100 times and the probabilities for each interaction are averaged. Overall, interactions exceeding a threshold of 0.95 are considered to be true interactions. We found that the use of Platt scores as opposed to raw SVM decision values are no sufficient way to tackle confidence value integration for *HDP*. It could reduce the need for confidence recalibration, yet the implicit correction is likely incomplete as an excessive number of targets is predicted for TFs with a high out-degree in the gold standard (see Section 3).

Bauer *et al.* (2011) introduce a novel machine learning approach, **RIP (Regulatory Interaction Predictor)** to infer interactions in human. Therefore, 4064 primary human tissue samples of 76 experimental conditions are obtained containing expression levels of 13,069 genes. For these genes 81 mid-range GO terms are determined for functional categorization. The gold standard and PWMs is derived from the TRANSFAC database (v2009.2). The gold standard contains 303 TFs with 2896 interactions for 949 targets. The remaining 248,641 interactions are considered to contain no interactions. The network topology, TF binding sites and expression data are used to calculate 10 features describing each TF:target pair. Overall, 2000 SVM models are trained using these feature to create an ensemble classifier for performance evaluation and prediction of new candidate regulatory interactions. From the description in Bauer *et al.* (2011), we could neither confirm nor disprove a susceptibility to Simpson's paradox. Furthermore, it is not clear whether TF binding sites (used for training) and gold standard interactions (used for evaluation) are truly separated as both were derived from TRANSFAC.

5.5 Unsupervised approaches

We used the approach of **Faith *et al.* (2007)** as an example of an unsupervised inference approach for performance comparison in the main paper. This approach introduces an extension of the class of relevance network algorithms called Context Likelihood of Relatedness (**CLR**). The approach is unsupervised and features an adaptive background correction step to eliminate false correlations and indirect influences inferred from expression data. CLR is applied to *E. coli* microarray expression data of 4345 genes from 445 profiles to identify the targets of 328 TFs. 3216 known interactions were derived from RegulonDB to assess the performance of CLR and three competing approaches: relevance networks, ARACNE and Bayesian networks. CLR performs best among unsupervised methods on prokaryote data, but is hardly better than guessing on yeast data and is, thus, outperformed by supervised approaches like SIRENE (Mordelet and Vert, 2008).

5.6 Review papers

The seminal review paper of **De Smet and Marchal (2010)** describes and compares network inference approaches across three categories depending on the input used: (i) supervised and semi-supervised *vs.* unsupervised learning, (ii) integrative *vs.* non-integrative and (iii) direct network inference (NI) *vs.* module-based NI *vs.* module inference methods. They compared 14 state-of-the-art module and network inference approaches. The methods are tested on an expression data set of *E. coli* and known interactions from RegulonDB. The authors show that predicted interactions differ to a great degree between the different approaches, which might offer a way to complement predictions. The comparison of the methods shows that performance differs for the various TFs depending on their properties, *e.g.*, the number of targets. A comparison of the predictions of CLR and SIRENE reveals that supervised methods preferentially predict targets for TFs with a high number of known targets.

In their tutorial, **Luts *et al.* (2010)** embed local, SVM-based methods for classification in the context of chemometrics. The SVM models are trained on yeast expression data using a gold standard containing 118 TFs and 4000 interactions. For the resulting regulatory network the interactions exceeding an induced probability of more than 0.95 are considered true interactions. The varying TF out-degrees have not been taken into account.

Vert (2010) reviews pattern recognition algorithms used to infer directed and undirected biological networks from heterogeneous genomic data. The algorithms comprise local and global model approaches to reconstruct metabolic, protein-protein interaction (PPI) and gene regulatory networks of model organisms. Different approaches are analyzed and the performance as indicated by AUC and AUPR is compared across different networks. For the reconstruction of a metabolic network in yeast a gold standard containing 2782 interactions among 668 enzymes from KEGG and three different data sets is used: 157 expression values per enzyme, a 23 bit vector representing localization information of the enzyme and a 145 bit phylogenetic profile defining the presence in fully sequenced organisms. Six methods are compared for the reconstruction of the metabolic network: local models, TPPK and MLPK kernels, *de novo* approach (unsupervised), KCCA, and an algorithm based on an EM procedure. The results show that local models outperform other supervised methods and the unsupervised approach performs worse than supervised methods. For the reconstruction of a PPI network in yeast, a gold standard of 2438 interactions between 984 proteins is used. In addition to the data for metabolic network reconstruction a yeast two-hybrid data set is used. The approaches for the metabolic network resemble that of the PPI reconstruction. The AUC and AUPR values show that the local model approach is reconstructs the PPI network best and the unsupervised method performs worse than supervised methods. The reconstruction of the *E. coli* GRN was tested using the gold standard (154 TFs, 1211 targets, 3293 interactions) and data set (445 microarray expression profiles for 4345 genes) of Mordelet and Vert (2008). The reconstruction of the gene regulatory network is conducted with SIRENE, Bayesian networks, ARACNE, CLR and an extended relevance network algorithm. SIRENE shows a higher recall at a precision of 60% and 80% than the other methods. SIRENE outperforms the unsupervised approach CLR.

Cerulo *et al.* (2010) examined the influence of the selection strategy of negative examples on the performance of supervised machine learning methods. Three different supervised approaches: PosOnly, SVMOnly, and PSEUDO-RANDOM. Here, SVMOnly resembles a default SVM treating all unlabeled interactions as negative. The PosOnly approach introduces an empirically estimated constant factor to correct the prediction results whereas the PSEUDO-RANDOM method selects negative examples from the transitive closure of known interactions. All approaches are examined using both simulated expression data (using

GeneNetWeaver) with varying percentages of known positive examples and experimental *E. coli* expression data. The performance of the supervised approaches is compared to the unsupervised approaches ARACNE and CLR.

Testing is done using simulated *E. coli* and *S. cerevisiae* gene interaction networks. Four different networks of 10, 50, 100, and 500 genes are simulated. The experimental data of *E. coli* contains 445 expression profiles of 4345 genes. The gold standard consists of 3293 interactions, 154 TFs and 1211 genes from RegulonDB. The supervised approaches perform better on simulated data with increasing known positive (10% to 100%) and a perform worse on *S. cerevisiae* than *E. coli* simulated data. PosOnly and PSEUDO-RANDOM perform superior to SVMOnly. PosOnly performs best on simulated data considering the F-Measure. On experimental data the SIRENE protocol is adopted. Here, PosOnly outperforms both SVMOnly and PSEUDO-RANDOM. The comparison to unsupervised approaches shows that PosOnly exhibits similar or better performance. Below a certain fraction of known positives both PSEUDO-RANDOM and SVMOnly perform worse than unsupervised methods, yet if enough interactions are known beforehand the supervised approaches outperform unsupervised methods. As the F-Measure is used for performance comparison the selection of a threshold is required to separate positive from negative predictions resulting in differences among PosOnly and SVMOnly due to the corrective procedure applied in PosOnly. By contrast, when using the AUC, PosOnly and SVMOnly are identical, indicating that the mere ranking of possible targets is not affected. We conclude therefore that this procedure does not rectify the *HDP* problem.

