Supplementary material

Supplementary Methods

In addition to ROC curves, we also analysed precision-recall (PR) curves for the cooperative as well as the inhibitory case. Due to the very low number of true positives, the individual PR curves often have atypical shapes (suppl. Fig. 1 b-d). Consequently, it is difficult to generate meaningful average curves from the many experiments on synthetic data which we performed to evaluate MONA.

In order to overcome the problem of atypical PR curves, we have repeated the evaluation on MONA and MGSA on synthetic data allowing for a larger number of terms being active. We did this for the cooperative as well as the inhibitory case, for average noise and microRNA influence. Therefore we sampled up to 30 independent[1] terms. It is worth noting that 30 independent active terms is more than what one may expect in a real world dataset; that is why the resulting PR and ROC curves should be interpreted with care. In order to illustrate the expected statistics across a large number of runs we generated a unified PR curve. This can be done by combining all individual runs in one large artificial dataset. Alternatively, the individual PR curves can be averaged.

Both methods have advantages and important drawbacks: First,the combined PR curve is very sensitive to outliers where the parameter p (prior probability for a term being active) is estimated too high, which leads to a high baseline-probability. Only one such case can result in a considerable degradation of the PR curve as all terms will be considered false positives for relatively low thresholds. Second, it is consistent with the standard procedure in the machine learning literature to compare the performance of two classifiers: here, significant differences are established by the pairwise comparison of an algorithm on a number of individual datasets [1, 2]. This is also how we determine the p-values in the main paper text. A natural visualization of this is either showing all individual PR curves or the median/average across all individual curves.

Averaging individual PR curves solves many of these drawbacks. However, information on the consistency of the scores is lost. That is why we illustrate the variation of the scores between runs using kernel density plots and show PR curves using the two aggregation techniques: via combination into one artificial dataset as well as via direct averaging of the individual PR curves.

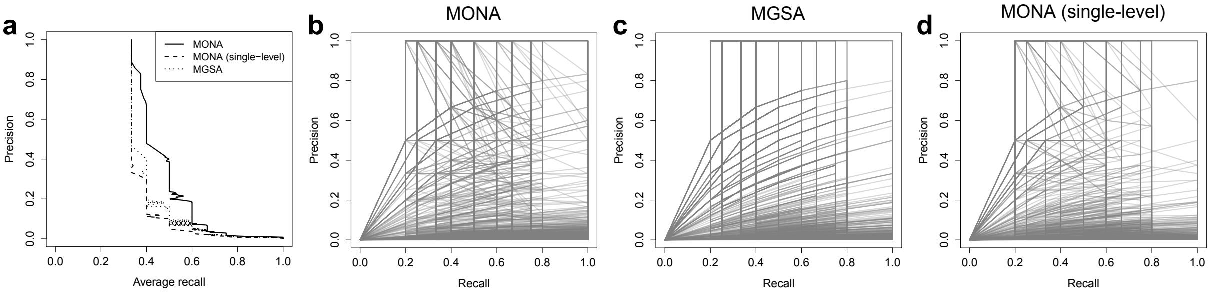
Supplementary Results

In supplementary Figure 2 we show the unified PR curves generated via averaging of the individual PR curves (from suppl. Figure 3) and via combination in one artificial dataset. It can be seen that the performance for MONA is superior to the single-level approaches MONA single and MGSA. When the PR curve is generated by combining all 500 individual datasets into a single one, it can be seen that MGSA performs better than MONA for very small levels of recall. As discussed in the main text, this is a consequence of systematic differences between the MCMC sampling approach and EP which results in a broad distribution of the baseline probability and in turn in a degradation of the PR curve (comp. supplementary Methods). This is also illustrated in suppl. Fig. 3 (d) where the kernel-density estimate of the mean posterior probability of a term being on is shown. As discussed above, it can be seen that while MGSA yields a sharp distribution of the baseline-probabilities, MONA infers a broader distribution of baseline-probabilities including some few outliers. This results in the degradation of the combined PR curve (suppl. Fig. 2 (b)). It is important to note that the ranks are inferred correctly (as illustrated in figure 2 (a)), and the inconsistency of the scores is rather small in absolute numbers (e.g. 0.02 instead of 0.003 in 'extreme' cases). Hence, for practical applications this only plays a minor role.

