Abstract

There has been much interest in reconstructing bi-directional regulatory networks linking the circadian clock to metabolism in plants. A variety of reverse engineering methods from machine learning and computational statistics have been proposed and evaluated. The emphasis of the present paper is on combining models in a model ensemble to boost the network reconstruction accuracy, and to explore various model combination strategies to maximize the improvement. Our results demonstrate that a rich ensemble of predictors outperforms the best individual model, even if the ensemble includes poor predictors with inferior individual reconstruction accuracy. For our application to metabolomic and transcriptomic time series from various mutagenesis plants grown in different light-dark cycles we also show how to determine the optimal time lag between interactions, and we identify significant interactions with a randomization test. Our study predicts new statistically significant interactions between circadian clock genes and metabolites in *Arabidopsis thaliana*, and thus provides independent statistical evidence that the regulation of metabolism by the circadian clock is not uni-directional, but that there is a statistically significant feedback mechanism aiming from metabolism back to the circadian clock.

1 Introduction

The global challenges of guaranteeing food and energy security in an expanding human population have led to an increased interest in understanding the molecular processes underlying biomass production in plants, with potential long-term applications aiming to improve the yield of crops and the quality of biofuels. A critical component of biomass production in plants is the interaction between circadian regulation and metabolism. The process of photosynthesis allows plants to utilize sunlight to produce essential carbohydrates during the day. However, the earth's rotation predictably removes sunlight, and hence the opportunity for photosynthesis, for a significant part of each day, and plants need to orchestrate the accumulation, utilization and storage of photosynthetic products in the form of starch over the daily cycle to avoid periods of starvation, and thus optimize growth rates. Plants therefore have evolved biological clocks – an endogenous circadian timing system that controls daily rhythms in transcriptional regulation and its control of metabolism – to adapt better to the 24 hour period of the solar day. In the last few years, substantial progress has been made to unravel the central processes of circadian regulation at the molecular level (Pokhilko et al., 2010, 2012, Guerriero et al., 2012). However, the detailed feedback mechanism between carbon metabolism and the circadian clock is less understood. Plants adjust the rates of starch accumulation and degradation in response to changes in the light-dark cycle with clues from the circadian clock, e.g. the starch degradation is controlled by the clock (Graf et al., 2010). Reversibly, the periodic, endogenous signals from the carbon metabolism seem to entrain the clock with up to half of the clock genes being affected (Dalchau et al., 2011, Bläsing et al., 2005). This suggests that a feedback mechanism exists between circadian clock and carbon metabolism and that metabolism plays a crucial role in regulating the clock (Haydon et al., 2013). Some basic models (Feugier and Satake, 2012) attempt to represent the dependence of starch turn-over on the circadian clock, but only capture generic high-level principles. Hence, a challenge for the plant systems biology community is the further elucidation of the

detailed structure of the bi-directional circadian regulatory networks and signalling pathways by systematic integration of transcriptomic, proteomic and metabolic concentration profiles.

2 Approach

The inference of molecular regulatory networks from post-genomic data has been a cornerstone of computational systems biology for over a decade. Following up on the seminal paper by Friedman et al. (2000), a plethora of approaches have been attempted. We have recently assessed the performance of a representative collection of state-of-the methods from machine learning and computational statistics for regulatory network reconstruction in the specific context of circadian regulation (Aderhold et al., 2014). This study has been based on a stochastic process model of molecular regulatory processes for generating realistic benchmark data from a known gold standard network (Guerriero et al., 2012), and the application of an ANOVA (analysis of variance) scheme for distinguishing the effect of the model from various confounding factors (network structure, missing values, and the effect of numerical differentiation). The present paper extends this study in four important respects. Firstly, following up on pioneering work in which we were involved (Marbach et al., 2012), we shift our focus from comparative model assessment to model ensembles, and we explore how to optimally combine predictors with unknown performance so as to maximize the overall network reconstruction accuracy. Secondly, we expand our application from the circadian clock to bi-directional regulation between circadian clock genes and metabolism, by ultilizing recent gene expression and metabolite concentration time series from various mutagenesis plants of *Arabidopsis thaliana* grown under different artificial light-dark cycles. Thirdly, we allow for the fact that transcriptional regulation and interactions between the transcriptome and the metabolome are subject to time delays, due to a series of intermediate steps related e.g. to post-translational modification, protein complex formation, translocation etc., and we explicitly include time delays into our modelling framework. Finally, we decide on the appropriate confidence threshold for network representation with a randomization test.

The shift to model ensembles was motivated by Marbach et al. (2012), where the superiority of an ensemble of predictors over the best individual predictor was first demonstrated. For the ensemble prediction, the authors chose the Borda count election method. This method was originally developed by 18th-century political scientist Jean-Charles de Borda as a method to select candidates in a democratic election. In this method, voters rank candidates in order of preference, and candidates are then ranked based on their average rank. Similarly, each prediction method provides a ranked list of regulator-regulatee interactions (corresponding to the list provided by a voter). From these lists, a list of average ranks can be computed. The authors additionally tried various weighting schemes based on the true gold standard, but pointed out that these weighting schemes are not viable in practice. In the present article, we extend the work of Marbach et al. (2012) in three respects. Firstly, our focus is on learning smaller networks at a higher level of accuracy. The networks learned by Marbach et al. (2012) contain several 100 nodes, and the high computational costs of learning such networks rule out the application of more advanced machine learning methods. Our networks are motivated by the circadian clock and its interaction with key metabolites, and contain between one and two dozen nodes. This allows

the application of more advanced Bayesian methods with MCMC sampling schemes, which have not been included in an ensemble prediction before. Secondly, we explore a variety of alternative methods for ensemble generation. Like Marbach et al. (2012), we study the standard Borda count election method, add two variants not studied before, and compare them to three popular algebraic combiners: the mean-rule, median-rule and trim (Polikar, 2006). We then explore various alternative schemes that are based on a distance measure in model output space. As opposed to the alternative weighting schemes explored by Marbach et al. (2012), which are based on the true gold standard and were merely included for deeper methodological insight, our measures draw on information that is actually available in the experiment and are therefore viable in practice. Thirdly, we apply the ensemble predictor to a challenging new real application: the prediction of bi-directional interactions between the circadian clock and metabolism.

3 Methods

3.1 Modelling transcriptional regulation

As in Aderhold et al. (2014), the starting point of our study is the mathematical formulation of transcriptional regulation introduced by Barenco et al. (2006),

$$y_{g,t} \;=\; \frac{dx_{g,t}}{dt} \;=\; \alpha_g + f_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}) - \lambda_g x_{g,t} \tag{1}$$

where $x_{g,t}$ is the mRNA concentration of gene g at time t, α_g is the basal transcription rate for gene g, λ_g is the mRNA degradation rate for gene g, $f_g(.)$ is an unknown regulation function, and $\mathbf{x}_{\boldsymbol{\pi}_g,t}$ is the set of concentrations of the regulating transcription factors $\boldsymbol{\pi}_g$ of gene g at time t. This fundamental equation provides the basis for learning and inference in systems biology, as e.g. described by Lawrence et al. (2010). A common approach is to approximate the time derivative on the left-hand side by a finite difference quotient:

$$\frac{dx_{g,t}}{dt} \;\approx\; \frac{x_{g,t+\Delta t} - x_{g,t}}{\Delta t} \tag{2}$$

For a unit time delay $\Delta t = 1$ this leads to the standard dynamical model:

$$x_{g,t+1} \;=\; x_{g,t} + \alpha_g + f_g(\mathbf{x}_{\boldsymbol{\pi}_g}, t) - \lambda_g x_{g,t} \;=\; h(x_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t}) \tag{3}$$

for some function $h(x_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t})$. This equation provides the basis for a variety of 'dynamic' algorithms, e.g. time-shifted regression methods (Morrissey et al., 2011). As in our earlier work (Aderhold et al., 2014) we adopt an alternative approach based on nonparametric Bayesian modelling with Gaussian processes. The idea is to exploit the fact that the derivative of a Gaussian process is also a Gaussian process (Solak et al., 2002); hence analytic expressions for the mean and the standard deviation of the derivative are available. For the covariance of the Gaussian process, we used the standard squared exponential kernel.[1] We compared this approach with the finite difference method in (2)using two discretization

[1]This is the standard default setting in the R package `gptk` (Kalaitzis et al., 2013).

levels: a "coarse" gradient ($\Delta t = 2$h) and a "fine" ($\Delta t = 24$ min). To model the regulation of metabolites, we modify eq. (1) as follows:

$$y_{g,t} \;=\; \frac{dx_{g,t}}{dt} \;=\; \alpha_g + f_g(\mathbf{x}_{\boldsymbol{\pi}_g,t-\tau}) - \lambda_g x_{g,t} \tag{4}$$

where $x_{g,t}$ now denotes the concentration of metabolite g at time t, and $\tau \geq 0$ is a time delay to allow for the fact that a concentration change of a regulating transcript will not have an immediate effect on the metabolite concentration, due to delays resulting from a series of intermediate steps, like post-translational modification, protein complex formation, translocation etc. We also use eq. (4) instead of eq. (1) for modelling the regulating influence of metabolites back on clock gene expression.

3.2 Network reconstruction methods

We have applied the 15 state-of-the-art network reconstruction methods compared in our previous study (Aderhold et al., 2014) to the data described in Section 4: Graphical Gaussian models (*GGMs*), as proposed by Schäfer and Strimmer (2005); two sparse regression methods: *Lasso* (Tibshirani, 1995) and *Elastic Nets* (Hastie et al., 2001); *Tesla*, which is a combination of Lasso with a changepoint process (Ahmed and Xing, 2009); four hierarchical Bayesian regression methods, as described in Grzegorczyk and Husmeier (2012), with the following modifications: without changepoints (*HBR*); with changepoints representing the light phase, supervised learning (*HBR-light*); with a changepoint segmentation of the target range (the mRNA transcription rate) to approximate Michaelis-Menten nonlinearities, unsupervised learning (*HBR-cp*); and with nonlinear transformations (quadratic and inverse terms) of the mRNA/protein concentration of the putative regulators included (*HBR-nl*), see Aderhold et al. (2014) for details; the Bayesian spline autoregression method proposed in Morrissey et al. (2011) (*BSA*); the graphical Gaussian model implementation of transcriptional regulation proposed in Äijö and Lähdesmäki (2009) (*GP*); state space models (*SSM*) with approximate Bayesian variational inference, as proposed in Beal (2003) and Beal et al. (2005); sparse Bayesian regression (*SBR*) with automatic relevance determination (Rogers and Girolami, 2005); mixture Bayesian networks (*MBN*), as proposed by Ko et al. (2009); a mutual information based approach from Margolin et al. (2006), referred to as *ARACNE*; conventional Gaussian Bayesian networks with the BGe scoring metric (*BGE*), as proposed by Geiger and Heckerman (1994). An overview of the methods is given in Table 1, and the way they are related can be glimpsed from Figure 3.

We used the authors' own software implementations with their default parameter settings where possible, making straightforward modifications to adapt a dynamical model of the form eq. (3) to eqns. (1,4); see Aderhold et al. (2014) for details. Based on eq. (1) or (4), the target variable is the derivative of the concentration at time t, and the explanatory variables are (possibly time-shifted) concentrations of the regulators. Each network thus becomes a bipartite graph, consisting of (i) potentially time-shifted concentrations of potential regulators and (ii) estimated time derivatives of target variables. Hence, the regulators of each individual target variable could be inferred separately and independently with each method, and for each method the corresponding target-specific results (i.e. the interaction

Abreviation	Full Name
HBR	Hierarchical Bayesian regression
HBR-cp	HBR with change-points on gradient
HBR-nl	HBR with additional non-linear terms
HBR-light	HBR with light dependent change-points
Lasso	Sparse regression with L_1 penalty
ElasticNet	Sparse regression with L_1 and L_2 penalty
Tesla	Sparse regression with time-varying change-points
GGM	Graphical Gaussian models
SBR	Sparse Bayesian regression (a.k.a. Automatic Relevance determination)
BSA	Bayesian spline autogression
SSM	State-space models
GP	Gaussian processes
ARACNE	Mutual information measure with pruning
MBN	Mixture Bayesian networks
BGe	Gaussian Bayesian networks

Table 1: **Overview of the methods included in our ensemble studies.** For detailed descriptions of these methods we refer to the original literature publications (see references given in the main text). In our earlier work (Aderhold et al. (2014)) we have provided summaries of these methods, including all implementation details.

strengths between the regulators and targets) could then be merged to obtain the method's overall network prediction.

3.3 Model ensembles

The formation of an ensemble involves two key steps: the selection of models to be included, and the weighting of their predictions. Polikar (2006) and Kuncheva (2004) have reviewed a variety of common approaches, including both weighted and unweighted schemes. From their selection, we have chosen three variants of the Borda count method, which has also been used by Marbach et al. (2012), and three variants of algebraic combination methods, including the mean, the median, and the trim rule. Depending on the type of model included in the ensemble, molecular interaction strengths are either predicted in terms of posterior probabilities (e.g. HBR) or absolute values of regression parameters (e.g. Lasso). For the Borda count methods, only the predicted ranks of the edges (i.e. potential molecular interactions) are needed, which can be accomplished straightforwardly by sorting the edge scores predicted by the individual models. The algebraic methods, on the other hand, are based on operations applied to the actual scores. To this end, all scores that do not already constitute probabilities (like absolute values of regression parameters) were rescaled to the interval $[0, 1]$.

