Supplementary Figures

Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models

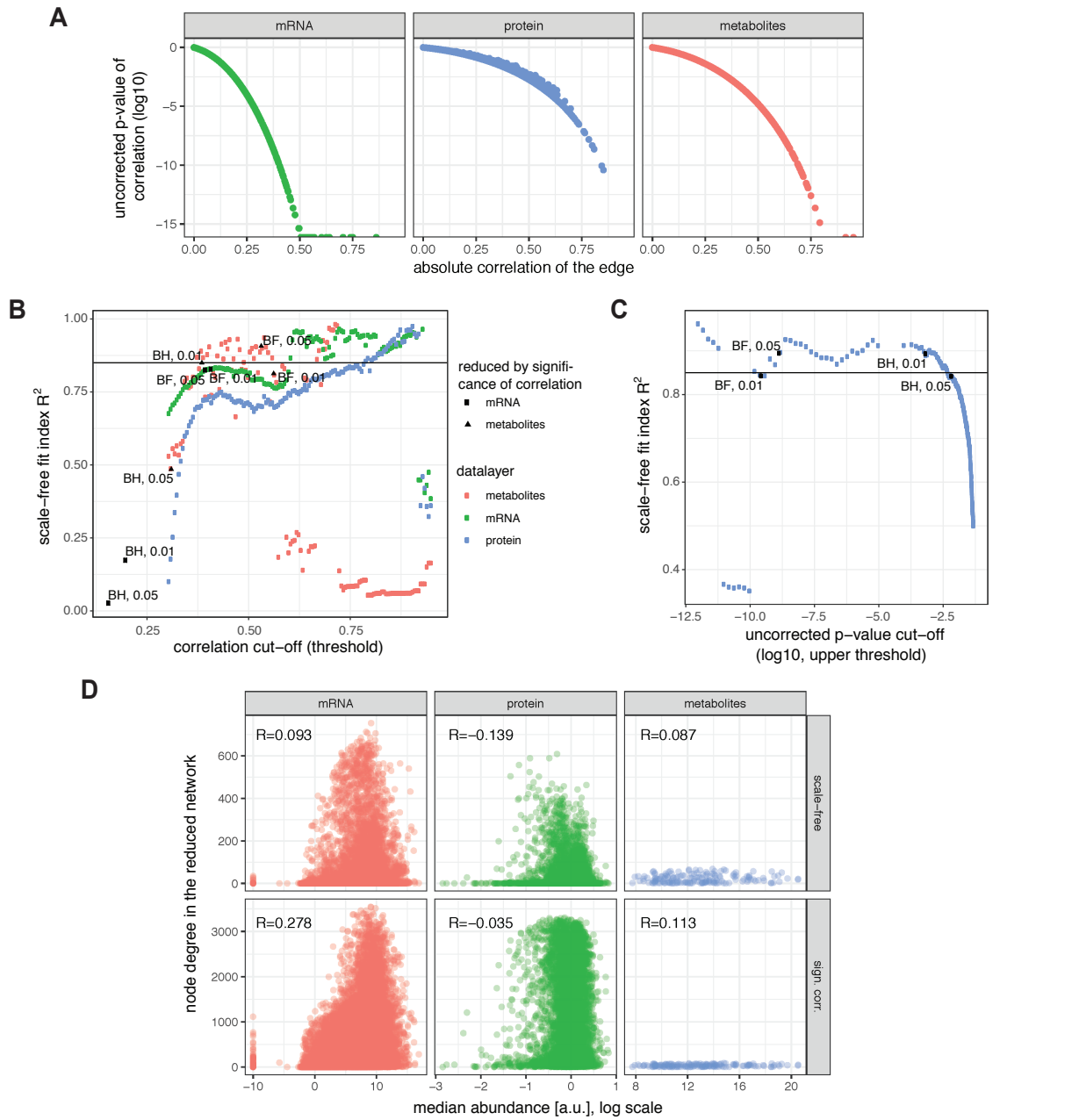
Katharina Baum[1,2], Jagath C. Rajapakse[3,*], Francisco Azuaje[1,*]

[1] Bioinformatics and Modelling, Luxembourg Institute of Health, Luxembourg, 1A-B rue Thomas Edison, 1445 Strassen, Luxembourg