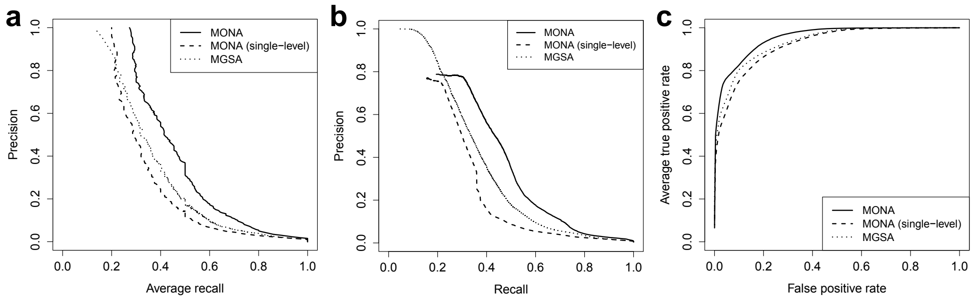
[1] Due to the large number of terms (resulting from the large assignment matrix) in the inhibitory case, we do not sample terms with a very large number of children in the GO hierarchy (largest 10%) in order to facilitate efficient sampling.

The PR curves also illustrate that MGSA tends to perform better than MONA single-level, especially for low levels of recall. As discussed in the main paper text, this is also due to the difference in inference algorithms (exact MCMC vs approximate EP).

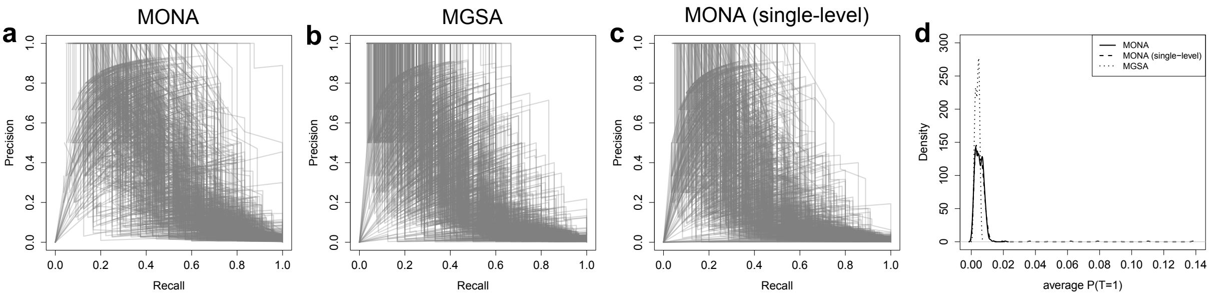
As for the cooperative model, the same trends can be seen for the inhibitory model (suppl. Fig. 5-6): MGSA tends to have a higher precision than MONA single-level, especially for small recall levels. Furthermore, due to the EP approximation the distribution of baseline probabilities is broader for MONA than for MGSA, resulting in a degradation of the joint PR curve for very small recalls. However, this does not affect the ranks and the overall performance of MONA is superior to MGSA, also in terms of precision-recall.



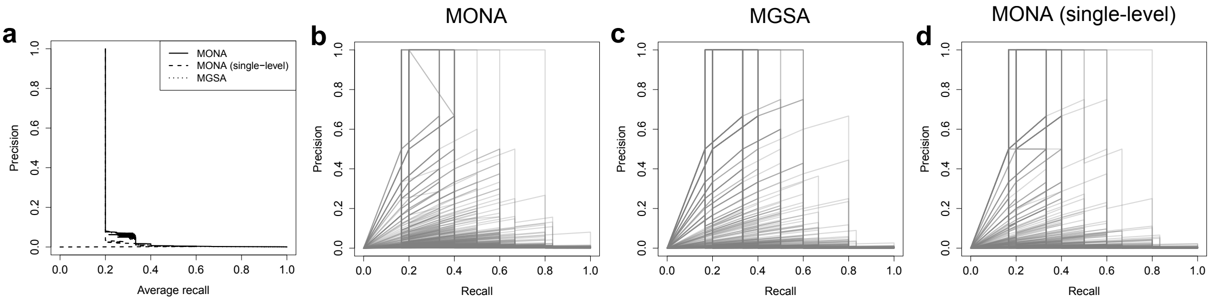
Supplementary Figure 1: Precision-recall (PR) curves of MONA and MGSA on the synthetic data for the cooperative model as described in the main text (medium α and β).