We have started our analysis with the Borda count method promoted by Marbach et al. (2012), which works as follows. If there are n predictors in an ensemble, then for each of the n predictors a regulator-regulatee interaction will receive n points for a first preference, $n-1$ points for a second preference, $n-2$ for a third, and so on. The final ranks predicted by the ensemble are obtained from the sums of these scores. This is a direct adaptation of the method developed by 18th-century political scientist Jean-Charles de Borda for selecting candidates in a democratic election. We additionally explored two alternative Borda count methods, which are adapted from various election systems (the parliamentary elections of Slovenian, Nauru and Oklahoma). In the first alternative, the score given to a candidate regulator-regulatee interaction by each predictor in the ensemble is equal to the number of candidates ranked below it, so that a candidate interaction receives $n-1$ points for a first preference, $n-2$ for a second, and so on, with zero points for being ranked last. In other words, a candidate ranked in ith place receives $(n-i)$ points. As a second alternative, each predictor awards the first-ranked candidate interaction with one point, while the second-ranked candidate receives half of a point, the third-ranked candidate receives one-third of a point, etc. Again, for both alternative methods, the final ranks predicted by the ensemble are obtained from the sums of the individual scores.

In addition to the Borda count method, we have evaluated the performance of three algebraic combination methods reviewed by Polikar (2006): the mean rule, the median rule, and the trim rule. The *mean rule* is a simple operation that averages the regulator-regulatee interaction scores over all the models in the ensemble. It can be considered to be objective, as it does not apply any prior knowledge in the form of a score transformation or model weighting, although implementing a weighted version is straightforward. The *median rule* only differs from the mean rule in that it applies the median operation to the ensemble of regulator-regulatee interaction scores instead of the mean. The *trim rule* is a modification of the mean rule that avoids extreme values by discarding the smallest and largest regulator-regulatee interaction scores.

All these methods, both of the Borda count or algebraic type, can be combined with weighting schemes, to represent prior confidence in the models that form the ensemble, or prior knowledge about the specific application. Marbach et al. (2012) used a weighting scheme based on a known gold standard. As the authors point out themselves, this is only feasible for synthetic toy problems, and it is obsolete for real applications. To achieve an objective evaluation, we have *not* applied any weighting scheme in our work.

A variety of other algebraic combination methods were proposed by Polikar (2006). The *minimum* and *maximum rules* are one-sided modifications of the trim rule, in which only the smallest or the largest scores are discarded, thereby effectively introducing a pessimistic or an optimistic bias. The *product rule* obtains the ensemble score by multiplying the scores from the individual predictors. Since ranking is invariant with respect to a monotonic transformation of the scores, and the logarithm of a product of scores is the sum of the logarithms of the individual scores, $\log(x_1 x_2) = \log(x_1) + \log(x_2)$, the product rule is equivalent to the mean rule applied to log-transformed scores. In the *majority voting* rule, each model selects one and only one of the candidates (in our case: potential regulator-regulatee interactions), in which the candidate with the highest value receives one vote and the remaining predictors no vote at all. The votes are summed over all models in the ensemble and the candidate with the majority of summed votes is selected as the winner. Other voting systems such

as the Condorcet election methods apply majority votes to pairs of candidates in order to elect a winner. The popular Copeland method orders candidates based on pairwise victories and defeats. We have not included any of these methods in our study, for the following reasons. We do not want to introduce an a priori pessimistic or optimistic bias, as inherent in the minimum and maximum rules. We have shown that the product rule is equivalent to the mean rule with the scores subjected to a log transformation. This transformation leads to numerical instabilities for scores close to zero, and also implies that the ensemble prediction is heavily dominated by the lowest scores; such a distortion is intrinsically suboptimal. Likewise, the majority rule is suboptimal, due to the substantial information loss inherent in discarding all scores that are different from the winning score. The Condorcet and Copeland methods stem from social choice theory concerned with collective decision. They also incur information loss and subjective biases (Chevaleyre et al., 2007), which does not appear appropriate in our application.

Instead of the alternative methods from Polikar (2006), we have explored three novel ensemble formation methods that are based on predictor-specific "interaction vectors" of all inferred interaction strengths. Note that as opposed to the alternative ensemble formation schemes studied by Marbach et al. (2012), which are based on the unknown true network structure, our methods draw on information that is directly available after completing the model training schemes and are therefore viable in practice. From the available models we selected those contributing to the ensemble in the following way. For each network reconstruction method m_i $(i = 1, \ldots, 15)$ we build a method-specific "interaction vector" $\mathbf{v}_i$ of all inferred interaction strengths, and we standardize these vectors to Euclidean norm $||\mathbf{v}_i||_e = 1$. We then determine the "median model" m_j, whose interaction vector $\mathbf{v}_j$ has the lowest average pairwise Euclidean distance to the other vectors,

$$j := \arg\min_k \left\{ \sum_{i=1}^{15} ||\mathbf{v}_k - \mathbf{v}_i||_e^2 \right\} \tag{5}$$

Next, we sort the methods with respect to their Euclidean distances, $D_i := ||\mathbf{v}_j - \mathbf{v}_i||_e$, from the median model m_j, to obtain an ordering $m_{\sigma(1)}, \ldots, m_{\sigma(15)}$, where $\sigma(1) = j$ and $D_{\sigma(i)} \leq D_{\sigma(k)}$ for $\sigma(i) < \sigma(k)$. We start with an ensemble consisting only of the median model $E_1 = \{m_{\sigma(1)}\}$, and extend it by successively adding further models until all models are included. Whenever we add a model we apply the mean rule to generate a single ensemble output as previously described. We have compared three different procedures: (1) add the model having the lowest distance from the median method: $E_i := E_{i-1} \cup \{m_{\sigma(i)}\}$, (2) add the model having the highest distance from the median method: $E_i := E_{i-1} \cup \{m_{\sigma(15+1-i)}\}$, (3) add the model having median distance from the median method. If the number of remaining methods is even, there are two candidate methods with median distance, and we then either select the less distant (3a) or the more distant (3b) candidate first. A comparison of these procedures is shown in Figure 4.

3.4 Network inference scoring scheme

All the methods described in Sections 3.2 and 3.3 provide a means by which interactions between genes and proteins can be ranked in terms of their significance or influence. If the

true network is known, this ranking defines the Receiver Operating Characteristic (ROC) curve (Hanley and McNeil, 1982), where for all possible threshold values, the *sensitivity* or recall is plotted against the complementary *specificity*.[2] By numerical integration we then obtain the area under the curve (AUROC) as a global measure of network reconstruction accuracy, where larger values indicate a better performance, starting from AUROC=0.5 to indicate random expectation, to AUROC=1 for perfect network reconstruction.

Another well established measure that is closely related to the AUROC score is the area under the precision recall curve (AUPREC), which is the area enclosed by the curve defined by the *precision* plotted against the *recall*.[3] AUPREC scores have the advantage over AUROC scores in that the influence of large quantities in false positives can be identified better through the precision. There have been suggestions that precision-recall curves indicate differences in network reconstruction performance more clearly than ROC curves (Davies and Goadrich, 2006). While this is true for large, genome-wide networks, we demonstrate here that for the network complexity of interest in our study the differences between the two scoring schemes are negligible (see Figure 7).

3.5 ANOVA

For our performance evaluation on the simulated data described in Section 4.1, we were running hundreds of simulations for a variety of different settings, related to the observation status of the molecular components (mRNA only versus mRNAs and proteins), the method for derivative estimation (described in Section 3.1), the regulatory network structure (shown in Figure 1), and the method applied for learning this structure from data (reviewed in Sections 3.2 and 3.3). The results from these studies are complex, and patterns are not easily discernible by visual inspection, as can be seen from the figures in the Appendix, Figures 12-14. In order to disentangle the different factors, and in particular distinguish the effect of the model from the other confounding factors, we proceeded as in our previous work (Aderhold et al. (2014)) and adopted the DELVE evaluation procedure for comparative assessment of classification and regression methods in Machine Learning (Rasmussen, 1996, Rasmussen et al., 1996) and set up a multi-way analysis of variance (ANOVA) scheme (e.g. Brandt, 1999).

Let Y_{ognmk} denote the AUROC score obtained for observability status o, gradient computation g, network topology n, network reconstruction method m, and data instantiation k. The range of these indices is as follows: $o \in \{0, 1\}$, where $o = 0$ indicates partial (mRNAs only) and $o = 1$ complete (mRNAs and proteins) observation; $g \in \{0, 1, 2\}$, where $g = 0, 1$ denotes derivative approximation with difference quotients according to eq. (2), with $\Delta t = 2h$ and $\Delta t = 24min$, respectively, and $g = 2$ denotes gradient approximation via GP interpolation; $n \in \{0, 1, 2, 3, 4, 5\}$, where $n = 0$ represents 'wildtype' or 'P2010' (the published network topology, shown in in the top left panel of Figure 1), and $n \neq 0$ one of five sparser network modifications in which increasing proportions of feedback interactions have

[2] *Sensitivity:* proportion of true interactions that have been detected; *specificity:* proportion of non-interactions that have been avoided.

[3] *Precision:* proportion of predicted interactions that are true; *recall:* proportion of true interactions that have been detected, i.e. sensitivity

been pruned by disabling certain proteins to function as transcription factors (Figure 1); $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of all models and model ensembles included in our study, and $k \in \{0, 1, 2, 3, 4\}$ for five different data instantiations. We model the AUROC scores with the following ANOVA approach:

$$y_{ognmk} \;=\; O_o + G_g + N_n + M_m + \varepsilon_{ognmk} \tag{6}$$

where $\varepsilon_{ognmk} \sim N(0, \sigma^2)$ is zero-mean white additive Gaussian noise, and O_o, G_g, N_n, and M_m are main effects associated with observation status, derivative approximation, true network structure, and network reconstruction method, respectively. As a sanity check, we carried out standard residual analysis; this did not indicate any violation of the model assumptions, as discussed in the appendix of Aderhold et al. (2014).

3.6 Randomization test

For the application to the real data described in Section 4.2, we decided on the significance threshold, i.e. the value above which a regulatory interaction is regarded as significant, with a randomization test. To this end, we generated 20 randomized data sets in which all time series were permuted; this keeps the marginal distributions of mRNA or metabolite concentrations invariant whilst destroying all genuine correlations. We then applied our model ensemble and determined the distribution of interaction scores (see Figure 10). From these distributions, two significance thresholds were obtained: the value above which 5% of the probability mass can be found; this is the standard threshold of $p = 0.05$ without correction for multiple testing. Alternatively, we recorded the largest value obtained; this threshold controls the family-wise type-I error at significance level $p = 0.05$, i.e. a correction for multiple testing is included.

3.7 Determining time delays

We want to apply a method ensemble to infer the interactions between genes involved in circadian regulation and metabolism in plants, as described in Subsection 4.2. For this real-world application we have to determine the optimal time delay τ in eq. (4). Denote by $\boldsymbol{\theta}$ the vector of model parameters, and by $\mathcal{D}$ the data. We want to determine the time lag τ that maximizes the posterior probability $p(\tau|\mathcal{D})$. From Bayes rule, $p(\tau|\mathcal{D}) \propto p(\mathcal{D}|\tau)p(\tau)$, and on the assumption of a uniform prior $p(\tau) = C$, this requires optimization of the marginal likelihood:

$$p(\mathcal{D}|\tau) = \sum_{\mathcal{M}} \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} p(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}, \boldsymbol{\phi}|\tau) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{7}$$

where the sum is over all valid network structures $\mathcal{M}$ and the integrals are over all network parameters $\boldsymbol{\theta}$ and all hyperparameters $\boldsymbol{\phi}$, respectively. The integral, given in eq. (7), is analytically intractable. One can resort to standard numerical procedures based on MCMC, like Chib's method (Chib and Jeliazkov, 2001). However, for the hierarchical Bayesian models

employed in our study, we can apply a numerically more efficient procedure. The marginal likelihood in eq. (7) can be written as:

$$p(\mathcal{D}|\tau) \;=\; \int_{\boldsymbol{\phi}} \left(\sum_{\mathcal{M}} \left(\int_{\boldsymbol{\theta}} p(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}|\boldsymbol{\phi}, \tau) d\boldsymbol{\theta} \right) \right) p(\boldsymbol{\phi}) d\boldsymbol{\phi} \tag{8}$$

and we apply the following three-prong approach. Firstly, for a given network structure $\mathcal{M}$ and fixed hyperparameters $\boldsymbol{\phi}$ and under certain regulatory conditions (Grzegorczyk and Husmeier (2013)), the inner marginal integral can be computed in closed form :

$$p(\mathcal{D}, \mathcal{M}|\boldsymbol{\phi}, \tau) = \int_{\boldsymbol{\theta}} p(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}|\boldsymbol{\phi}, \tau) d\boldsymbol{\theta} \tag{9}$$

Secondly, under exploitation of model modularity and a modest restriction on the network connectivity (fan-in limitation), the marginalisation over network structures can be accomplished with polynomial time complexity. We obtain:

$$p(\mathcal{D}|\boldsymbol{\phi}, \tau) = \sum_{\mathcal{M}} p(\mathcal{D}, \mathcal{M}|\boldsymbol{\phi}, \tau) \tag{10}$$

where the sum is over all valid network structures. The only remaining part that requires a numerical approximation is the third: the integration over the hyperparameter(s):

$$p(\mathcal{D}|\tau) = \int_{\boldsymbol{\phi}} p(\mathcal{D}|\boldsymbol{\phi}, \tau) p(\boldsymbol{\phi}) d\boldsymbol{\phi} \tag{11}$$

This is a low-dimensional (one-dimensional) integral, which can be efficiently solved numerically with Monte Carlo integration by repeatedly sampling $\boldsymbol{\phi}$ $(g = 1, \ldots, G)$ from the prior distribution $p(\boldsymbol{\phi})$ and computing the following Monte-Carlo estimator:

$$\widehat{p(\mathcal{D}|\tau)} = \frac{1}{I} \sum_{i=1}^{I} p(\mathcal{D}|\boldsymbol{\phi}^{(i)}, \tau) \tag{12}$$

where $\boldsymbol{\phi}^{(i)}$ $(i = 1, \ldots, I)$ is the i-th sample from $p(\boldsymbol{\phi})$. For the hierarchical Bayesian regression (HBR) model, we provide the mathematical details of this three-prong approach in the appendix.