Madhamshettiwar *et al.* (2012) evaluate the performance of unsupervised and supervised machine learning methods on different data sets. The performance of 8 unsupervised methods is evaluated on three types of networks: simulated knock-down and multi-factorial gene expression data sets from the DREAM3 and DREAM4 competitions (*http://www.the-dream-project.org/*), simulated data sets using SynTReN, sampling from known yeast and *E. coli* networks to create sub-networks, and on an ovarian cancer microarray data set. The supervised method SIRENE is applied on DREAM3, DREAM4, the ovarian cancer microarray and on an adenocarcinoma data set. The gold standard of the ovarian cancer network contains 280 TFs, 2170 targets and 6330 interactions. The performance of SIRENE is compared to the best unsupervised method for each data set. SIRENE is found to obtain an increased AUC for DREAM3 and the ovarian cancer data set, but a lower AUC for the DREAM4 data set than the best unsupervised method GENIE. The predictions preferentially belong to TFs with a high out-degree.

Marbach *et al.* (2012) recently conducted a comprehensive comparison of 35 individual unsupervised network inference methods. Among these, 29 have been submitted by participants of the DREAM5 challenge and 6 of them are commonly used methods. These approaches have been tested on four different data sets: (i) an *in silico* network with a gold standard of 195 TF and 1643 genes, and a data set of 805 arrays and 487 conditions, (ii) an *E. coli* network with 296 TFs and 4297 genes, and a data set of the same size than the *in silico* network, (iii) a yeast network with 183 TFs and 5567 genes, and a data set of 536 arrays and 321 conditions, and (iv) a network of the human pathogen *S. aureus* with 90 TFs and 2677 genes, and a data set of 160 arrays and 53 conditions. The approaches are classified into 6 categories: regression, mutual information, correlation, Bayesian networks, meta (combination of approaches) and other (not in previous categories) approaches. The analysis of the methods for different network motifs shows that methods within the same category have a similar performance on the same motifs. To improve the predictions, the results of multiple inference approaches are combined by re-scoring interactions according to their rank. This community network shows better performance than individual methods. The combination of methods from different categories outperforms the consensus performance of similar methods. The combination of strong and weak methods reveals that the weak predictor does not affect the performance. Results on yeast are hardly better than guessing, thus supporting our claim that unsupervised approaches are not suitable to infer interactions in eukaryotes.

6 Discussion of the yeast network

This section provides a detailed literature review on yeast gene regulation focusing on the genes and regulatory interactions contained in the network of Figure 9. Briefly, this modular network representation was derived by applying graph clustering to regulatory interactions (compare Section 2.9). Thereby, TFs regulating overlapping sets of targets were combined into TF modules, targets regulated by overlapping sets of TFs were combined into target modules, and individual interactions were combined into abstract interactions that link the two types of modules. Via this grouping, 11232 individual interactions were represented by a simplified network of 9 TF modules that, via 13 abstract interactions, regulate 9 target modules.

The following literature review consists of three parts. In the first part, we briefly describe examples for the detection of novel active and the pruning of quiet regulations. In the second part, we focus on the target clusters, their biological functions, their expression patterns, and their interactions with the specific TF clusters. The corresponding 9 TF clusters are described in the third part showing that TFs in TF modules jointly regulate specific biological processes. This third part describes the regulatory functions of the TF clusters and thus contains most of the literature references. The order in which clusters are discussed as well as the chosen cluster representative gene correspond to Figure 9.

This review demonstrates that genes in target gene modules constitute defined biological processes and that their observed expression behaviour can be plausibly derived and interpreted from the corresponding TF modules and their action under the measured experimental conditions.

6.1 Novel predictions

In the following, we briefly describe examples (i) for novel predictions missing in current gold standards (the activation of *cat2* and *tes1* by *pip2/oaf1* and *adr1*) as well as (ii) for an interaction contained in the gold standard not supported by our predictions (the regulation of *hap4* by *cat8*). This latter interaction may thus be an example for a 'quiet' interaction not associated with expression changes of the target.

Genes involved in peroxisomal beta-oxidation in *S. cerevisiae* are repressed in the presence of glucose, de-repressed on non-fermentable carbon sources such as ethanol, and further induced by more than ten-fold in the presence of oleate (Gurvitz and Rottensteiner, 2006). Examples of gene products involved in the breakdown of fatty acids include *pot1*, *pox1*, *fox2*, *sps19*, and *cta1*. The transcriptional up-regulation of these genes is driven by the *pip2/oaf1* transcription factor, binding to the oleate response element (ORE), and by *adr1*, binding to another upstream activating site, UAS1 (Hiltunen *et al.*, 2003). *Cat2*, a carnitine O-acetyltransferase, and *tes1*, an acyl-CoA thioesterase are also enzymes involved in fatty acid breakdown, currently postulated to be regulated by *pip2/oaf1* (Hiltunen *et al.*, 2003). We predicted that the transcription of *cat2* and *tes1* is also activated by *adr1*, which has not been reported before (or only indirectly as for *tes1* (Smith *et al.*, 2007)) but seems plausible given the known regulation of beta-oxidation genes by *pip2/oaf1* and *adr1*.

Cat8 and *hap4* are major transcriptional regulators of the diauxic shift (Schüller, 2003). *cat8* especially activates the transcription of gluconeogenic genes via binding to a carbon source responsive element (CSRE) in their promoter. *Cat8* itself is transcriptionally regulated in dependence on the carbon source, where positive regulation on non-fermentable carbon sources is carried out by the *hap2/3/4/5* complex (Turcotte *et al.*, 2010). *Hap4* is the activator subunit of the *hap2/3/4/5 complex*, especially driving the expression of genes involved in respiration and the TCA-cycle. *hap4* is also the regulatory subunit of the complex, as it is the only one whose level is regulated by the carbon source itself. Interestingly, it seems that *hap4* and *cat8* are mutually activating each other, as *hap4* transcription has been shown to be *cat8*-dependent (Brons *et al.*, 2002). In our network, the regulation of *hap4* by *cat8* was not predicted. This is in agreement with current studies, which assign the carbon source dependent regulation of *hap4* rather to *rds2* (Turcotte *et al.*, 2010).

6.2 Target clusters

IMP biosynthesis & peptidase inhibition (*ade1*). The majority of genes in this cluster perform functions in the *'de novo' IMP biosynthetic process* and *peptidase inhibition*. As judged from the expression data, we observed genes with peptidase inhibitor activity to be strongly up- and down-regulated under