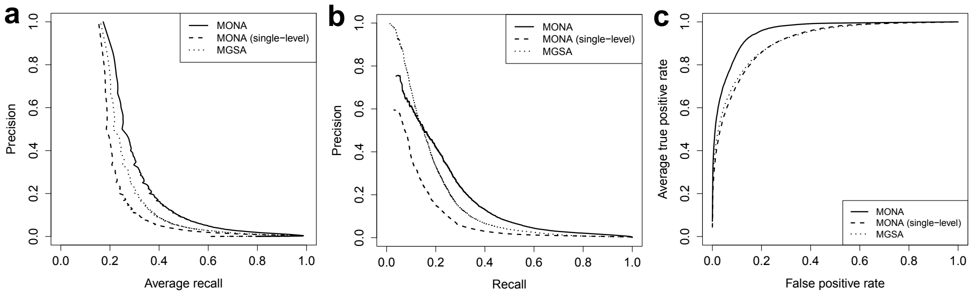
Supplementary Figure 2: Unified PR curves generated by averaging individual PR curves (a) as well as by generating one joint curve (b) for the cooperative model. In (c) the averaged ROC curves are shown. All curves are based on the synthetic dataset with up to 30 active terms.



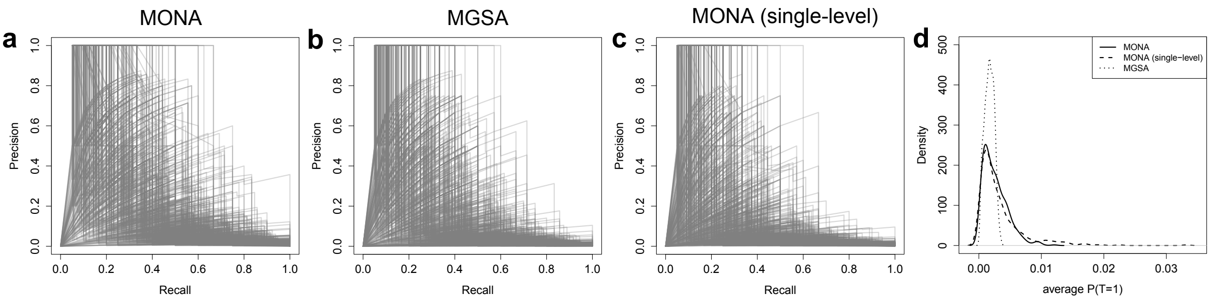
Supplementary Figure 3: Individual PR curves with up to 30 active terms (a-c) for the cooperative model and kernel density plots for the mean posterior probabilities of a term being on as a measure for the consistency of the scores (d).