4 Data

For the purpose of assessing single model and model ensemble learning accuracy we adapted a system of Markov Jump processes that realistically simulate molecular interactions. These include the transcription of mRNA and translation of corresponding proteins in the central circadian clock of *Arabidopsis thaliana*. Since the ground truth is known from the mathematical model, the generated data can be used for an objective performance evaluation. The Biopepa framework simulates the expression of mRNA and translation of proteins of the circadian core clock in the plant *Arabidopsis thaliana*. Furthermore we applied the findings from the synthetic evaluation to data that involve core clock genes and growth related metabolites in *Arabidopsis thaliana*. The data was derived from the TiMet project (Flis et al., 2013) and constitutes qRT-PCR measurements of mRNA and metabolite measurements in plant leaves.

4.1 Simulated mRNA and protein concentration time series

For the purpose of assessing single model and model ensemble network reconstruction accuracy, we followed Guerriero et al. (2012) and generated realistic mRNA and protein concentration time series from a published gene regulatory network of circadian regulation in *Arabidopsis thaliana* (Pokhilko et al., 2010). The simulations are based on Markov jump processes and describe the elementary molecular reactions of transcription initiation, translation, post-translational modification, degradation etc. as discrete stochastic events. This avoids the artificial and unrealistic limit cycle behaviour of models based on ordinary differential equations. Various mathematical models based on ordinary differential equations (ODE) have been previously employed in modeling the circadian clock in *Arabidopsis thaliana* (Locke et al., 2006, Pokhilko et al., 2012, 2013). The resulting molecular profiles commonly exhibit pronounced regular oscillations that lack any stochasticity normally observed in biological data. For a more realistic approach, we model the individual molecular processes of transcription, translation, degradation, dimerization etc. as individual discrete and stochastic events. Like Guerriero et al. (2012) we adopted the Bio-PEPA[4] framework to simulate gene and protein expression profiles for the core circadian clock of *Arabidopsis thaliana* with the Gillespie algorithm (Ciocchetta and Hillston, 2009). A full list of reactions and their corresponding mathematical descriptions is available from the supplementary material of Guerriero et al. (2012). The underlying regulatory network, taken from Pokhilko et al. (2010), is shown in Fig 1, and we additionally took 5 less densely connected network variants, from Aderhold et al. (2014) (see Figure 3 in that paper), in which various feedback interactions had been pruned via complete targeted gene knockouts. For each of these networks we created 11 interventions, in emulation of the biological protocols from Flis et al. (2013). These interventions include knock-outs of the genes GI, LHY, TOC1, and the double knockout PRR7/PRR9, and varying photo-periods of 4, 6, 8, 12, and 18 hours as well as a full dark (DD) and a full light (LL) cycle, each following a 12h-12h light-dark cycle entrainment phase over 5 days. Protein and mRNA concentration samples were taken after entrainment in 2h intervals for 1 day for each intervention, resulting in a total of 143 observations for each gene and associated protein displayed in Fig 1. In addition to keeping all data, we emulated transcription-only profiling assays and discarded the protein concentrations.

4.2 Arabidopsis metabolite and mRNA time series

A major objective of our study is to improve the reconstruction accuracy of interactions between genes involved in circadian regulation and metabolism in plants. We have used recent concentration time series from the EU project Timet (Flis et al., 2013), which include 10 core circadian clock genes and 11 growth related metabolites of *Arabidopsis thaliana*, measured at the same time points and conditions. The metabolite data were obtained from spectrophotometric assays and include measures of total protein content, starch, glucose-6-phosphate (g6p), nitrate, total amino acids (aa), fructose, fumarate, glucose, chlorophyll-A & B, and malate. The transcriptional profiles were obtained with qRT-PCR, and include the 10 core clock genes identified in the literature (Pokhilko et al., 2010, 2012): LHY, CCA1, NI

[4]http://www.biopepa.org

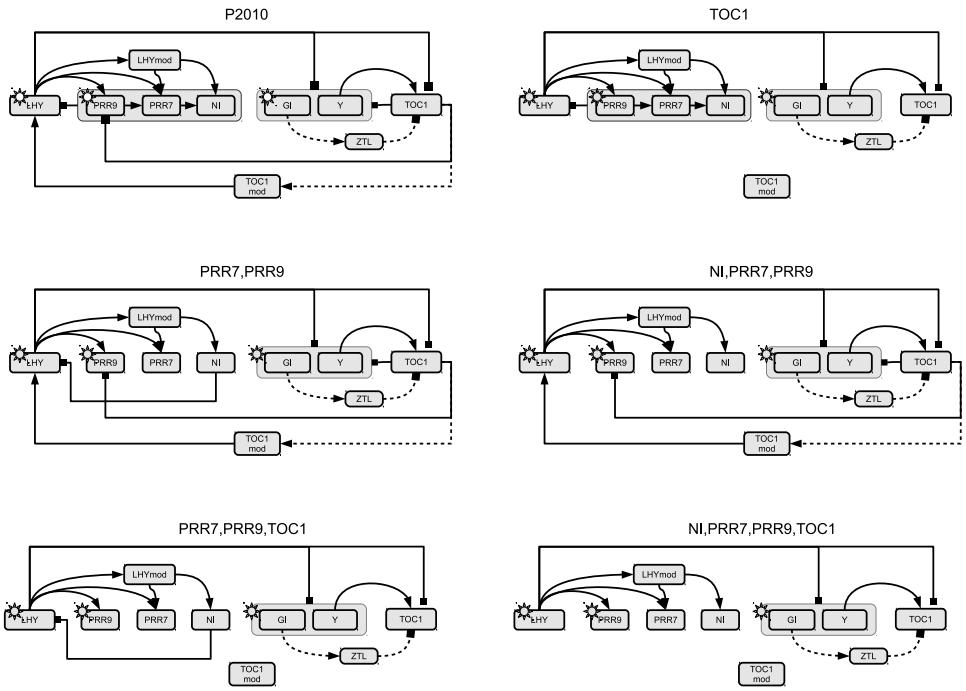

Figure 1: **Circadian clock network and modifications.** The top left panel (P2010) shows the circadian clock network from Pokhilko et al. (2010). The remaining networks are modifications, where certain transcription factors have been disabled to decrease the network complexity and inhibit certain feedback loops (illustration adapted from Aderhold et al. (2014)). The benchmark data described in Section 4.1 were generated for each of the displayed networks with the stochastic process model described in Guerriero et al. (2012). Solid lines show interactions between genes via transcription factors, dashed lines represent protein modifications, and the sun symbol indicates the direct regulatory influence of light. Arrow heads: activation. Squares: inhibition.

(PRR5), PRR7, PRR9, TOC1, ELF3, ELF4, LUX, and GI. The data stem from the leaves of various genetic variants of *Arabidopsis thaliana*, and encompass two wildtypes of the strains Columbia (Col-0) and Wasilewski (WS) and 5 clock mutants, namely a double knock-out 'LHY/CCA1' in the WS strain, a single knock-out of 'GI' and 'TOC1' in the strain Col-0, a double-knockout 'PRR7/PRR9' in strain Col-0, and a single knock-out of 'ELF3'. The plants were grown in varying light conditions: a diurnal cycle with 12 hours light and 12 hour darkness (12L/12D), an extended night with full darkness for 24 hours (DD), and an extended light with constant light (LL) for 24 hours. An exception is the 'ELF3' mutant, which was grown only in 12L/12D condition. Samples were taken every 2 hours to measure mRNA and metabolite concentrations, yielding a total of 266 observations from all listed interventions. Further information on the data and the experimental protocols is available from Flis et al. (2013). For our study, we extracted the transcription profiles of the core clock mRNA that are included in the models from the literature (Guerriero et al., 2012, Pokhilko et al., 2010): LHY, CCA1, NI (PRR5), PRR7, PRR9, TOC1, ELF3, ELF4, LUX, and GI. The metabolite data consists of the profiles protein, starch, glucose-6-phosphate (g6p), nitrate, amino acids (aa), fructose, fumarate, glucose, chlorophyll-A & B, and malate. An additional binary light indicator variable with 0 for darkness and 1 for light was included to indicate the status of the experimentally controlled light condition.

5 Results and Discussion

5.1 Model ensemble

Combination approaches We have applied six combination strategies, discussed in Section 3.3, to combine the 15 methods from Table 1 into an ensemble. We have then used the outputs from the ensemble to reconstruct the regulatory networks. Figure 2 shows the comparison of the six approaches on the simulated data described in Section 4.1. The left panel displays the AUROC and the right panel the AUPREC scores, including confidence intervals, as obtained with the ANOVA scheme from Section 3.5. It can be observed that the mean-rule outperforms the remaining combination methods, in particular the Borda count scheme from Marbach et al. (2012). The difference to the trim rule is negligible, which indicates that the most extreme models do not exert a disturbing influence on the ensemble. The median-rule shows a significantly lower performance than the mean-rule. This slightly unexpected finding suggests that the more extreme scores include relevant information, and that effectively eliminating their influence with a "robust" combination method is counterproductive. The most important finding is that the mean-rule outperforms the majority of the Borda count methods, in particular the Borda count variant that was applied in Marbach et al. (2012) (Borda-1). This improvement in performance is a consequence of the information loss inherent in ranking. For the Borda count methods, all model scores are first ranked, and the ranks are then averaged. For the algebraic mean-rule, the model scores are first averaged, and then ranked. When an ensemble consists of M models, the Borda count scheme incurs an information loss M times (for each of the models in the ensemble), whereas the algebraic mean-rule incurs an information loss only once. This difference explains the boost in performance with the mean-rule. The best Borda count method, and the only Borda

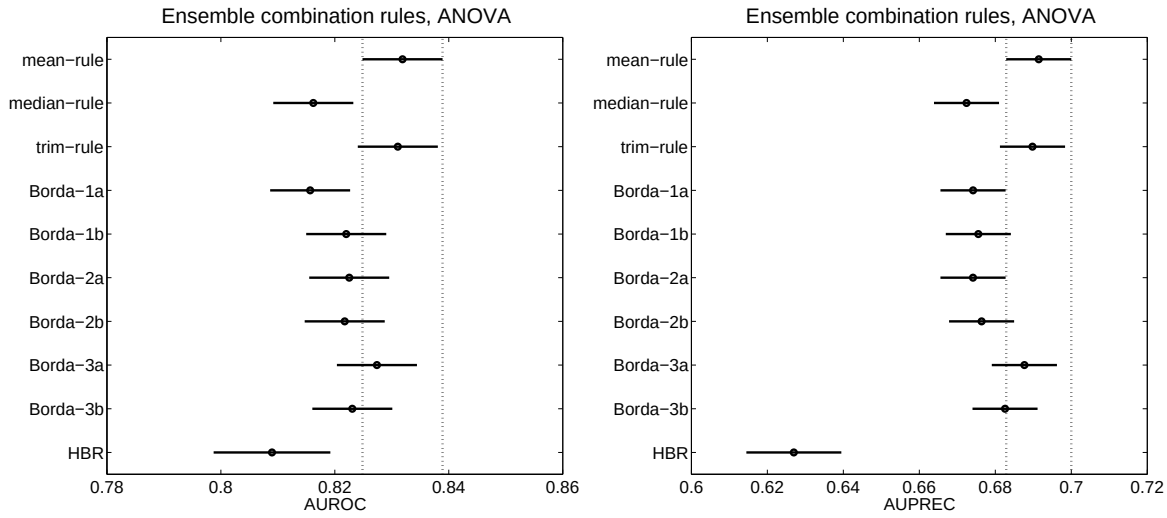

Figure 2: **Comparison between different model combination strategies for ensemble formation.** The top rows show three algebraic combination methods: the mean rule, the median rule, and the trim rule. The remaining rows show three Borda count variants (Borda1-3). All of these methods are described in Section 3.3. The Borda count methods were applied in two subvariants, with (symbol b) or without (symbol a) randomly breaking ties between tied model outputs. The bottom row serves as a reference and presents the performance of the best individual model (HBR) in the ensemble. *Left panel:* AUROC scores and confidence intervals. *Right panel:* AUPREC scores and confidence intervals. In both cases, the ANOVA scheme of Section 3.5 was applied, and the scores correspond to M_m in equation (6).

count variant for which the deterioration over the mean-rule is not significant, is *Borda-3a*, which assigns fractions to the ranks and thereby puts slightly higher weight on the higher scores. It is encouraging to observe that all combination strategies outperform the best method in the ensemble, HBR in the bottom row of Figure 2; these findings are consistent with Marbach et al. (2012). For the remaining studies, we have chosen the mean-rule as the best performing ensemble method.