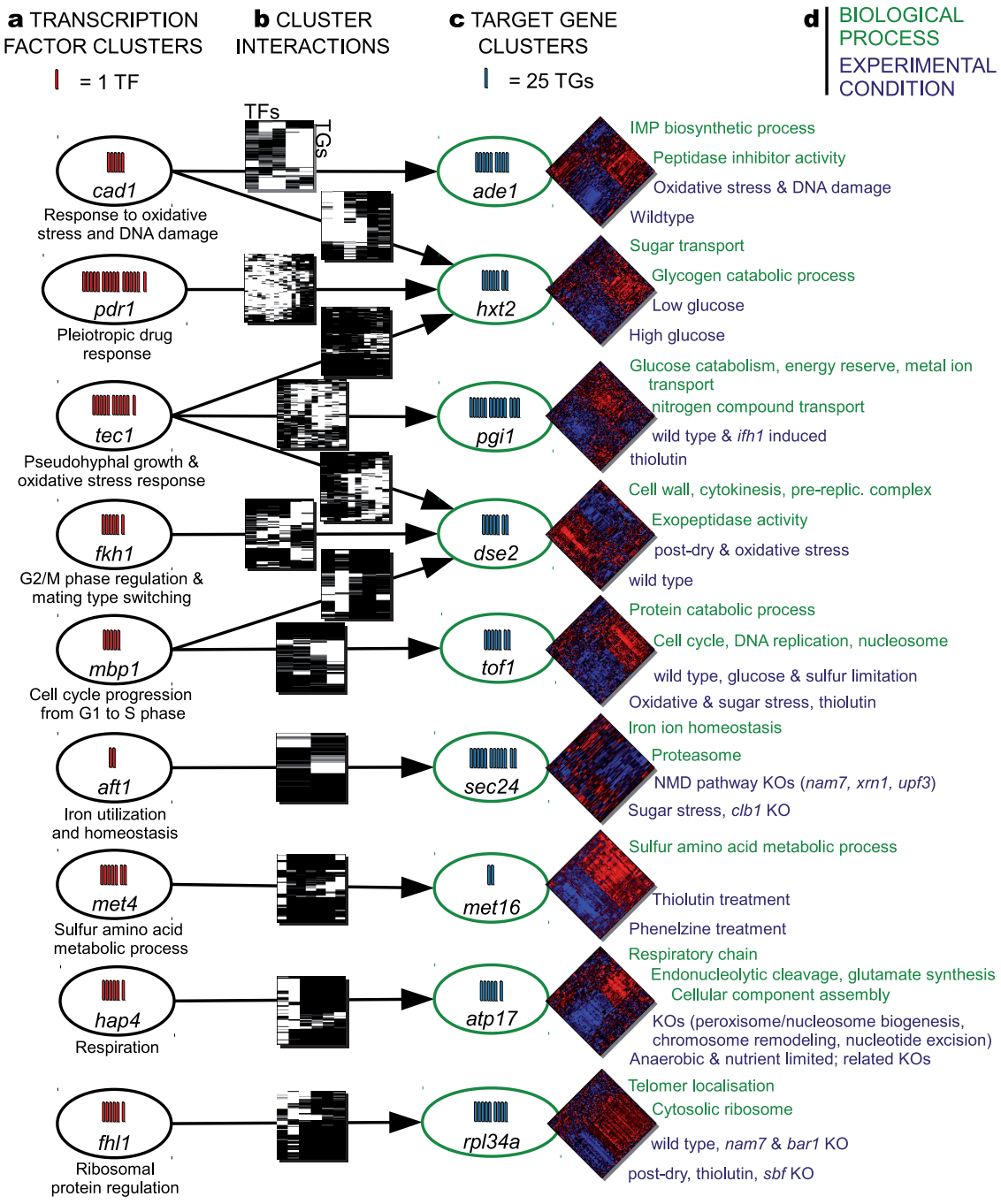

Figure 9: **Interactions and expression profiles.** We partitioned our network of 22,231 gene regulatory interactions for visualization and identification of network modules. We derived **(a)** 9 clusters of 61 TFs that, via **(b)** 13 interactions between clusters (arrows), regulate **(c)** 9 clusters of 1758 target genes. A representative gene is displayed for each TF and target cluster. Cluster interaction maps (black=interaction, white=no interaction) comprise a total of 11232 (50.5%) interactions. **(d)** Thus, depicted TF modules are likely to trigger expression responses (heatmaps: red=up-, blue=down-regulation) in respective target modules and associated biological processes (green annotation). The heatmaps display the differential expression of these target modules under the indicated knockout (KO) and other experimental conditions (blue annotation).

oxidative stress and wild type growth in YEP medium, respectively. The genes involved in IMP biosynthetic process display the reversed expression pattern. In our network, these genes are subject to transcriptional control by the *cad1* TF-cluster. As explained in the subsequent section on TF clusters, these regulators drive the response to oxidative/osmotic stress and resulting DNA-damage. This initiates repair mechanisms of the DNA and targeted protein degradation and explains the observed expression pattern of the target genes strongly reacting to oxidative stress. Among them is the cluster representative *ade1*, required for 'de novo' purine nucleotide biosynthesis (Shakoury-Elizeh *et al.*, 2004a). Other genes of the ADE family, namely *ade4,5,7,13,17* are also contained in the cluster making it the gene family with the most members present in the cluster.

Sugar transport & glycogen catabolism (*hxt2*). The *hxt2*-cluster represents sugar transporter responding to different extra-cellular sugar concentrations, and genes involved in glycogen catabolism required under glucose limitation. Matching these functions, we observed that the genes involved in glycogen catabolism are strongly up-regulated under thiolutin treatment and strongly down-regulated in media supplied with high glucose concentrations. On the other hand, we could confirm that the genes involved in sugar transport are not uniformly expressed – some are up- (e.g. *hxt1*) and some are down-regulated (e.g. *hxt2*) under high glucose concentrations (as previously described in (Ozcan and Johnston, 1999)). We found the *hxt2*-cluster among the two most heavily regulated clusters in our network (besides the *dse2*-cluster): it is regulated by the *cad1*-cluster (oxidative stress), the *pdr1*-cluster (drug response), and the *tec1*-cluster (pseudohypal growth). Thus, we linked the regulation by the *cad1*-cluster under oxidative stress with higher glucose uptake rates and/or release of glucose from the glycogen storage in order to respond to higher energy requirements under stress. The same holds for the *pdr1*-cluster to ensure the pleiotropic drug resistance of yeast cells. On the other hand, we predicted regulation by the *tec1*-cluster under pseudohyphal growth, i.e. growth under abundant fermentable glucose, so that the expression of glucose transporters is up-regulated and the expression of glycogen catabolic enzymes down-regulated. This matches the actual observed expression behavior. *hxt2*, a member of the hexose transporter HXT family (Ozcan and Johnston, 1999), is the representative of the cluster. The HXT family is also the gene group with the most members in this cluster (four additional hexose transporter genes *hxt1,3,7,9* are included as well).

Glucose catabolism, energy reserve & nitrogen compound transport (*pgi1*). The genes in this cluster have overrepresented functions in the *glucose catabolic process* (like *glycolysis*), *metal ion transport*, *energy reserve metabolic process*, and *nitrogen compound transport*. Thus, this cluster is functionally related to the *hxt2*-cluster. However, while the *hxt2*-cluster is responsible for uptake and release of glucose from the glycogen storage, the *pgi1*-cluster performs the energy production via catabolism of glucose, and feeds the energy reserve with superfluous glucose. Interestingly, these processes are also controlled by the *tec1*-cluster like the *hxt2*-cluster. Thus, we found these regulators to be tightly linked to genes involved in glucose usage, i.e. uptake and catabolism, in times of glucose starvation and abundance, where it is released from and fed to the energy reserve, respectively. Among the target genes is the cluster representative, *pgi1*, the glycolytic enzyme phosphoglucose isomerase (Aguilera and Zimmermann, 1986). In addition, the genes functioning in metal ion transport, energy reserve metabolic process, and protein refolding are significantly up-regulated under thiolutin treatment and down-regulated under wild-type and *ifh1* (activator of ribosomal protein expression) induced conditions. Among them is the gene group most strongly represented – the heat shock proteins *hsp26,31,42,78,104* annotated to have chaperone activity.