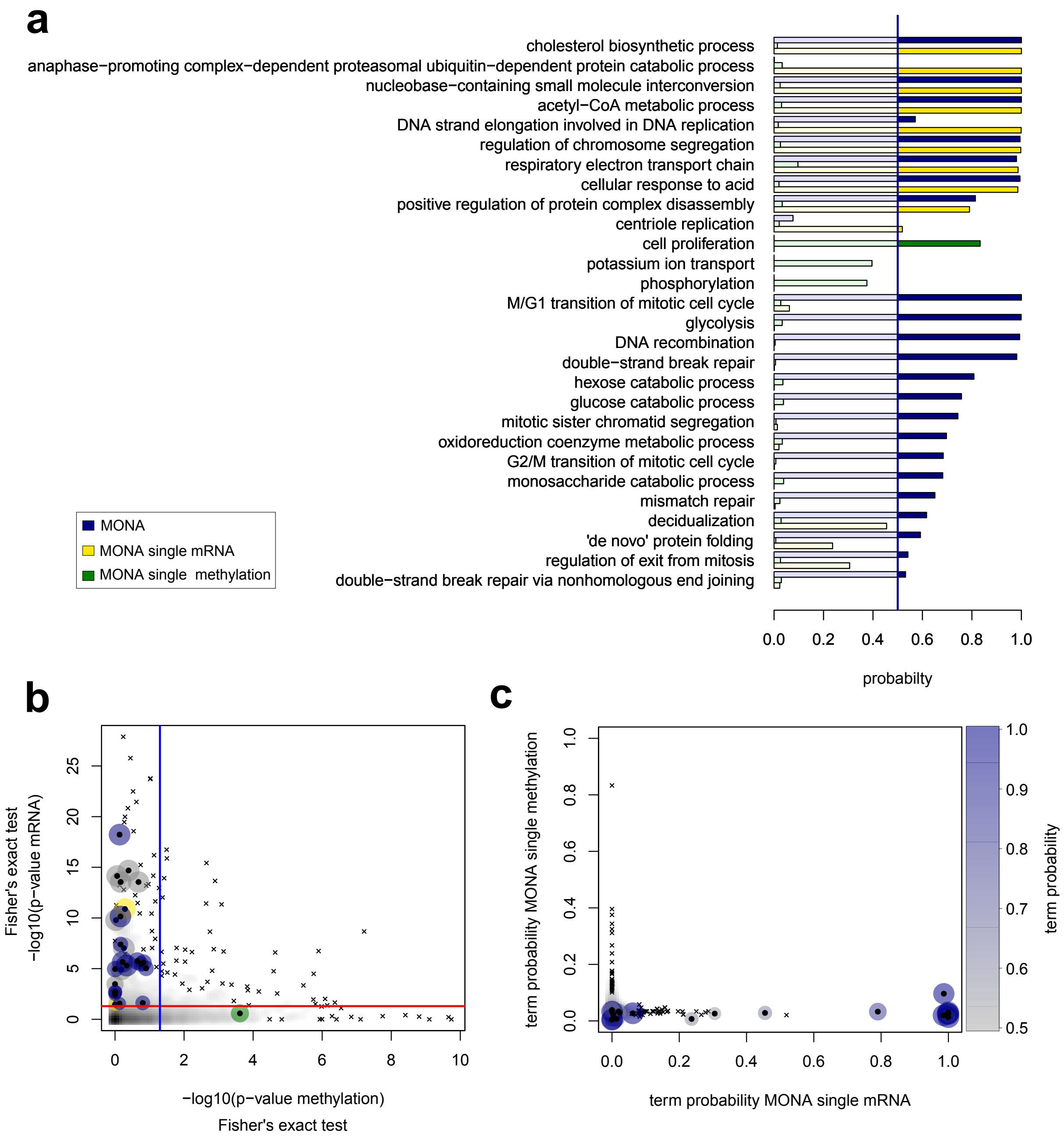
Supplementary Figure 4: Precision-recall (PR) curves of MONA and MGSA on the synthetic data for the inhibitory model as described in the main text (medium α and β, medium miRNA influence).



Supplementary Figure 5: Unified PR curves generated by averaging individual PR curves (a) as well as by generating one joint curve (b) for the inhibitory model. In (c) the averaged ROC curves are shown.



Supplementary Figure 6: Individual PR curves with up to 30 active terms (a-c) for the inhibitory model and kernel density plots for the mean posterior probabilities of a term being on as a measure for the consistency of the scores (d).



Supplementary Figure 7: Comparison of results obtained from MONA, single species MONA on mRNA level and single species MONA on methylation level. **(a)** Term probabilities from MONA (blue), single species MONA on mRNA level (yellow) and single species MONA on methylation level (green). **(b)** For each GO term, p-values of Fisher's exact test on mRNA and methylation level are plotted against each other. Active terms resulting from MONA are marked as blue dots, from single species MONA on mRNA level as yellow dots and from single species MONA on methylation level as green dots. The grey dots represent terms that were identified by both, MONA and single species MONA on mRNA level. The size of the dots represents the term probability. **(c)** Term probabilities from single species MONA on mRNA level and single species MONA on methylation level are plotted against each other. Active terms resulting from MONA are marked as dots and are colour- and size-coded by its respective MONA term probability.

References

[1] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[2] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.