Ensemble formation For the simulated data from Subsection 4.1, we set the time delay fixed at $\tau = 0$, corresponding to eq. (1), and we first analyze the data with the individual network reconstruction methods, listed in Table 1. Subsequently we build the method-specific predicted interaction strength vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{15}$ and determine the "median" model, as described in Subsection 3.3.[5]

The "median" method is given by the Gaussian Graphical models (GGM). To visualize the (dis-)similarities among the 15 individual network reconstruction methods, we apply a standard principal component analysis (PCA). Projecting the method-specific interaction strength vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{15}$ onto the first two principle components yields the PCA plot shown in Figure 3. As one would expect, similar methods (like Lasso and elastic nets, or different HBR variants) are close together and the median model (GGM) has a central position, indicating that the 2-dimensional PCA plot conserves most of the information. In particular it can be seen from the PCA plot that there are two outliers (MBN and ARACNE), which are located far away from the other methods. Further investigation revealed that these two methods tend to infer sparser networks and, thus, overall systematically lower interaction strengths than the other methods.

Figures 4 and 5 show the network reconstruction performance of different model ensembles on the simulated data of Section 4.1. The reconstruction performance was quantified in terms of AUROC and AUPREC scores (see Subsection 3.4), and the ANOVA scheme (see Subsection 3.5) was applied, where each model ensemble constitutes a separate main effect M_m ($m = 1, \ldots, 15$) in eq. (6). In the following description of the results we focus on the AUROC scores, i.e. Figure 4. We first ranked the individual models according to their performance (Figure 4a) and added them according to a best-first schedule to the ensemble. Figure 4b shows that the ensemble performance steadily increases with increasing ensemble size until about half the models are included. At this point the performance reaches a plateau, with only weak variations as the ensemble size further increases. Adding models to an ensemble according to a best-first schedule is not viable in practice, as the actual model performance is typically unknown. We therefore tried the strategies described in Section 3.3. From the vectors of predicted interaction strengths, $\mathbf{v}$, we determined the "median" model (GGM), as defined in eq. (5), and added further models to the ensemble based on their distance in $\mathbf{v}$-space. When adding the closest models first (Figure 4c) or most distant models first (Figure 4d), the ensemble performance initially deteriorates. Both strategies are suboptimal in a "bias-variance decomposition" sense. The "closest-first" strategy (Figure 4c) does not give the ensemble sufficient variance, while the "most distant first" strategy

[5] As described in Subsection 3.5, our simulation setting features 6 network topologies n (different Arabidopsis mutants), 2 different data types o (partial and complete), 3 different gradients g ($\Delta t = 2h$, $\Delta t = 24min$, and GP interpolation), and 5 independent data instantiations k for each scenario, so that each method-specific vector $\mathbf{v_i}$ contains 9900 individual interaction strengths in total.

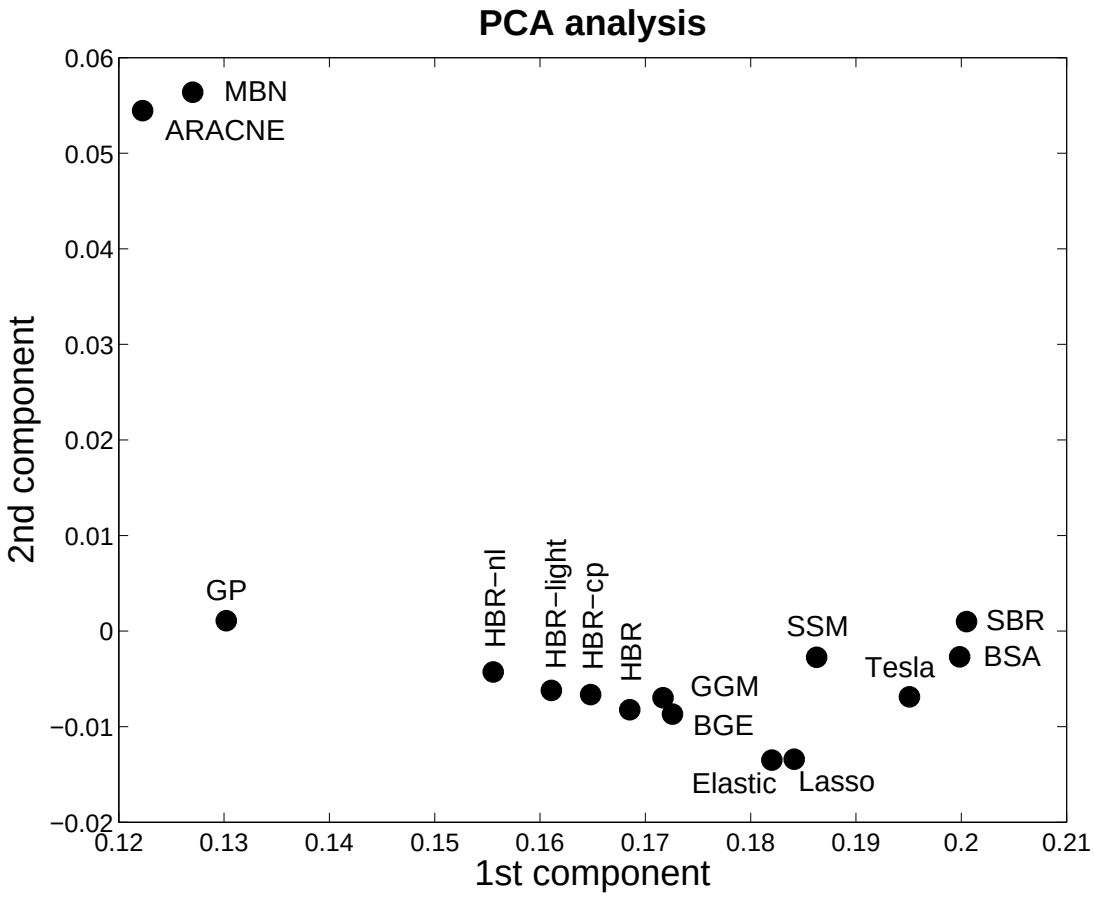

Figure 3: **Principle component plot of the 15 methods included in our ensemble study.** For a method overview see Table 1. The two principle components were computed from the high-dimensional vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{15}$ of predicted interaction strengths. As one would expect, related methods, like the group of HBR methods, or Lasso and elastic nets, are closely grouped together. Note the large distance to the outliers MBN and ARACNE, which were the methods with the poorest network reconstruction accuracy (see top left panels in Figures 4 and 5).

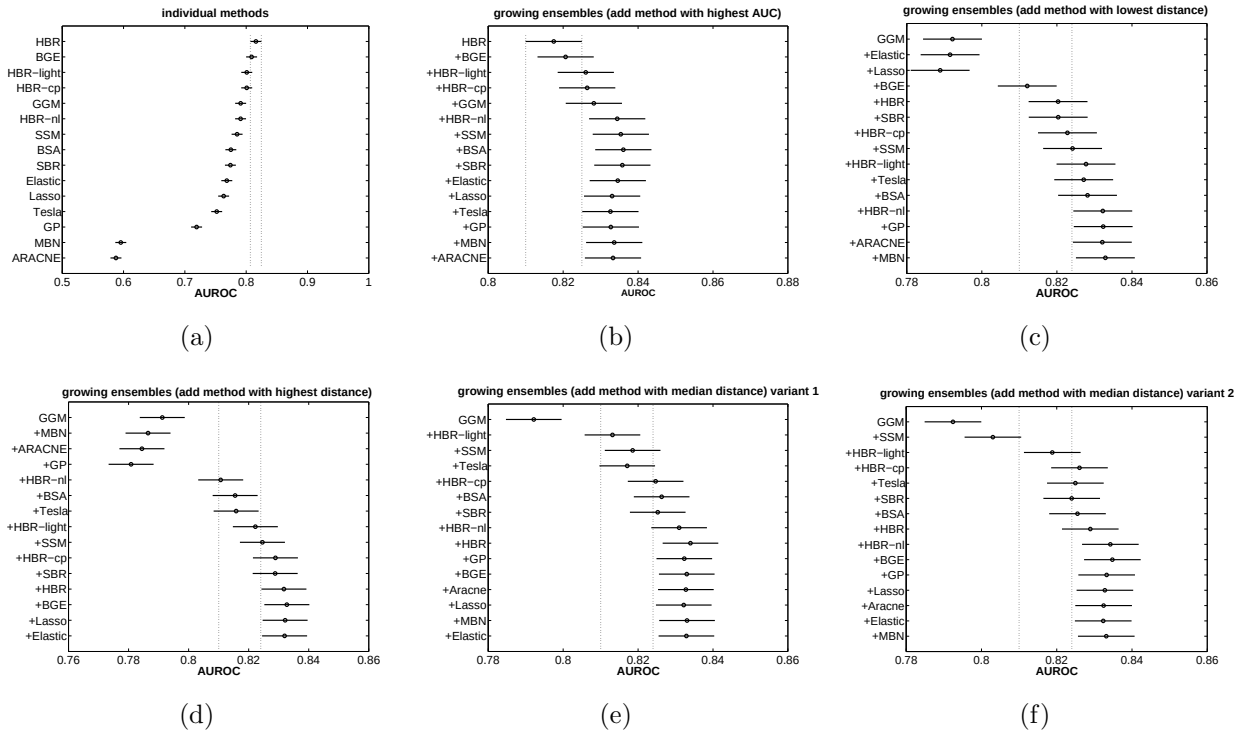

Figure 4: **Performance of different model ensemble formation strategies in terms of AUROC scores.** For the full method names behind the abbreviations see Table 1. The panels show AUROC confidence intervals for the ANOVA main effect related to the model or ensemble, M_m in eq. (6). **Panel (a)** shows the AUROC scores of the individual models, reproducing the results from Aderhold et al. (2014). **Panels (b-e)** show how the AUROC values depend on ensembles of growing size (from top to bottom), for different ensemble formation schemes. Starting with a single model (top of each panel), new models (indicated by the labels on the vertical axis) are successively included in the ensemble as one descends from the top of the panel, until one reaches the complete ensemble (including all models) at the bottom. **Panel (b):** Models are included in the ensemble according to their individual AUROCs scores (in descending order). **Panels (c-f):** Starting with the median model, as defined in eq. (5), further models are added according to the different strategies described in Section 3.3: **Panel (c):** minimal distance first; **Panel (d):** maximal distance first; **Panels (e-f):** median distance first (two different ways of breaking the tie for even numbers). For comparison, the dotted reference lines indicate the confidence interval for the best-performing method (HBR) from **Panel (a)**. Complementary AUPREC score confidence intervals are provided in Figure 5

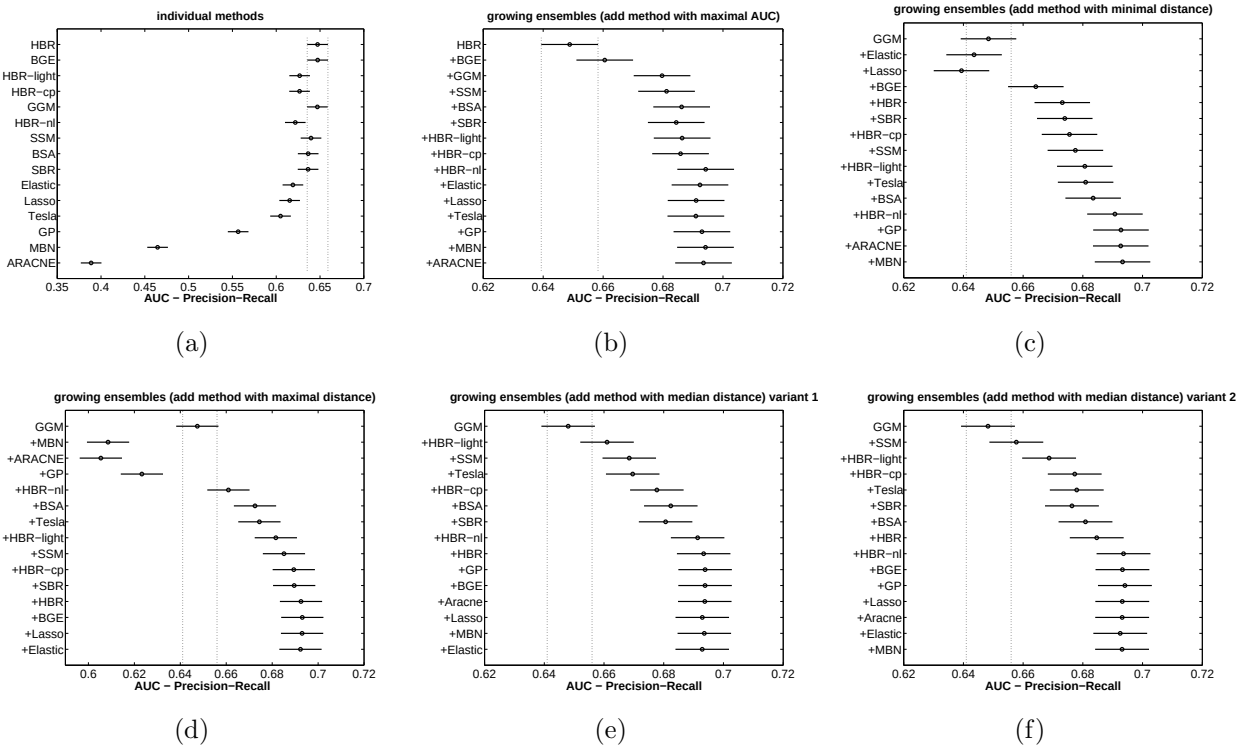

Figure 5: **Performance of different model ensemble formation strategies in terms of AUPREC scores.** This figure is identical to Figure 4 except that AUPREC scores have been used. See caption of Figure 4 for further details.