Cell wall, cell division & peptidase activity (*dse2*). Target genes in the *dse2*-cluster encode predominantly proteins with cell cycle / cell division related functions such as *fungal-type cell wall* and *cytokinesis, completion of separation*, but also *pre-replicative complex assembly* and *exopeptidase activity*. In our expression compendia, the genes with functions in cell wall organization, cytokinesis, and pre-replicative complex assembly appear strongly up-regulated under wild type growth in YEP medium and strongly down-regulated in post-dry and oxidative stress conditions. This indicates that the cell cycle is arrested under harmful conditions by down-regulation of the genes responsible for its execution under favorable conditions. Arrest of the cell cycle occurs especially when the cell fails to pass the G1-checkpoint (G1-S transition; ensures proper DNA synthesis) or the G2-checkpoint (G2-M transition; ensures proper mitosis). We exactly predicted regulators of the G1-S transition (*mbp1*-cluster) and G2-M transition (*fkh1*-cluster) to control the *dse2*-cluster. Thus, the function and expression of target genes appropriately matches the predicted regulations. While the *mbp1*-cluster predominantly controls G1-related proteins promoting growth and S-phase preparation (e.g. *pre-replicative complex assembly*), the *fkh1*-cluster especially targets genes involved

in *cytokinesis, completion of separation*. In addition, the *tec1*-cluster promoting pseudohyphal growth and filamentation introduces a third facet of the *dse2*-cluster regulation. Here, the *tec1* cluster mainly regulates genes involved in cell wall and cell membrane disassembly. Such a gene is the cluster representative *dse2*, which degrades the cell wall from the daughter cell causing the daughter to separate from the mother cell (Colman-Lerner *et al.*, 2001). Groups having more than two members within the cluster are: *cdc5*,6,20; chitin synthases involved in cytokinesis (*chs1,2,7*); B-type cyclins (*clb1,2,6*); G1 cyclins (*cln1-3*); daughter cell specific genes (*dse1,2,4*); and MCM and SWI genes.

Protein catabolism & cell cycle (*tof1*). Related to the *dse2*-cluster, the *tof1*-cluster contains genes involved in S phase related cell cycle events (overrepresented GO-terms: *heteroduplex formation*, *cell cycle*, *replication fork*, *DNA replication*, *nucleosome*). On the other hand, the second major component of this cluster are protein catabolic genes (overrepresented GO-terms: *endopeptidase regulator activity*, *proteosomal protein catabolic process* and *peptide catabolic process*). The *tof1*-cluster is similar to the *dse2*-cluster in two additional attributes besides similar functions of their member genes. We observed for both gene clusters striking expression patterns under oxidative stress as compared to wildtype growth. And both are regulated in our predicted network by the *mbp1*-cluster that regulates the cell cycle transition from G1- to S-phase. However, while the *tof1*-cluster is regulated by the *mbp1*-cluster alone, we predicted two additional regulator clusters for the *dse2*-cluster (*fkh1* and *tec1*; see previous section). This was the reason for separating the two clusters from one another. The cluster representative is *tof1*, a subunit of the DNA replication pausing checkpoint complex Tof1p-Mrc1p-Csm3p (Katou *et al.*, 2003). *mrc1* is also included in the cluster. Most strongly represented are CDC and POL genes with five (*cdc5,7,9,21,45*) and four (*pol1,12,30,32*) members encoding S phase specific cell division cycle genes and subunits of α and δ DNA polymerase, respectively.

Iron ion homeostasis & proteasome (*sec24*). Half of the genes in the *sec24*-cluster are involved in *iron ion homeostasis* and the other half are components of the *proteasome regulatory particle, base subcomplex*. While genes annotated to *iron ion homeostasis* appear predominantly up-regulated in experiments where members of the nonsense mediated mRNA decay (NMD) pathway have been knocked out (e.g. *nam7*, *xrn1*, and UPFs), the majority of genes annotated to *proteasome regulatory particle, base subcomplex* are down-regulated under these conditions. In contrast, the latter are up-regulated under sugar stress and cell cycle related knockouts (*clb1* and *cdc36,39,40*) implying higher proteolytic activity under these conditions. The cluster representative is *sec24*, which is required for cargo selection during vesicle formation in ER to Golgi transport Duden (2003). Regulated by *aft1* and *reb1*, most strongly represented in this cluster are the SEC and FRE gene families: there are five additional SEC genes (*sec4,8,9,23,61*) and five FRE genes (*fre1-3,5-6* encoding ferric reductase) contained in the cluster.

Sulfur amino acid metabolic process (*met16*). The genes of the *met16*-cluster encode enzymes of the *sulfur amino acid metabolic process*. In our expression compendia, these genes are significantly up- and down-regulated under thiolutin and phenelzine treatment, respectively. Thiolutin is a sulfur-containing antibiotic, which is a potent inhibitor of yeast RNA polymerases. While thiolutin blocks in general transcription in yeast, Pelechano and JE (2008) observed that genes functioning in the sulphur amino acid metabolic process are induced in response to the weak stress induced by low concentrations of thiolutin (see also Grigull *et al.* (2004)). Phenelzine, on the other hand, is a non-selective and irreversible monoamine oxidase inhibitor, which has been linked to decreased concentrations of sulfur amino acids in rat brains (Benedetti *et al.*, 1990, 1991). As judged from gene expression in yeast, we argued that this seems to be due to down-regulation of the responsible MET genes. The regulation in both cases is carried out through combined control by the *met4,28,31,32* and *cbf1* transcription factors of the *met4*-cluster as explained below. The cluster representative is *met16*, a 3'-phosphoadenylsulfate reductase involved in sulfate assimilation and methionine metabolism (Thomas *et al.*, 1990). In addition, 19 other genes of the MET gene family are also included in the cluster.

Respiratory chain (*atp17*). The majority of genes in this cluster perform functions in the *mitochondrial respiratory chain* such as *hydrogen ion transmembrane transporter activity*. Experimental conditions associated with differential expression of these genes are oxygen availability (aerobis: up; anaerobis: down) and corresponding transcription factor perturbations (e.g. *hap4* over-expression). Our predicted regulatory interactions place the *atp17*-cluster under control of *hap1* and the *hap2/3/4/5* complex, which agrees well with existing knowledge (Zitomer and Lowry, 1992). *atp17* itself encodes the subunit f of mitochondrial ATP synthase (Spannagel *et al.*, 1997). Other subunits of the ATP synthase, a key enzyme of the respiratory chain, are also contained in the cluster (*atp1-5,7,14-18,20*). In addition, the cluster contains eight subunits

of cytochrome c oxidase (*cox4-9,12,13*), and all seven subunits of the ubiquinol cytochrome-c reductase complex (*qcr1,2,6-10*).