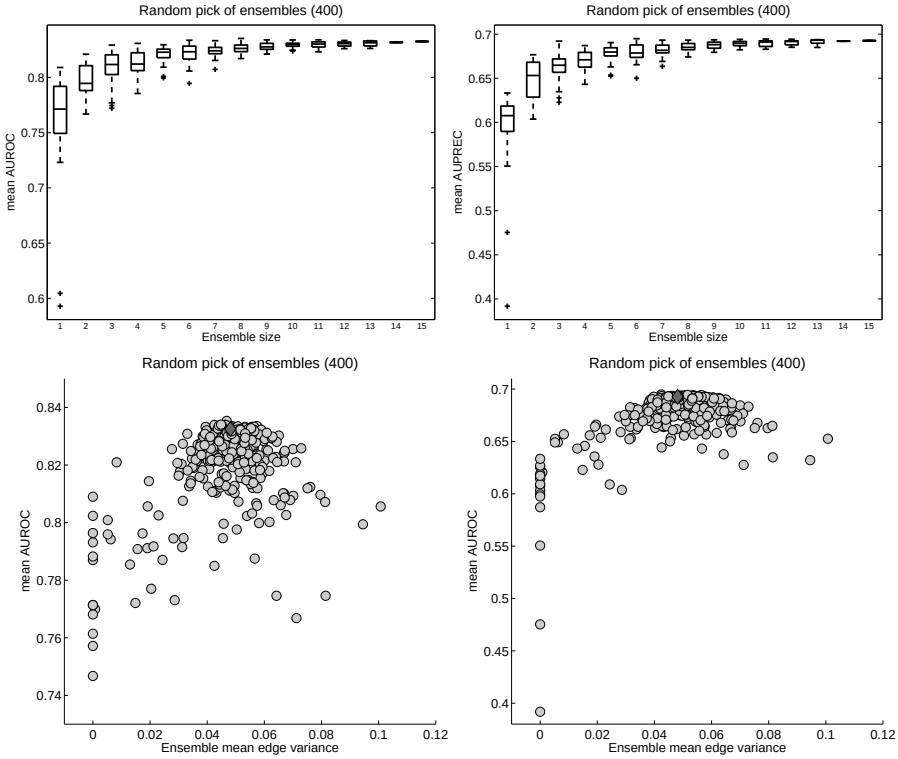

Figure 6: **Model ensemble size and network reconstruction accuracy. Top row:** Distribution of network reconstruction accuracies M_m in eq. (6) in terms of AUROC (top left) and AUPREC (top right) scores, for model ensembles of different size. A random sample of 400 different model ensembles was generated, and the panel shows the distribution of M_m (in terms of AUROC) for different ensemble sizes. **Bottom row:** Alternative representation, in which the AUROC (bottom left) and AUPREC (bottom right) scores M_m are plotted against the mean edge variance; this is the variance of the interaction strength across all models in the ensemble, averaged over all interactions in the network. The complete ensemble including all models is marked by a darker diamond.

METHOD	HBR	BGe	HBR-light	HBR-cp	GGM	HBR-nl	SSM	BSA
AUROC	0.016	0.023	0.031	0.031	0.041	0.041	0.047	0.057
AUPREC	0.045	0.045	0.065	0.065	0.045	0.070	0.052	0.055
METHOD	SBR	Elastic	Lasso	Tesla	GP	MBN	ARACNE	**MEAN**
AUROC	0.057	0.063	0.068	0.080	0.113	0.236	0.244	**0.077**
AUPREC	0.056	0.073	0.077	0.087	0.135	0.227	0.303	**0.093**

Table 2: **Improvements achieved with the method ensemble.** This table provides the average improvements that the ensemble of all methods yields over the 15 individual methods in terms of AUROC and AUPREC differences. All differences are significant at the level $p = 0.05$. The average AUROC and AUPREC gains over all methods are provided in the last panels (**'MEAN'**). Figure 7 shows a scatter plot of the total AUROC and AUPREC scores. For the full method names behind the abbreviations see Table 1.

(Figure 4d) leads to the early inclusion of very poor models in the ensemble ("bias"). A reasonable compromise between these two extremes seems to be the "median first" strategy (Figures 4e-f), which avoids the initial decline in performance. Interestingly, none of these methods achieves a clear peak in performance before the maximum ensemble size is reached. Figure 5 shows the complementary AUPREC score presentation of Figure 4. A comparison of the AUROC (Figure 4) and the AUPREC (Figure 5) scores reveals very similar trends. Only for the 15 individual models (see panels (a)) the ranking of changes; with SSM, BSA, SBR performing slightly better and HBR-light and HBR-cp performing slightly worse in terms of AUPREC scores than in terms of AUROC scores. However, with respect to the ensemble building strategies exactly the same trends can be observed. Adding successively either the closest models (Figure 5(c)) or the most distant models (Figure 5(d)), the ensemble performance initially deteriorates, while the "median first" strategy avoids the initial decline (Figure 5(e-f)). Finally, all four strategies reach a plateau. To relax the restriction of the ensemble growth path, we generated a large random sample of model ensembles. The top row of Figure 6 shows the distribution of ensemble performance scores as a function of the ensemble size. The bottom row of Figure 6 shows an alternative representation, where the ensemble performance is plotted against the mean edge variance in the ensemble. Figure 6 suggests that there is not much room for improvement over the complete ensemble, in which all models are included.

All panels in Figures 4-5 demonstrate that the complete model ensemble outperforms all individual methods. To make that more explicit Figure 7 shows a scatter plot of the total AUROC and AUPREC scores of the individual methods, and thereby also indicates the performance of the complete method ensemble and the average performance of all individual methods. From the scatter plot it can be seen that both network inference scoring schemes (AUROC and AUPREC) described in Subsection 3.4 are strongly correlated (Pearson correlation coefficient: 0.9704) and that the method ensemble performs better than the average method and also outperforms all individual methods. The average AUROC and AUPREC differences between the method ensemble and the individual methods are summarized in Table 2. Even the best model (HBR) is outperformed by the ensemble despite the inclusion of poor models with inferior network reconstruction performance. In terms of average

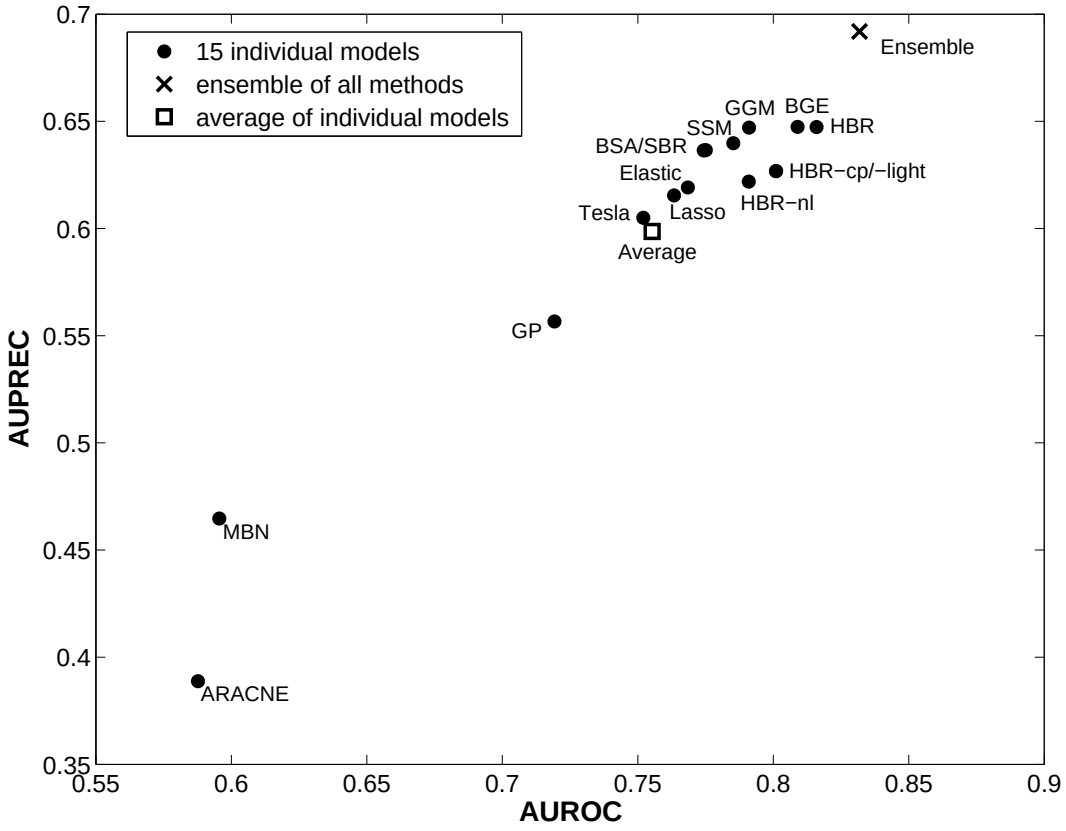

Figure 7: **Scatter plot of the AUROC and AUPREC scores of the individual methods.** The plot shows that the AUROC and AUPREC scores of the 15 individual methods, listed in Table 1, are strongly correlated (Pearson correlation coefficient: 0.9704). There are two additional points indicating the average performance of all 15 methods (cross symbol) and the performance of the ensemble of all 15 methods (square symbol). The AUROC and AUPREC score differences between the method ensemble and the individual methods are provided in Table 2.

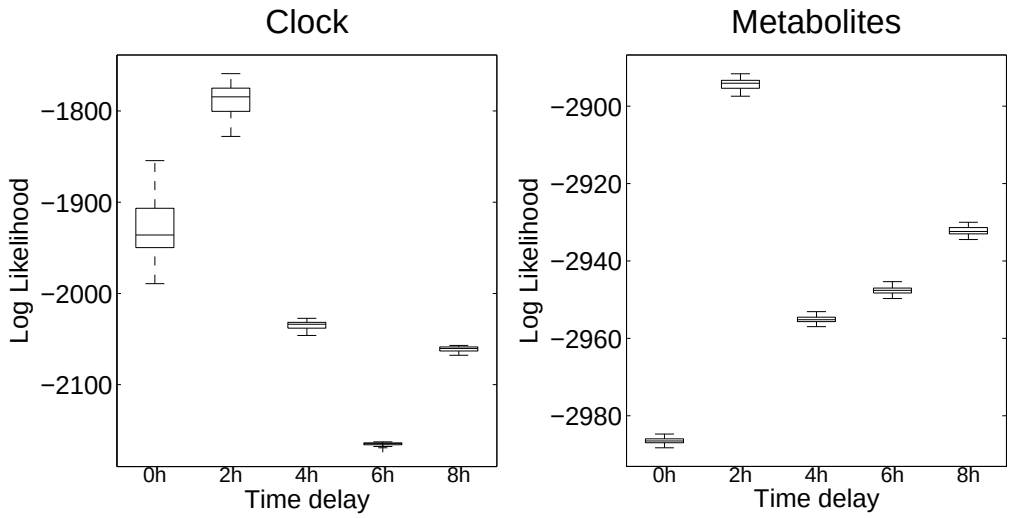

Figure 8: **Determination of adequate time delays for clock genes and metabolites.** The figure shows boxplot representations of our Monte Carlo estimates of the marginal log likelihood for different time delays $\tau \in \{0h, 2h, 4h, 6h, 8h\}$. The technical details of our Monte Carlo estimator are provided in Subsection 3.7 and the appendix. *Left panel:* Regulation of clock gene mRNAs. *Right panel:* Regulation of metabolites. Both panels clearly suggest that the maximal marginal likelihood values are reached for the time lag $\tau = 2h$.

AUROC and AUPREC differences the gain over the best model (HBR) is given by 0.016 (AUROC) and 0.045 (AUPREC). In the first instance, the gain may not appear substantial, though these differences are statistically significant. However, in practical applications the best method is usually unknown. Hence, a fairer figure of merit is the average performance gain over the individual methods. In terms of AUROC and AUPREC differences the average gain over all methods is given by 0.077 (AUROC) and 0.093 (AUPREC), see Table 2.