Cytosolic ribosome (*rpl34a*). The *rpl34* A-cluster contains genes coding for ribosomal proteins. These genes are significantly up-regulated over a wide range of favorable growth conditions, including wild type growth and knockout of *nam7* (an ATP-dependent RNA helicase involved in nonsense mediated mRNA decay) and *bar1* (an aspartyl protease helping yeast cells to find mating partners). On the other hand, we observed a significant down-regulation under harmful conditions where growth is slowed or stopped, e.g. under post-dry and thiolutin treatment conditions and also in experiments where the components of the SBF complex (*swi4* and *swi6*) have been knocked out. In our network, we predicted the majority of RPL genes to be regulated in concert by *fhl1*, *ifh1*, *sfp1*, and *rap1*. As explained in the section on TF clusters below that matches their known functionality. Among the ribosomal genes is also the cluster representative *rpl34* A, which encodes a component of the large 60S ribosomal subunit (Planta and Mager, 1998). In total, the cluster contains 64 RPL-genes, coding for protein components of the large 60S ribosomal subunit, and 45 RPS-genes, coding for protein components of the small 40S ribosomal subunit.

6.3 Transcription factor clusters

Response to oxidative stress & DNA damage (*cad1*). This cluster contains major regulators of the response to oxidative stress and resulting DNA damage. Among them are five of eight members of the Yap family, namely *yap1* and *cad1* (*yap2*), major oxidative stress regulators; *cin5* (*yap4*) and *yap6*, involved in osmotic stress response; and *yap7*, of currently unknown function (Rodrigues-Pousada *et al.*, 2010). In addition, the cluster contains *rfx1*, a major transcriptional repressor of DNA-damage-regulated genes (Zaim *et al.*, 2005), and *xbp1*, a transcriptional repressor that binds to promoter sequences of the cyclin genes and that is induced by stress or starvation during mitosis (Mai and Breeden, 1997).

Pleiotropic drug response (*pdr1*). This TF cluster consists of four major functional subgroups, all tightly connected to the cellular response to various kinds of drug and nutrition stress.
The first group contains *pdr1*, *pdr3*, *stb5* and *msn1*. While *msn1* is a general regulator of drug response (Chang *et al.*, 2003), *pdr1*, *pdr3* and *stb5* (known to build homodimers, and heterodimers with each other) are very specific regulators of the regulation of the multi-drug resistance in yeast (Akache *et al.*, 2004).
The second group contains *nrg1*, *mga1* and *ash1*, known to regulate pseudohyphal growth, i.e. growth in conditions of nitrogen limitation and abundant fermentable carbon sources like glucose (Arkowitz and Bassilana, 2011). *nrg1* is a negative regulator of glucose-repressed genes (Zhou and Winston, 2001), *mga1* a suppressor of pseudohyphal growth defects (Lorenz and Heitman, 1998), and *ash1* a positive regulator of pseudohyphal growth linked to mating switching and cell cycle (Cosma, 2004).
The third functional component of this TF cluster consists of *sut1*, *rox1*, *ixr1*, and *rim101*. All four TFs are involved in gene expression under hypoxia, i.e. deprivation of sufficient oxygen levels. *sut1* induces (Regnacq *et al.*, 2001), whereas *rox1* represses hypoxic gene expression (Kastaniotis and Zitomer, 2000). *ixr1* is also annotated to play a role in the cellular response to hypoxia (Lambert *et al.*, 1994), whereas *rim101* generally contributes to the response to anoxic, anaerobic, and pH stress (Lamb *et al.*, 2003).
The fourth subgroup is only loosely connected, as all members are involved in different nutrition stress responses: *ino4* derepresses inositol/choline-regulated genes (Santiago and Mamoun, 2003), *pho4* activates transcription of phosphate metabolism in response to phosphate limitation (Zhou and EK, 2011), *smp1* contributes to the response to osmotic stress (de Nadal *et al.*, 2003), *cup9* represses peptide transport (Byrd *et al.*, 1998), and *rlm1* responds to stress in order to maintain cell integrity (Jung *et al.*, 2002).

Pseudohyphal growth & oxidative stress response (*tec1*). This TF cluster consists of three major subgroups, which are functionally interconnected with each other. The first group contains *tec1*, *ste12*, *sok2*, *phd1*, and *flo8*, which are all regulators of pseudohyphal and invasive growth (Liu *et al.*, 1996; Pan and Heitman, 2000; Chou *et al.*, 2006; Brueckner *et al.*, 2011). The second group contains *skn7*, *sko1*, and *msn2*, regulators of the response to osmotic and oxidative stress (Martinez-Pastor *et al.*, 1996; Rep *et al.*, 2001; He *et al.*, 2009). And the third group contains the general repressor-activator protein *rap1*, and the regulator of ribosomal protein transcription *fhl1*. The three groups work tightly together in the response to osmotic and oxidative stress, which is known to induce pseudohypal and invasive growth in yeast (Zaragoza and Gancedo, 2000). In response to stress and growth conditions, ribosomal protein gene transcription is then regulated by *fhl1* and *rap1* (Zhao *et al.*, 2006); in its role as a transcription activating factor, the largest group of *rap1*

target genes are those that encode ribosomal proteins(Lieb *et al.*, 2001).

G2/M regulation & mating type switching (*fkh1*). This TF cluster contains six TFs – *fkh1*, *fkh2*, *mcm1*, *ndd1*, *swi5*, and *ace1* – that are all involved in the regulation of specific cell cycle phases and the mating type switching (initiated in G1). The forkhead transcription factors *fkh1* and *fkh2* regulate the expression of G2/M phase genes and donor preference during mating type switching (Zhu *et al.*, 2000; Coïc *et al.*, 2006). *mcm1* is a pleiotropic regulator of cell-type-specific transcription and pheromone response. *mcm1* also activates transcription of genes involved in G2/M-phase of mitosis and regulates mating type switching in cooperation with *fkh1* and *fkh2* (Kumar *et al.*, 2000). *ndd1* positively regulates G2/M-phase genes and is recruited by the forkhead TFs and *mcm1* to G2/M-specific promoters (Koranda *et al.*, 2000). Among the genes regulated by *fkh1* and *fkh2* are also *swi5* and *ace1*, which are transcription factors required for the subsequent temporal wave of cell cycle regulated gene expression in the M/G1 phase interval Spellman *et al.* (1998).

Cell cycle progression from G1 to S (*mbp1*). This TF cluster contains four TFs: *mbp1*, *swi4*, *swi6*, and *mal33*. *mbp1*, *swi4* and *swi6* are known to form regulatory complexes that drive the expression of genes of the G1/S transition, including cyclins and genes required for DNA synthesis and repair (Bean *et al.*, 2005). *mal33*, regulator of maltose fermentation, is a further downstream transcription factor, regulated itself by *mbp1* (Iyer *et al.*, 2001).

Iron utilization & homeostasis (*aft1*). This is the smallest TF cluster consisting of only two TFs: *aft1* and *reb1*. *aft1* is involved in iron utilization and homeostasis (Shakoury-Elizeh *et al.*, 2004b), whereas *reb1* is the general RNA polymerase I enhancer binding protein (Bordi *et al.*, 2001). The clustering indicates how *aft1* recruits (or is recruited to) the RNA polymerase, presumably via physical interaction with *reb1*.

Sulfur amino acid metabolic process (*met4*). This cluster contains *met4*, the major activator of the sulfur amino acid metabolic process, along with stabilizing (*met28*) and DNA-binding cofactors (*met31*, *met32*, and *cbf1*), which are known to work in a regulatory complex (Lee *et al.*, 2010). The cluster also contains *gcn4*, which facilitates general amino acid control and, thus, also regulates *met4* (Mountain *et al.*, 1993). A TF involved in branched-chain amino acid synthesis, namely *leu3*, is also contained (Friden and Schimmel, 1988).

Respiration (*hap4*). This cluster represents the TFs that control respiration in dependence of heme via *hap1*, or independent from heme via *hap4*, in complex with *hap2*, *hap3*, and *hap5* (Zitomer and Lowry, 1992). The cluster also contains *gln3*, a transcriptional activator of genes regulated by nitrogen catabolite repression (Magasanik and Kaiser, 2002). The link between *gln3* and the HAPs is indirect and is established by the retrograde regulators *rtg1* and *rtg3*. These TFs are known to regulate specific HAP target genes in response to mitochondrial dysfunction and particular nitrogen sources discriminated via the DAL/NAP genes under *gln3* control (Zaman *et al.*, 2008).

Ribosomal protein regulation (*fhl1*). This cluster contains the major regulators of ribosomal protein transcription *fhl1* and *sfp1*, and the coactivator of *fhl1*, namely *ifh1* (Marion *et al.*, 2004; Zhao *et al.*, 2006). Interestingly, *fhl1* and *sfp1* are known to bind via the general regulatory proteins *rap1* and *ste12*, which are also contained in the cluster.

References

Abdulrehman, D. *et al.* (2011). YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*, **39**(Database issue), D136–D140.

Aguilera, A. and Zimmermann, F. (1986). Isolation and Molecular Analysis of the Phosphoglucose Isomerase Structural Gene of Saccharomyces Cerevisiae. *Mol Gen Genet*, **202**(1), 83–9.

Akache, B. *et al.* (2004). Complex Interplay Among Regulators of Drug Resistance Genes in Saccharomyces Cerevisiae. *J Biol Chem*, **279**(27), 7855–60. 2.

Ambroise, J. *et al.* (2012). Transcriptional network inference from functional similarity and expression data: a global supervised approach. *Stat Appl Genet Mol Biol*, **11**(1), 1–24.

Arkowitz, R. and Bassilana, M. (2011). Polarized Growth in Fungi: Symmetry Breaking and Hyphal Formation. *Semin Cell Dev Biol*, **22**(8), 806–15.

Barla, A. *et al.* (2008). A method for robust variable selection with significance assessment. In *ESANN*, pages 83–88.

Barzel, B. *et al.* (2013). Network link prediction by global silencing of indirect correlations. *Nat Biotechnol*, **31**(8), 720–725.

Bauer, T. *et al.* (2011). RIP: the regulatory interaction predictor–a machine learning-based approach for predicting target genes of transcription factors. *Bioinformatics*, **27**(16), 2239–2247.

Bean, J. *et al.* (2005). High Functional Overlap Between MluI Cell-cycle Box Binding Factor and Swi4/6 Cell-cycle Box Binding Factor in the G1/S Transcriptional Program in Saccharomyces Cerevisiae. *Genetics*, **171**(1), 49–61.

Benedetti, M. *et al.* (1990). The Effects of Aging on MAO Activity and Amino Acid Levels in Rat Brain. *J Neural Transm Suppl*, **29**, 259–68.

Benedetti, M. *et al.* (1991). Effects of Ageing on the Content in Sulfur-containing Amino Acids in Rat Brain. *J Neural Transm Gen Sect*, **86**(3), 191–203.

Bleakley, K. *et al.* (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics*, **23**(13), i57–i65.

Bleakley, K. *et al.* (2009). Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, **25**(18), 2397–2403.

Bordi, L. *et al.* (2001). In Vivo Binding and Hierarchy of Assembly of the Yeast RNA Polymerase I Transcription Factors. *Mol Biol Cell*, **12**(3), 753–60.

Brons, J. F. *et al.* (2002). Dissection of the promoter of the HAP4 gene in S. cerevisiae unveils a complex regulatory framework of transcriptional regulation. *Yeast*, **19**(11), 923–932.

Brown, M. P. S. *et al.* (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, **97**(1), 262–267.

Brueckner, S. *et al.* (2011). The TEA Transcription Factor Tec1 Links TOR and MAPK Pathways to Coordinate Yeast Development. *Genetics*, **189**(2), 479–94.

Bryne, J. C. *et al.* (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, **36**(Database issue), D102–D106.

Byrd, C. *et al.* (1998). The N-end Rule Pathway Controls the Import of Peptides Through Degradation of a Transcriptional Repressor. *EMBO J*, **17**(1), 269–77.

Cai, Y.-D. *et al.* (2007). Prediction of regulatory networks: identification of transcriptionfactor-target relationship from gene ontology information and gene expression data. Technical report, Manchester Institute for Mathematical Sciences, The University of Manchester. MIMS Preprint.

Cerulo, L. *et al.* (2010). Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics*, **11**, 228.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chang, K. *et al.* (2003). The Putative Transcriptional Activator MSN1 Promotes Chromium Accumulation in Saccharomyces Cerevisiae. *Mol Cells*, **16**(3), 291–6.

Chou, S. *et al.* (2006). Regulation of Mating and Filamentation Genes by Two Distinct Ste12 Complexes in Saccharomyces Cerevisiae. *Mol Cell Biol*, **26**(13), 4794–805.

Chua, G. *et al.* (2006). Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci USA*, **103**(32), 12045–12050.

Ciofani, M. *et al.* (2012). A validated regulatory network for th17 cell specification. *Cell*, **151**(2), 289–303.

Coïc, E. *et al.* (2006). Cell cycle-dependent regulation of saccharomyces cerevisiae donor preference during mating-type switching by sbf (swi4/swi6) and fkh1. *Mol Cell Biol*, **26**(14), 5470–5480.

Colman-Lerner, A. *et al.* (2001). Yeast Cbk1 and Mob2 Activate Daughter-specific Genetic Programs to Induce Asymmetric Cell Fates. *Cell*, **107**(6), 739–50.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Cosma, M. (2004). Daughter-specific Repression of Saccharomyces Cerevisiae HO: Ash1 Is the Commander. *EMBO Rep*, **5**(10), 953–7.

de la Fuente, A. *et al.* (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**(18), 3565–3574.

de Nadal, E. *et al.* (2003). Targeting the MEF2-like Transcription Factor Smp1 by the Stress-activated Hog1 Mitogen-activated Protein Kinase. *Mol Cell Biol*, **23**(1), 229–37.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat Rev Microbiol*, **8**(10), 717–729.

Deng, M. *et al.* (2004). An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*, **11**(2-3), 463–475.

Dimitriadou, E. *et al.* (2011). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6.

Duden, R. (2003). ER-to-Golgi Transport: COP I And COP II Function (Review). *Mol Membr Biol*, **20**(3), 197–207.

Eisen, M. B. *et al.* (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, **95**(25), 14863–14868.

Ernst, J. *et al.* (2008). A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comput Biol*, **4**(3), e1000044.

Faith, J. J. *et al.* (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, **5**(1), e8.

Faith, J. J. *et al.* (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res*, **36**(Database issue), D866–D870.