5.2 Predicted clock-metabolite interactions

For the real data from Subsection 4.2, we allowed for non-zero time delays, generalizing eq. (1) by eq. (4). Figure 8 shows that the optimal time delay for both metabolic and transcriptional regulation is about two hours ($\tau = 2h$). The non-zero time delay in the latter case results from the fact that the protein concentrations are missing, and mRNA concentrations have to be taken as proxy for missing transcription factor activities.

The results of our randomization tests to determine the significance thresholds for the network interaction strengths are given in Figures 9 and 10. As described in Subsection 3.6, we generated randomized data sets to determine the distribution of spurious network interaction scores. From these distributions, represented as histograms in Figure 10, we determined thresholds to extract the significant network interactions. Figure 9 shows the proportions of edges selected under varying settings of the threshold for the different groups of interactions. The graphs show that the proportion of selected edges corresponding to interactions

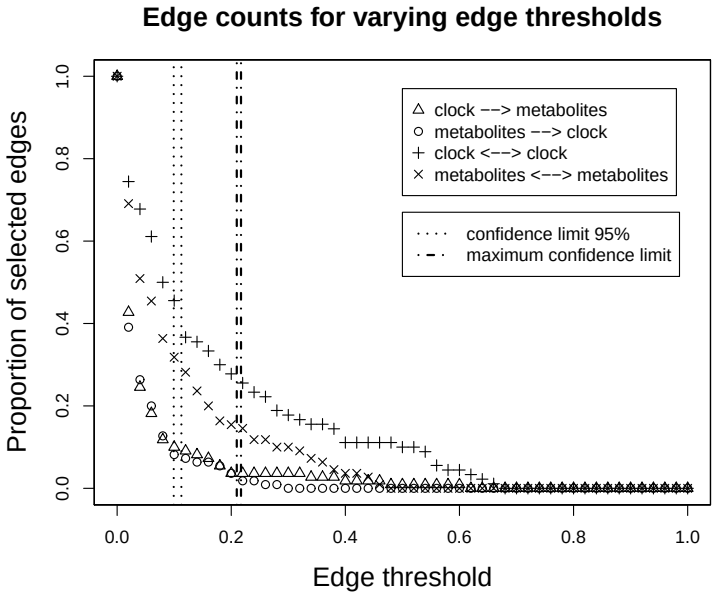

Figure 9: **Distribution of predicted interaction strengths.** The figure shows the proportion of molecular interactions (vertical axis) exceeding a cutoff threshold (horizontal axis) for four molecular groups: (i) within the clock genes, (ii) within the metabolites, (iii) clock genes regulating metabolites, and (iv) metabolites acting back on the clock genes. The vertical lines indicate the $p = 0.05$ significance thresholds for groups 3 and 4 without (dotted lines) and with (dash-dotted lines) correction for multiple testing, using the randomization test from Section 3.6 (results displayed in Figure 10).

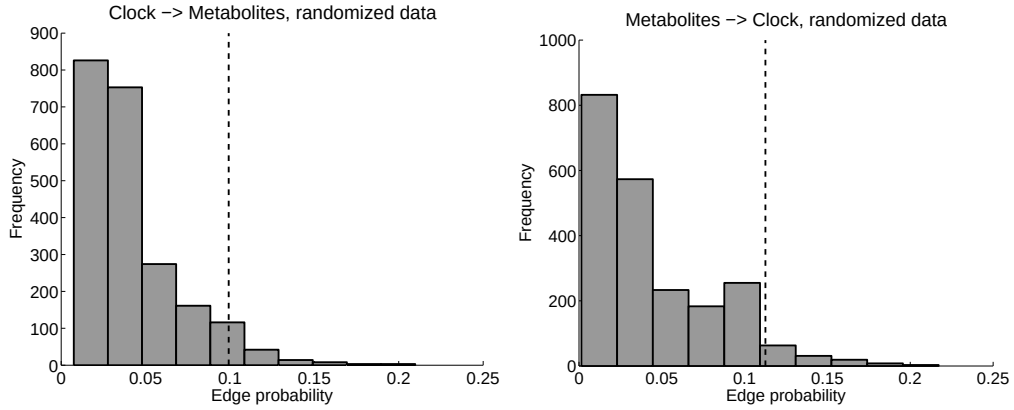

Figure 10: **Significance threshold determination by randomization tests.** The model ensemble described in Section 3.3 was applied to the mRNA and metabolite concentration time series described in Section 4.2. *Left panel:* Distribution of the strengths of regulatory interactions from clock genes to metabolites, obtained from randomized data. The vertical dashed line indicates the point above which 5% of the probability mass can be found. *Right panel:* Idem, for regulatory interactions from metabolites to genes.

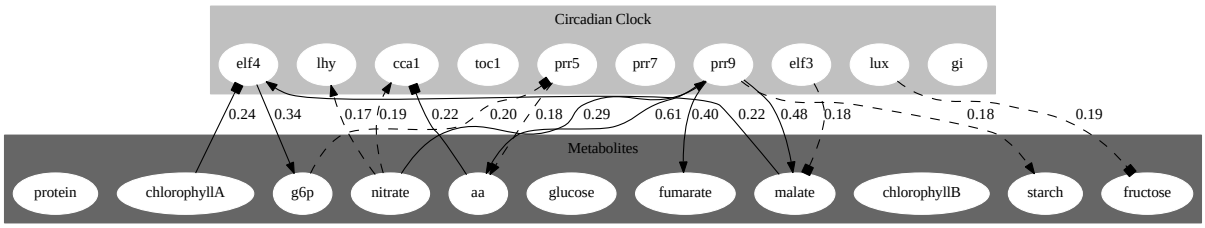

Figure 11: **Predicted bi-directional interactions between the circadian clock and metabolites.** The model ensemble described in Section 3.3 was applied to the mRNA and metabolite concentration time series from Subsection 4.2. Interactions between metabolites and clock genes that are significant at the $p = 0.05$ level without (dashed lines) and with (solid lines) correction for multiple testing. Arrow heads: activation. Circles: inhibition. For improved clarity and to avoid extensive clutter in the diagram we focus on the feedback mechanism between the circadian clock and metabolism, given that interactions within the circadian clock and within the metabolism have already been more widely studied in the literature. For interactions within the clock genes, see Locke et al. (2006), Kolmos et al. (2009), Herrero et al. (2012), Pokhilko et al. (2010), Pokhilko et al. (2012), and Pokhilko et al. (2013) as well as Figure 12 in Aderhold et al. (2014). Interactions among the metabolites have for example been illustrated in Figure 4 in Gille et al. (2011).

among clock genes and among metabolites dominate, and that there are far fewer interactions between clock genes and metabolites.

Figure 11 shows the bi-directional regulatory network between the circadian clock genes and various metabolites, as predicted by our model ensemble. Dashed lines indicate interactions that are significant at the $p = 0.05$ level with correction for multiple testing; dashed lines are interactions that are significant at the $p = 0.05$ level without correction for multiple testing (see Subsection 3.6). The threshold values have been determined by randomization tests.

The first striking feature is that three clock genes are connected to components of the starch synthesis pathway, including *fructose*, *glucose-6-phosphate*, and *starch*. This seems to be in line with the main function of the circadian clock, which is to control what proportion of carbon assimilated during the day is to be accumulated as starch in the leaves (Feugier and Satake, 2012). The regulation of the monosaccharide *fructose* is located near the root of the starch biosynthesis pathway; see e.g. Figure 4 in Gille et al. (2011). Further downstream is *glucose-6-phosphate*, which is regulated by ELF4 and occupies a central position in the cascade that converts fructose derivative *fructose-6-phosphate* and *glucose* into *glucose-1-phospate*, which is the first committed step in starch synthesis (Geigenberger, 2011) . Hence, our study suggests that the circadian clock controls starch synthesis by regulating early, middle and late stages of the starch biosynthesis pathway. *Malate* and *fumarate* are tricarboxylic acid (TCA) cycle metabolites, and our models predict that both are controlled by PRR9. This prediction is consistent with the findings in Fukushima et al. (2009) of a tight link between the circadian clock and the TCA cycle with the difference that our prediction points to ELF3 rather than PRR9 as a repressing factor; in fact, PRR9 is inferred as activating

malate and fumarate. *Nitrate* is known to enhance plant growth in functioning as a nutrient and a signal that reprograms carbon metabolism and resource allocation (Wang et al., 2000). In the latter context it seems intriguing that *nitrate* has an activating influence on the morning genes LHY and CCA1, and also effects PRR9, which in turn controls two of the TCA metabolites, as noted above. The interaction pointing from *chlorophyll a* to ELF4 suggests a light related signal to the clock. In fact, ELF4 has been proposed to play an important role in phytochrome signalling to the clock (Kikis et al., 2005). Since phytochromes are involved in the synthesis of chlorophyll, this interaction might constitute a proxy to a photosensitive negative signal regulating ELF4. Finally, our models predict a direct regulating influence of the clock genes on amino acid content, but not on protein content, which is consistent with the fact that our models are designed to learn direct interactions and suppress indirect ones.

6 Conclusion

The focus of our study has been the reconstruction of regulatory networks related to circadian regulation. Rather than aiming to indentify the "best" reconstruction method, as we pursued in our previous work (Aderhold et al., 2014), we have shown that a significant boost in performance can be achieved with a model ensemble, in independent confirmation of related work in which we have been involved (Marbach et al., 2012). Whilst it is well-known in the machine learning community that the performance of a model ensemble is typically better than the average model performance, we have demonstrated that the ensemble also outperforms the *best* individual model, despite the inclusion of models with inferior performance. In extension of Marbach et al. (2012) we have explored various alternative schemes for selecting the models to be included in the ensemble. We have not found any clear performance optimum along the ensemble growth paths that we have investigated, suggesting that the maximum-size ensemble, which includes all models, has near-optimum performance. In addition, we have compared different strategies for combining models in an ensemble. We have found that the algebraic mean-rule outperforms the Borda count voting scheme used by Marbach et al. (2012). This is presumably a consequence of the reduction in information loss effected by commuting the order of the operations *ranking* and *averaging*.

An application of our model ensemble to metabolomic and transcriptomic time series from various mutagenesis plants grown in different light-dark cycles has predicted several statistically significant interactions between circadian clock genes and metabolites in Arabidopsis. This provides independent statistical evidence that the regulation of metabolism by the circadian clock is not uni-directional, but that there is a statistically significant feedback mechanism aiming from metabolism back to the circadian clock. The present work suggests new hypotheses for specific forms of bi-directional interactions, which are plausible in light of our current biological understanding.

Acknowledgements: The work described in the present article is part of the TiMet project on linking the circadian clock to metabolism in plants. TiMet is a collaborative project (Grant Agreement 245143) funded by the European Commission FP7, in response to call FP7-KBBE-2009-3. A.A. is supported by the TiMet project. Parts of the work were

done while M.G. was supported by the German Research Foundation (DFG), research grant GR3853/1-1. We would like to thank Catherine Higham for helpful discussions.

Appendix

Marginal likelihood for time delays

In the hierarchical Bayesian regression (HBR) model, described in Subsection 2.5 of Aderhold et al. (2014), the target observations $\mathbf{y}_g$ are modelled independently for each target g ($g = 1, \ldots, G$):

$$\mathbf{y}_g|(\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \mathbf{w}_g, \boldsymbol{\pi}_g) \sim \mathcal{N}(\mathbf{X}_{\boldsymbol{\pi}_g}^T \mathbf{w}_g, \sigma_g^2 \mathbf{I}) \tag{13}$$

where $\mathbf{w}_g$ are the regression parameter vectors, determined by the sets of regulators (covariates) $\boldsymbol{\pi}_g$, $\mathbf{X}_{\boldsymbol{\pi}_g}$ is the regressor matrix implied by the regulator set $\boldsymbol{\pi}_g$, and σ_g^2 is the node-specific noise variance parameter. Gaussian priors are imposed on the regression parameter vectors:

$$\mathbf{w}_g|(\sigma_g^2, \delta_g, \boldsymbol{\pi}_g) \sim \mathcal{N}(\mathbf{0}, \delta_g \sigma_g^2 \mathbf{I}) \tag{14}$$

where δ_g is the target-specific SNR-hyperparameter. The noise variance parameters σ_g^2 and the SNR-hyperparameters δ_g are assumed to be inverse Gamma distributed with fixed (hyper-)hyperparameters ($g = 1, \ldots, G$). A more detailed model description can be found in subsection 2.5 of Aderhold et al. (2014). Here, we introduce a new 'time lag' parameter $\tau \in \{0, \ldots, \tau_{MAX}\}$, which indicates the time lag between the target observations $\mathbf{y}_g$ and the observations of the regulators in $\boldsymbol{\pi}_g$. The time lag parameter τ describes how to shift the observations of the regulators in eq. (4), i.e. τ describes how to build the regressor matrices $\mathbf{X}_{\boldsymbol{\pi}_g}$ from the data $\mathcal{D}$. Although $\mathbf{X}_{\boldsymbol{\pi}_g}$ depends on τ, we do not make this explicit in our notation.