Feizi, S. *et al.* (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat Biotechnol*, **31**(8), 726–733.

Friden, P. and Schimmel, P. (1988). LEU3 of Saccharomyces Cerevisiae Activates Multiple Genes for Branched-chain Amino Acid Biosynthesis by Binding to a Common Decanucleotide Core Sequence. *Mol Cell Biol*, **8**(7), 2690–7.

Friedman, J. H. *et al.* (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1–22.

Gasch, A. P. *et al.* (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**(12), 4241–4257.

Geurts, P. (2011). Learning from positive and unlabeled examples by enforcing statistical significance. In *JMLR: Workshop and Conference Proceedings*, volume 15.

Gillis, J. and Pavlidis, P. (2011). The impact of multifunctional genes on "guilt by association" analysis. *PLoS One*, **6**(2), e17258.

Greenfield, A. *et al.* (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, **29**(8), 1060–1067.

Grigull, J. *et al.* (2004). Genome-wide Analysis of MRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *Mol Cell Biol*, **4**(12), 5534–47.

Gurvitz, A. and Rottensteiner, H. (2006). The biochemistry of oleate induction: transcriptional upregulation and peroxisome proliferation. *Biochim Biophys Acta*, **1763**(12), 1392–1402.

Gustafsson, M. and Hörnquist, M. (2010). Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge. *PLoS One*, **5**(2), e9134.

Harbison, C. T. *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**(7004), 99–104.

Hastie, T. *et al.* (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag.

Haynes, B. C. *et al.* (2013). Mapping functional transcription factor networks from gene expression data. *Genome Res*, **23**(8), 1319–1328.

He, X. *et al.* (2009). Oxidative Stress Function of the Saccharomyces Cerevisiae Skn7 Receiver Domain. *Eukaryot Cell*, **8**(5), 768–78.

Hill, A. V. (1910). Proceedings of the physiological society: January 22, 1910. *The Journal of Physiology*, **40**(Suppl), i–vii.

Hiltunen, J. K. *et al.* (2003). The biochemistry of peroxisomal beta-oxidation in the yeast Saccharomyces cerevisiae. *FEMS Microbiol Rev*, **27**(1), 35–64.

Holloway, D. T. *et al.* (2007). Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Syst Synth Biol*, **1**(1), 25–46.

Holloway, D. T. *et al.* (2008). Classifying transcription factor targets and discovering relevant biological features. *Biol Direct*, **3**, 22.

Hu, Z. *et al.* (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet*, **39**(5), 683–687.

Huh, W.-K. *et al.* (2003). Global analysis of protein localization in budding yeast. *Nature*, **425**(6959), 686–691.

Huynh-Thu, V. A. *et al.* (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**(9), e12776.

Ito, T. *et al.* (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, **98**(8), 4569–4574.

Iyer, V. *et al.* (2001). Genomic Binding Sites of the Yeast Cell-cycle Transcription Factors SBF and MBF. *Nature*, **409**(6819), 533–8.

Jung, U. *et al.* (2002). Regulation of the Yeast Rlm1 Transcription Factor by the Mpk1 Cell Wall Integrity MAP Kinase. *Mol Microbiol*, **46**(3), 781–9.

Kanehisa, M. *et al.* (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**(Database issue), D277–D280.

Kastaniotis, A. and Zitomer, R. (2000). Rox1 Mediated Repression. Oxygen dependent repression in yeast. . *Adv Exp Med Biol*, **475**, 185–95.

Katou, Y. *et al.* (2003). S-phase Checkpoint Proteins Tof1 and Mrc1 Form a Stable Replication-pausing Complex. *Nature*, **424**(6952), 1078–83.

Koranda, M. *et al.* (2000). Forkhead-like Transcription Factors Recruit Ndd1 to the Chromatin of G2/M-specific Promoters. *Nature*, **406**(6791), 94–8.

Küffner, R. *et al.* (2012). Inferring gene regulatory networks by ANOVA. *Bioinformatics*, **28**(10), 1376–1382.

Kumar, R. *et al.* (2000). Forkhead Transcription Factors, Fkh1p and Fkh2p, Collaborate with Mcm1p to Control Transcription Required for M-phase. *Curr Biol*, **10**(15), 896–906.

Lamb, T. *et al.* (2003). The Transcription Factor Rim101p Governs Ion Tolerance and Cell Differentiation by Direct Repression of the Regulatory Genes NRG1 and SMP1 in Saccharomyces Cerevisiae. *Mol Cell Biol*, **23**(2), 677–86.

Lambert, J. *et al.* (1994). The ORD1 Gene Encodes a Transcription Factor Involved in Oxygen Regulation and Is Identical to IXR1, a Gene That Confers Cisplatin Sensitivity to Saccharomyces Cerevisiae. *Proc Natl Acad Sci USA*, **91**(15), 7345–9.

Lanckriet, G. R. G. *et al.* (2004). Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputings*, volume 9, pages 300–311. World Scientific Singapore.

Lee, T. *et al.* (2010). Dissection of Combinatorial Control by the Met4 Transcriptional Complex. *Mol Biol Cell*, **21**(3), 456–69.

Lee, T. I. *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**(5594), 799–804.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, **2**(3), 18–22.

Lieb, J. *et al.* (2001). Promoter-specific Binding of Rap1 Revealed by Genome-wide Maps of Protein-DNA Association. *Nat Genet*, **28**(4), 327–34.

Liu, H. *et al.* (1996). Saccharomyces Cerevisiae S288C Has a Mutation in FLO8, a Gene Required for Filamentous Growth. *Genetics*, **144**(3), 967–78.

Lorenz, M. C. and Heitman, J. (1998). Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics*, **150**(4), 1443–1457.

Luts, J. *et al.* (2010). A tutorial on support vector machine-based methods for classification problems in chemometrics. *Anal Chim Acta*, **665**(2), 129–145.

Madhamshettiwar, P. B. *et al.* (2012). Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med*, **4**(5), 41.

Magasanik, B. and Kaiser, C. (2002). Nitrogen Regulation in Saccharomyces Cerevisiae. *Gene*, **290(1-2)**, 1–18.

Mai, B. and Breeden, L. (1997). Xbp1, a Stress-induced Transcriptional Repressor of the Saccharomyces Cerevisiae Swi4/Mbp1 Family. *Mol Cell Biol*, **17**(11), 6491–501.

Marbach, D. *et al.* (2012). Wisdom of crowds for robust gene network inference. *Nat Methods*, **9**(8), 796–804.

Margolin, A. A. *et al.* (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.

Marion, R. *et al.* (2004). Sfp1 Is a Stress- and Nutrient-sensitive Regulator of Ribosomal Protein Gene Expression. *Proc Natl Acad Sci USA*, **101**(40), 4315–22. 1.

Martinez-Pastor, M. *et al.* (1996). The Saccharomyces Cerevisiae Zinc Finger Proteins Msn2p and Msn4p Are Required for Transcriptional Induction Through the Stress Response Element (STRE). *EMBO J*, **5**(9), 2227–35.

Matys, V. *et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, **34**(Database issue), D108–D110.

Mordelet, F. and Vert, J.-P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24**(16), i76–i82.

Mordelet, F. and Vert, J.-P. (2010). A bagging SVM to learn from positive and unlabeled examples. Technical report, Cornell University Library.

Mostafavi, S. *et al.* (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*, **9 Suppl 1**, S4.

Mountain, H. *et al.* (1993). The General Amino Acid Control Regulates MET4, Which Encodes a Methionine-pathway-specific Transcriptional Activator of Saccharomyces Cerevisiae. *Mol Microbiol*, **7**(2), 215–28.

Myers, C. L. *et al.* (2006). Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

Ozcan, S. and Johnston, M. (1999). Function and Regulation of Yeast Hexose Transporters. *Microbiol Mol Biol Rev*, **63**(3), 554–69.

Özgür, A. *et al.* (2005). Text categorization with class-based and corpus-based keyword selection. In *Proceedings of the 20th international conference on Computer and Information Sciences*, ISCIS'05, pages 606–615, Berlin, Heidelberg. Springer-Verlag.

Pan, X. and Heitman, J. (2000). Sok2 Regulates Yeast Pseudohyphal Differentiation Via a Transcription Factor Cascade That Regulates Cell-cell Adhesion. *Mol Cell Biol*, **20**(22), 8364–72.

Pavlidis, P. and Gillis, J. (2013). Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update. *F1000Res*, **2**, 230.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.

Pelechano, V. and JE, P.-O. (2008). The Transcriptional Inhibitor Thiolutin Blocks MRNA Degradation in Yeast. *Yeast*, **25**(2), 85–92.

Planta, R. and Mager, W. (1998). The List of Cytoplasmic Ribosomal Proteins of Saccharomyces Cerevisiae. *Yeast*, **14**(5), 471–7.

Qian, J. *et al.* (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**(15), 1917–1926.

Quinlan, J. R. (1996). Bagging, Boosting, and C4.5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730. AAAI Press.

Rahmann, S. *et al.* (2003). On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, **2**, Article7.

Regnacq, M. *et al.* (2001). SUT1p Interaction with Cyc8p(Ssn6p) Relieves Hypoxic Genes from Cyc8p-Tup1p Repression in Saccharomyces Cerevisiae. *Mol Microbiol*, **40**(5), 1085–96.

Reményi, A. *et al.* (2004). Combinatorial control of gene expression. *Nat Struct Mol Biol*, **11**(9), 812–815.

Rep, M. *et al.* (2001). The Saccharomyces Cerevisiae Sko1p Transcription Factor Mediates HOG Pathway-dependent Osmotic Regulation of a Set of Genes Encoding Enzymes Implicated in Protection from Oxidative Damage. *Mol Microbiol*, **40**(5), 1067–83.

Rodrigues-Pousada, C. *et al.* (2010). The Yap Family and Its Role in Stress Response. *Yeast*, **27**(5), 245–58.

Santiago, T. and Mamoun, C. (2003). Genome Expression Analysis in Yeast Reveals Novel Transcriptional Regulation by Inositol and Choline and New Regulatory Functions for Opi1p, Ino2p, and Ino4p. *J Biol Chem*, **278**(40), 8723–30. 3.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press.

Schüller, H.-J. (2003). Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr Genet*, **43**(3), 139–160.

Sebastiani, F. (2005). Text categorization. In *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press.

Seok, J. *et al.* (2010). Knowledge-based analysis of microarrays for the discovery of transcriptional regulation relationships. *BMC Bioinformatics*, **11 Suppl 1**, S8.

Shakoury-Elizeh, M. *et al.* (2004a). DNA-bound Bas1 recruits Pho2 to activate ADE genes in Saccharomyces cerevisiae. *Eukaryot Cell*, **4**(10), 1725–35.

Shakoury-Elizeh, M. *et al.* (2004b). Transcriptional Remodeling in Response to Iron Deprivation in Saccharomyces Cerevisiae. *Mol Biol Cell*, **15**(3), 1233–43.

Shimamura, T. *et al.* (2009). Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Syst Biol*, **3**, 41.

Simon, N. *et al.* (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, **39**(5), 1–13.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **13**(2), 238–241.

Smith, J. J. *et al.* (2007). Transcriptional responses to fatty acid are coordinated by combinatorial control. *Mol Syst Biol*, **3**, 115.

Spannagel, C. *et al.* (1997). The Subunit F of Mitochondrial Yeast ATP Synthase – Characterization of the Protein and Disruption of the Structural Gene ATP17. *Eur J Biochem*, **247**(3), 1111–7.

Spellman, P. T. *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, **9**(12), 3273–3297.

The Gene Ontology Consortium (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, **38**(Database issue), D331–D335.

Thomas, D. *et al.* (1990). Gene-enzyme Relationship in the Sulfate Assimilation Pathway of Saccharomyces Cerevisiae. Study of the 3'-phosphoadenylylsulfate reductase structural gene. *J Biol Chem*, **265**(26), 5518–24.

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Turcotte, B. *et al.* (2010). Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS Yeast Res*, **10**(1), 2–13.

Uetz, P. *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Vert, J.-P. (2010). *Reconstruction of Biological Networks by Supervised Machine Learning Approaches*, chapter 7, pages 163–188. John Wiley & Sons, Inc.

von Mering, C. *et al.* (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Wingender, E. *et al.* (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, **29**(1), 281–283.

Winston, P. H. (1992). *Artificial Intelligence, 3rd edition*. Addison Wesley.

Yamanishi, Y. *et al.* (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, **21 Suppl 1**, i468–i477.

Yip, K. Y. *et al.* (2009). Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC Bioinformatics*, **10**, 241.

Yip, K. Y. and Gerstein, M. (2009). Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics*, **25**(2), 243–250.

Zaim, J. *et al.* (2005). Identification of New Genes Regulated by the Crt1 Transcription Factor, an Effector of the DNA Damage Checkpoint Pathway in Saccharomyces Cerevisiae. *J Biol Chem*, **280**(1), 28–37.

Zaman, S. *et al.* (2008). How Saccharomyces Responds to Nutrients. *Annu Rev Genet*, **42**, 27–81.

Zaragoza, O. and Gancedo, J. (2000). Pseudohyphal Growth Is Induced in Saccharomyces Cerevisiae by a Combination of Stress and CAMP Signalling. *Antonie van Leeuwenhoek*, **78**(2), 187–94.

Zhao, Y. *et al.* (2006). Fine-structure Analysis of Ribosomal Protein Gene Transcription. *Mol Cell Biol*, **26**(13), 4853–62.

Zhou, H. and Winston, F. (2001). NRG1 Is Required for Glucose Repression of the SUC2 and GAL Genes of Saccharomyces Cerevisiae. *BMC Genet*, **2**, 5. 5.

Zhou, X. and EK, O. (2011). Integrated Approaches Reveal Determinants of Genome-wide Binding and Function of the Transcription Factor Pho4. *Mol Cell*, **42**(6), 826–36.

Zhu, G. *et al.* (2000). Two Yeast Forkhead Genes Regulate the Cell Cycle and Pseudohyphal Growth. *Nature*, **406**(6791), 90–4.

Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**(7-8), 607–611.

Zitomer, R. and Lowry, C. (1992). Regulation of Gene Expression by Oxygen in Saccharomyces Cerevisiae. *Microbiol Rev*, **56**(1), 1–11.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.