As the targets $g = 1, \ldots, G$ are modelled independently and the overall network structure $\mathcal{M}$ is determined by the individual regulator sets, symbolically $\mathcal{M} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_G)$, the joint marginal likelihood has a modular form:

$$p(\mathcal{D}|\tau) = \prod_{g=1}^{G} \int_{\delta_g} \sum_{\boldsymbol{\pi}_g} \left(\int_{\mathbf{w}_g} \int_{\sigma_g^2} p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \mathbf{w}_g, \boldsymbol{\pi}_g, \delta_g, \tau) p(\sigma_g^2) p(\mathbf{w}_g|\sigma_g^2, \delta_g, \boldsymbol{\pi}_g) d\sigma_g^2 d\mathbf{w}_g \right) p(\boldsymbol{\pi}_g) p(\delta_g) d\delta_g \tag{15}$$

For a given regulator set $\boldsymbol{\pi}_g$ and a fixed SNR-hyperparameter δ_g marginalizing the HBR likelihood over the regression parameters and the noise variances:

$$p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\pi}_g, \delta_g, \tau) = \int_{\sigma_g^2} \left(\int_{\mathbf{w}_g} p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \mathbf{w}_g, \boldsymbol{\pi}_g, \delta_g, \tau) p(\mathbf{w}_g|\sigma_g^2, \delta_g, \boldsymbol{\pi}_g) d\mathbf{w}_g \right) p(\sigma_g^2) d\sigma_g^2 \tag{16}$$

yields a closed-form solution, see eq (15) in Aderhold et al. (2014).[6] It then follows from eqns. (15-16):

$$p(\mathcal{D}|\tau) = \prod_{g=1}^{G} \int_{\delta_g} \left(\sum_{\boldsymbol{\pi}_g} p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\pi}_g, \delta_g, \tau) p(\boldsymbol{\pi}_g) \right) p(\delta_g) d\delta_g \tag{17}$$

[6] Recalling the notation from subsection 3.7, we have: $\boldsymbol{\theta}_g = (\sigma_g^2, \mathbf{w}_g)$.

The prior, $p(\boldsymbol{\pi}_g)$, on the regulator sets, $\boldsymbol{\pi}_g$, in the HBR model is assumed to be a uniform distribution subject to a constraint on the maximal cardinality $\mathcal{F}$ of the number of regulators (see Grzegorczyk and Husmeier (2012) and Aderhold et al. (2014)). We thus obtain:

$$p(\mathcal{D}|\tau) = \prod_{g=1}^{G} \int_{\delta_g} \left(\frac{1}{T_g} \sum_{\pi_g : |\pi_g| \leq \mathcal{F}} p(\mathbf{y}_g | \mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\pi}_g, \delta_g, \tau) \right) p(\delta_g) d\delta_g \tag{18}$$

where T_g is the number of all valid regulator sets. Hence, if there are N potential regulators for a target g, then the number of valid parent sets T_g is given by:

$$T_g = \sum_{f=0}^{\mathcal{F}} \binom{N}{f} = O(N^{\mathcal{F}}) \tag{19}$$

so that T_g grows polynomially with the power of $\mathcal{F}$.[7] If the inner sums can be computed by full-enumeration, the marginal likelihood can be estimated by repeatedly sampling δ_g $(g = 1, \ldots, G)$ from their inverse Gamma prior distributions and computing the following Monte-Carlo estimator:

$$\widehat{p(\mathcal{D}|\tau)} = \prod_{g=1}^{G} \left(\frac{1}{I} \sum_{i=1}^{I} \left(\frac{1}{T_g} \sum_{\pi_g : |\pi_g| \leq \mathcal{F}} p(\mathbf{y}_g | \mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\pi}_g, \delta_g^{(i)}, \tau) \right) \right) \tag{20}$$

where $\delta_g^{(i)}$ $(g = 1, ...G;\ i = 1, \ldots, I)$ is the i-th sample drawn for target g from an inverse Gamma prior.[8] Alternatively, and in particular if the inner sums in eq. (18) are not computationally feasible, one can resort to standard numerical procedures based on MCMC, like Chib's method (Chib and Jeliazkov, 2001).

To get an idea of the approximation error of our marginal likelihood estimator, we consider J independent Monte-Carlo estimators $\hat{\psi}_{\tau,1}, \ldots, \hat{\psi}_{\tau,J}$ of size I, each computed with eq. (20). For each lag τ we can then consider the distributional form of these estimators to get an impression of (i.e. to bound) the approximation error. For the HBR model we have provided the technical details above, and we note that the marginal likelihoods of the HBR-nl and the HBR-light model can also be approximated along these lines. As we found that the marginal likelihoods of these three models show the same trends and peak at the same lag (a time shift of $\tau = 2h$, referring to one single data point shift), we show only the results for the HBR model in Figure 8.

In our application the data set $\mathcal{D}$ consists of a set of individual time series. When the network interactions are modelled subject to a time lag corresponding to τ data points,

[7] In our application, described in Subsection 4.2, we have 10 genes and 11 metabolites. For each target g we enforce the self-feedback loop, $g \to g$, to take the degradation processes into account, before we infer target-specific regulator sets with cardinalities up to $\mathcal{F} = 3$ from the remaining $N = 20$ variables (metabolites and genes). From eq. (19) we obtain that there are $T_g = 1562$ valid regulator sets $\boldsymbol{\pi}_g$ for each target g.

[8] Recalling the notation from subsection 3.7, we have: $\boldsymbol{\phi}_g = \delta_g$.

then the first τ target observations have to be withdrawn from each time series.[9] For a fair comparison among different time lags τ we first choose a maximal lag τ_{MAX}, and we withdraw the first τ_{MAX} observations from all time series. Subsequently, the remaining target observations $\mathbf{y}_g$ $(g = 1, \ldots, G)$ can be used for all lags $\tau \in \{0, \ldots, \tau_{MAX}\}$, i.e. this approach ensures that the target observations do not differ from τ to τ and that exactly the same target observations have to be modelled for *all* lags τ.[10]

Detailed simulation results

For our performance evaluation on the simulated data described in Section 4.1, we were running hundreds of simulations for a variety of different settings, related to the observation status of the molecular components (mRNA only versus mRNAs and proteins), the method for derivative estimation (described in Section 3.1), the regulatory network structure (shown in Figure 1), and the method applied for learning this structure from data (reviewed in Sections 3.2 and 3.3). The results from these studies are shown in Figures 12-14. These results are complex, and patterns are not easily discernible by visual inspection. This has motivated the application of the ANOVA scheme described in Section 3.5.

[9]This is due to the fact that the values of the potential regulators for the first τ target values are not available.

[10]Note that the data set $\mathcal{D}$ depends on τ_{MAX}, as the first τ_{MAX} observations of the original data set have to be withdrawn. We have therefore employed different values of τ_{MAX}, and we found identical trends for $\tau_{MAX} = 3, 4, 5$ (i.e. 6, 8, and 10 hours). In the paper we report the results obtained for $\tau_{MAX} = 4$ (i.e. 8 hours).

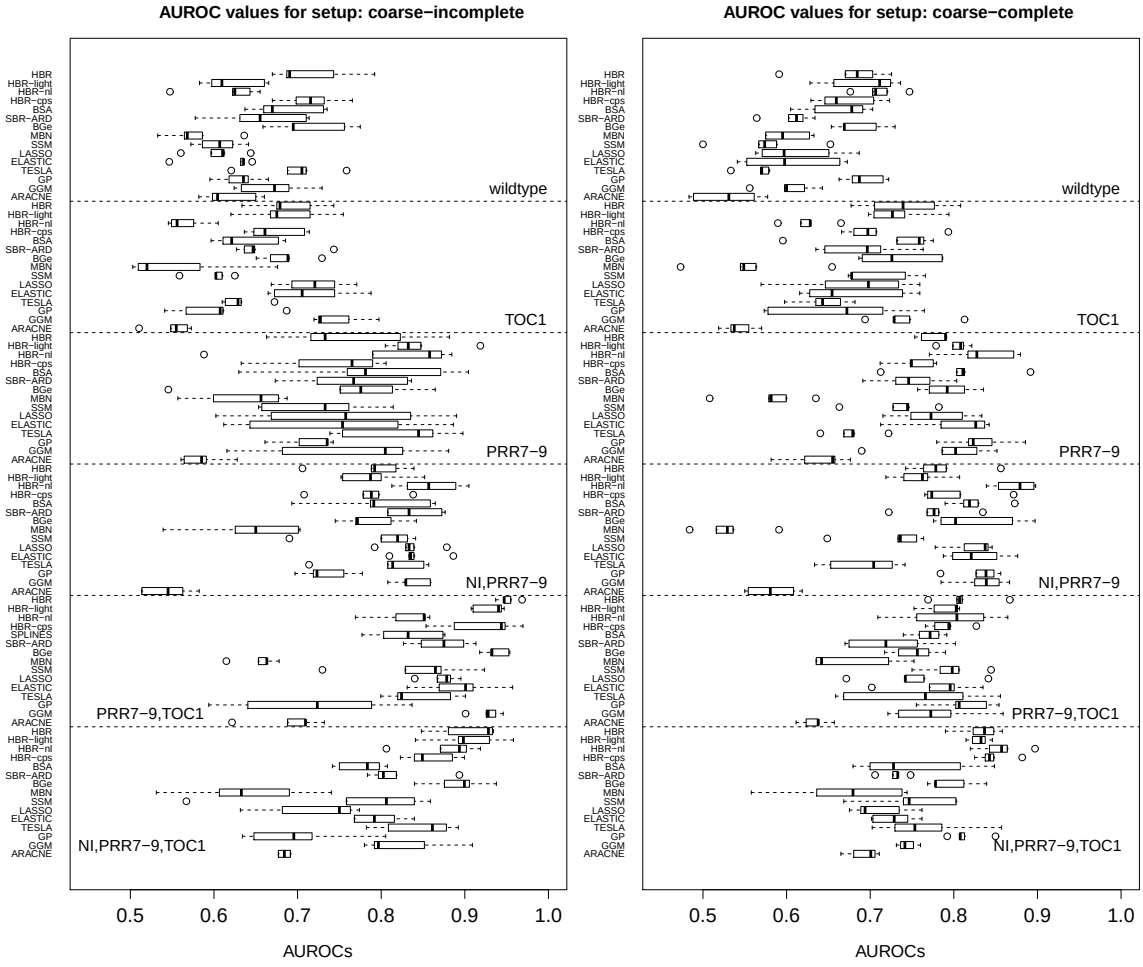

Figure 12: **Detailed AUROC scores: coarse gradient.** The boxplots show a fraction of the results obtained from the simulated data described in Section 4.1. The AUROC scores were obtained from the coarse response gradients with 2-hour intervals. The corresponding results for the fine gradient and the Gaussian process interpolation are displayed in Figures 13-14. *Left panel:* Incomplete data, with mRNA but no protein concentrations. *Right panel:* Complete data that include both protein and mRNA concentrations. Each panel contains six subpanels, representing the six different network topologies shown in Figure 1.

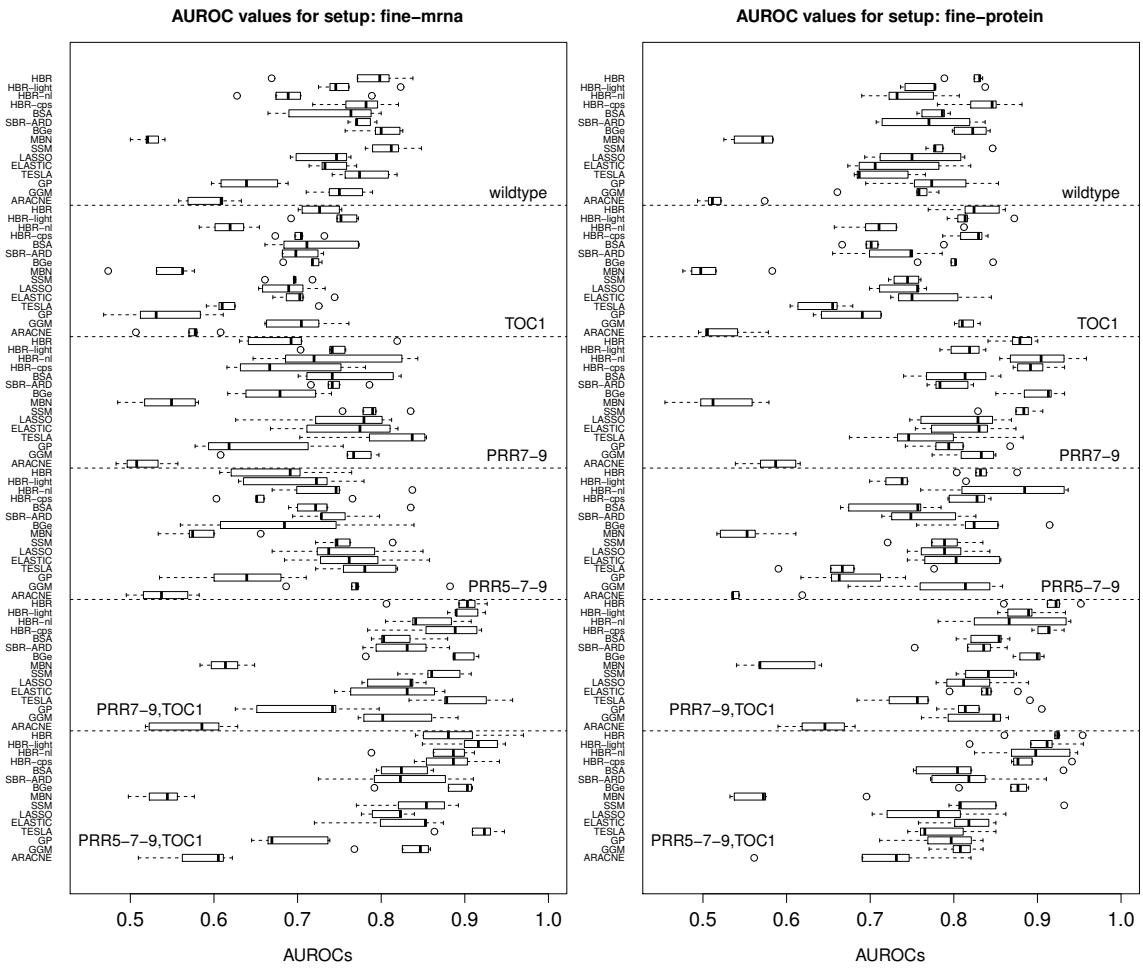

Figure 13: **Detailed AUROC scores: fine gradient.** This figure corresponds to Figure 12, but shows the AUROC scores obtained with the fine gradient (24-minute intervals) rather than the coarse gradient (2-h intervals). See Figure 12 for details.

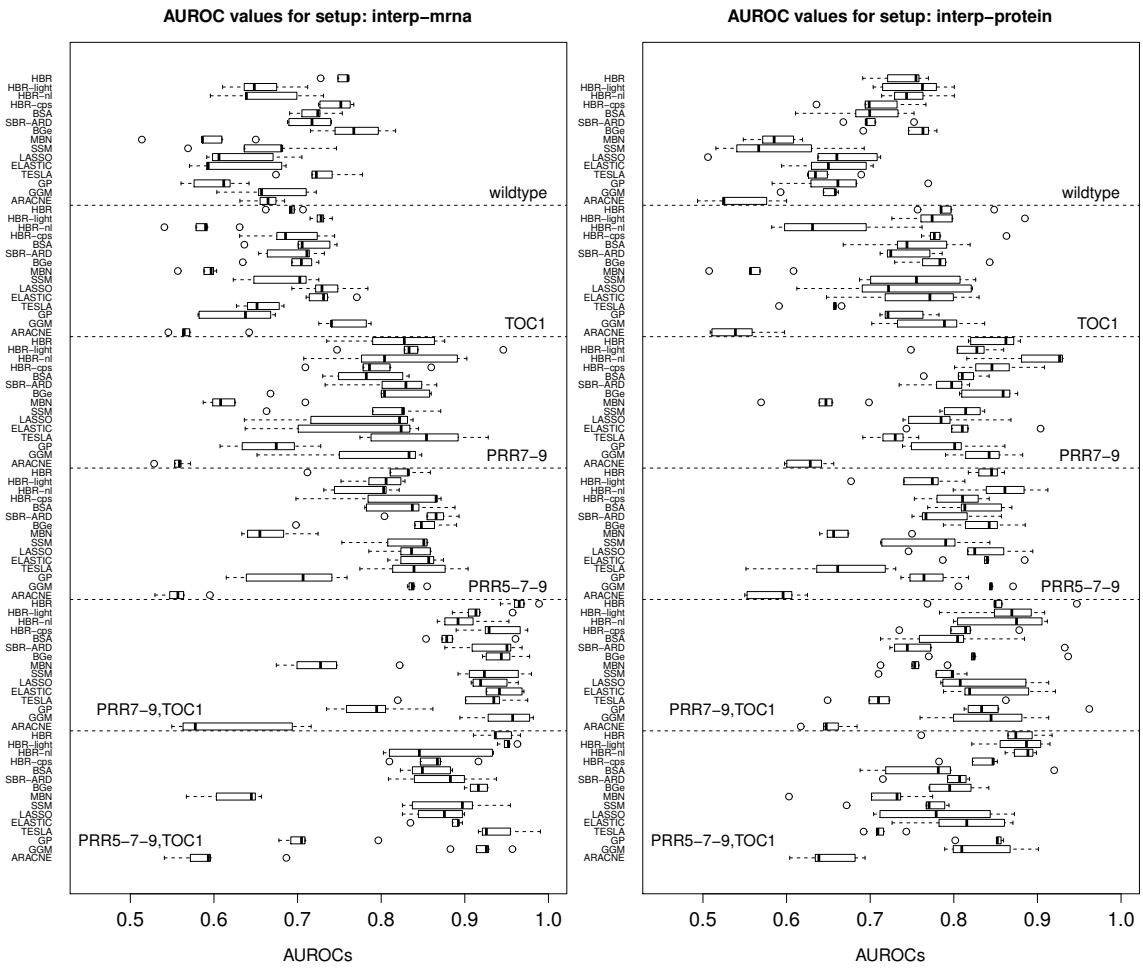

Figure 14: **Detailed AUROC scores: fine gradient.** This figure corresponds to Figure 12, but shows the AUROC scores obtained with the gradient from the Gaussian process interpolation rather than the finite difference method. See Figure 12 for details.

References

Aderhold, A., D. Husmeier, and M. Grzegorczyk (2014): "Statistical inference of regulatory networks for circadian regulation," *Statistical Applications in Genetics and Molecular Biology (SAGMB)*, 13, 227–273.

Ahmed, A. and E. P. Xing (2009): "Recovering time-varying networks of dependencies in social and biological studies," *Proceedings of the National Academy of Sciences*, 106, 11878–11883.

Äijö, T. and H. Lähdesmäki (2009): "Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics," *Bioinformatics*, 25, 2937–2944.

Barenco, M., D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank (2006): "Ranked prediction of p53 targets using hidden variable dynamic modeling," *Genome Biology*, 7, R25.

Beal, M. (2003): *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, UK.

Beal, M., F. Falciani, Z. Ghahramani, C. Rangel, and D. Wild (2005): "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors," *Bioinformatics*, 21, 349–356.

Bläsing, O. E., Y. Gibon, M. Günther, M. Höhne, R. Morcuende, D. Osuna, O. Thimm, B. Usadel, W.-R. Scheible, and M. Stitt (2005): "Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in arabidopsis," *The Plant Cell Online*, 17, 3257–3281.

Brandt, S. (1999): *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, New York, USA: Springer.

Chevaleyre, Y., U. Endriss, J. Lang, and N. Maudet (2007): *A short introduction to computational social choice*, Springer.

Chib, S. and I. Jeliazkov (2001): "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, 96, 270–281.

Ciocchetta, F. and J. Hillston (2009): "Bio-PEPA: A framework for the modelling and analysis of biological systems," *Theoretical Computer Science*, 410, 3065–3084.

Dalchau, N., S. J. Baek, H. M. Briggs, F. C. Robertson, A. N. Dodd, M. J. Gardner, M. A. Stancombe, M. J. Haydon, G.-B. Stan, J. M. Gonçalves, et al. (2011): "The circadian oscillator gene gigantea mediates a long-term response of the arabidopsis thaliana circadian clock to sucrose," *Proceedings of the National Academy of Sciences*, 108, 5104–5109.

Davies, J. and M. Goadrich (2006): "The relationship between Precision-Recall and ROC curves," *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Feugier, F. and A. Satake (2012): "Dynamical feedback between circadian clock and sucrose availability explains adaptive response of starch metabolism to various photoperiods," *Frontiers in Plant Science*, 3.

Flis, A., P. Fernandez, T. Zielinski, R. Sulpice, A. Pokhilko, H. G. McWatters, A. J. Millar, M. Stitt, and K. J. Halliday (2013): "Biological regulation identified by sharing timeseries data outside the 'omics," *Submitted*.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000): "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, 7, 601–620.

Fukushima, A., M. Kusano, N. Nakamichi, M. Kobayashi, N. Hayashi, H. Sakakibara,

T. Mizuno, and K. Saito (2009): "Impact of clock-associated arabidopsis pseudo-response regulators in metabolic coordination," *Proceedings of the National Academy of Sciences*, 106, 7251–7256.

Geigenberger, P. (2011): "Regulation of starch biosynthesis in response to a fluctuating environment," *Plant Physiology*, 155, 1566–1577.

Geiger, D. and D. Heckerman (1994): "Learning gaussian networks," in *International Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, 235–243.

Gille, S., K. Cheng, M. E. Skinner, A. H. Liepman, C. G. Wilkerson, and M. Pauly (2011): "Deep sequencing of voodoo lily (Amorphophallus konjac): an approach to identify relevant genes involved in the synthesis of the hemicellulose glucomannan," *Planta*, 234, 515–526.

Graf, A., A. Schlereth, M. Stitt, and A. M. Smith (2010): "Circadian control of carbohydrate availability for growth in Arabidopsis plants at night," *Proceedings of the National Academy of Sciences*, 107, 9458–9463.

Grzegorczyk, M. and D. Husmeier (2012): "A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology," *Statistical Applications in Genetics and Molecular Biology (SAGMB)*, 11, article 7.

Grzegorczyk, M. and D. Husmeier (2013): "Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models," *Machine Learning*, 1–50.

Guerriero, M., A. Pokhilko, A. Fernández, K. Halliday, A. Millar, and J. Hillston (2012): "Stochastic properties of the plant circadian clock," *Journal of The Royal Society Interface*, 9, 744–756.

Hanley, J. A. and B. J. McNeil (1982): "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, 143, 29–36.

Hastie, T., R. Tibshirani, and J. J. H. Friedman (2001): *The Elements of Statistical Learning*, volume 1, Springer New York.

Haydon, M. J., O. Mielczarek, F. C. Robertson, K. E. Hubbard, and A. A. Webb (2013): "Photosynthetic entrainment of the arabidopsis thaliana circadian clock," *Nature*.

Herrero, E., E. Kolmos, N. Bujdoso, Y. Yuan, M. Wang, M. C. Berns, H. Uhlworm, G. Coupland, R. Saini, M. Jaskolski, et al. (2012): "EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock," *Plant Cell Online*, 24, 428–443.

Kalaitzis, A. A., A. Honkela, P. Gao, and N. D. Lawrence (2013): *gptk: Gaussian processes tool-kit*, URL `http://CRAN.R-project.org/package=gptk`, R package version 1.06.

Kikis, E. A., R. Khanna, and P. H. Quail (2005): "ELF4 is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components CCA1 and LHY," *The Plant Journal*, 44, 300–313.

Ko, Y., C. Zhai, and S. Rodriguez-Zas (2009): "Inference of gene pathways using mixture Bayesian networks," *BMC Systems Biology*, 3, 54.

Kolmos, E., M. Nowak, M. Werner, K. Fischer, G. Schwarz, S. Mathews, H. Schoof, F. Nagy, J. M. Bujnicki, and S. J. Davis (2009): "Integrating ELF4 into the circadian system through combined structural and functional studies," *HFSP Journal*, 3, 350–366.

Kuncheva, L. I. (2004): *Combining pattern classifiers: methods and algorithms*, John Wiley & Sons.

Lawrence, N. D., M. Girolami, M. Rattray, and G. Sanguinetti (2010): *Learning and inference in computational systems biology*, MIT Press Cambridge.

Locke, J. C. W., L. Kozma-Bognár, P. D. Gould, B. Fehér, E. Kevei, F. Nagy, M. S. Turner, A. Hall, and A. J. Millar (2006): "Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*," *Molecular Systems Biology*, 2.

Marbach, D., J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky, et al. (2012): "Wisdom of crowds for robust gene network inference," *Nature Methods*, 9, 796–804.

Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano (2006): "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC Bioinformatics*, 7.

Morrissey, E. R., M. A. Juárez, K. J. Denby, and N. J. Burroughs (2011): "Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression," *Biostatistics*, 12, 682–694.

Pokhilko, A., A. Fernández, K. Edwards, M. Southern, K. Halliday, and A. Millar (2012): "The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops," *Molecular Systems Biology*, 8, 574.

Pokhilko, A., S. Hodge, K. Stratford, K. Knox, K. Edwards, A. Thomson, T. Mizuno, and A. Millar (2010): "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," *Molecular Systems Biology*, 6.

Pokhilko, A., P. Mas, A. J. Millar, et al. (2013): "Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs," *BMC Systems Biology*, 7, 1–12.

Polikar, R. (2006): "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, 6, 21–45.

Rasmussen, C. E. (1996): *Evaluation of Gaussian processes and other methods for non-linear regression*, Ph.D. thesis, Citeseer.

Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani (1996): "The DELVE manual," URL `http://www.cs.toronto.edu/~delve`.

Rogers, S. and M. Girolami (2005): "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, 21, 3131–3137.

Schäfer, J. and K. Strimmer (2005): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genomics and Molecular Biology*, 4.

Solak, E., R. Murray-Smith, W. E. Leithead, D. J. Leith, and C. E. Rasmussen (2002): "Derivative observations in Gaussian process models of dynamic systems," *Advances in Neural Information Processing Systems*.

Tibshirani, R. (1995): "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wang, R., K. Guegler, S. T. LaBrie, and N. M. Crawford (2000): "Genomic analysis of a nutrient response in arabidopsis reveals diverse expression patterns and novel metabolic and potential regulatory genes induced by nitrate," *The Plant Cell Online*, 12, 1491–1509.