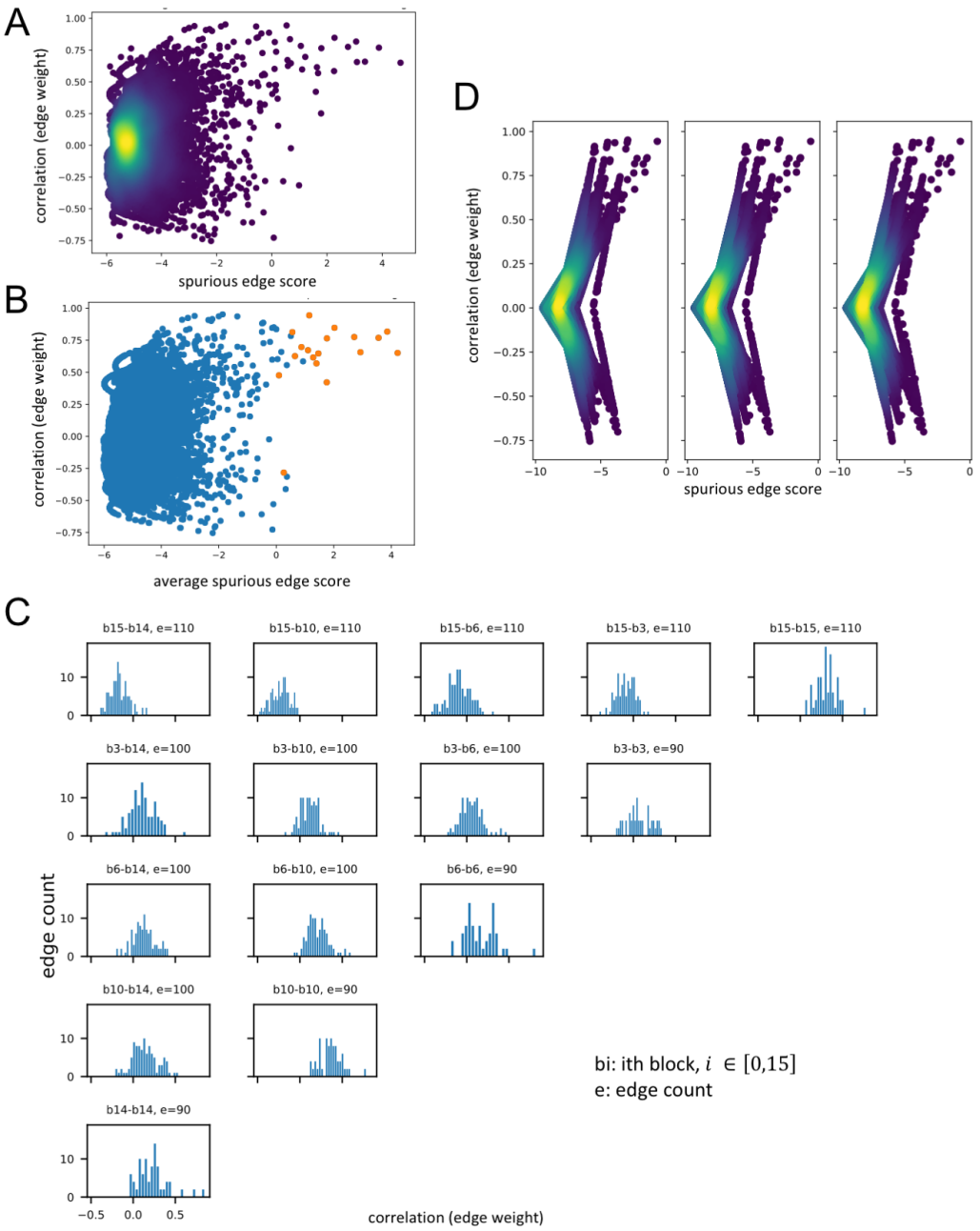
[2] Mathematical Modelling of Cellular Processes, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125 Berlin, Germany

[3] School of Computer Science and Engineering, Nanyang Technological University, N4-2a06, 50 Nanyang Avenue, Singapore 639798

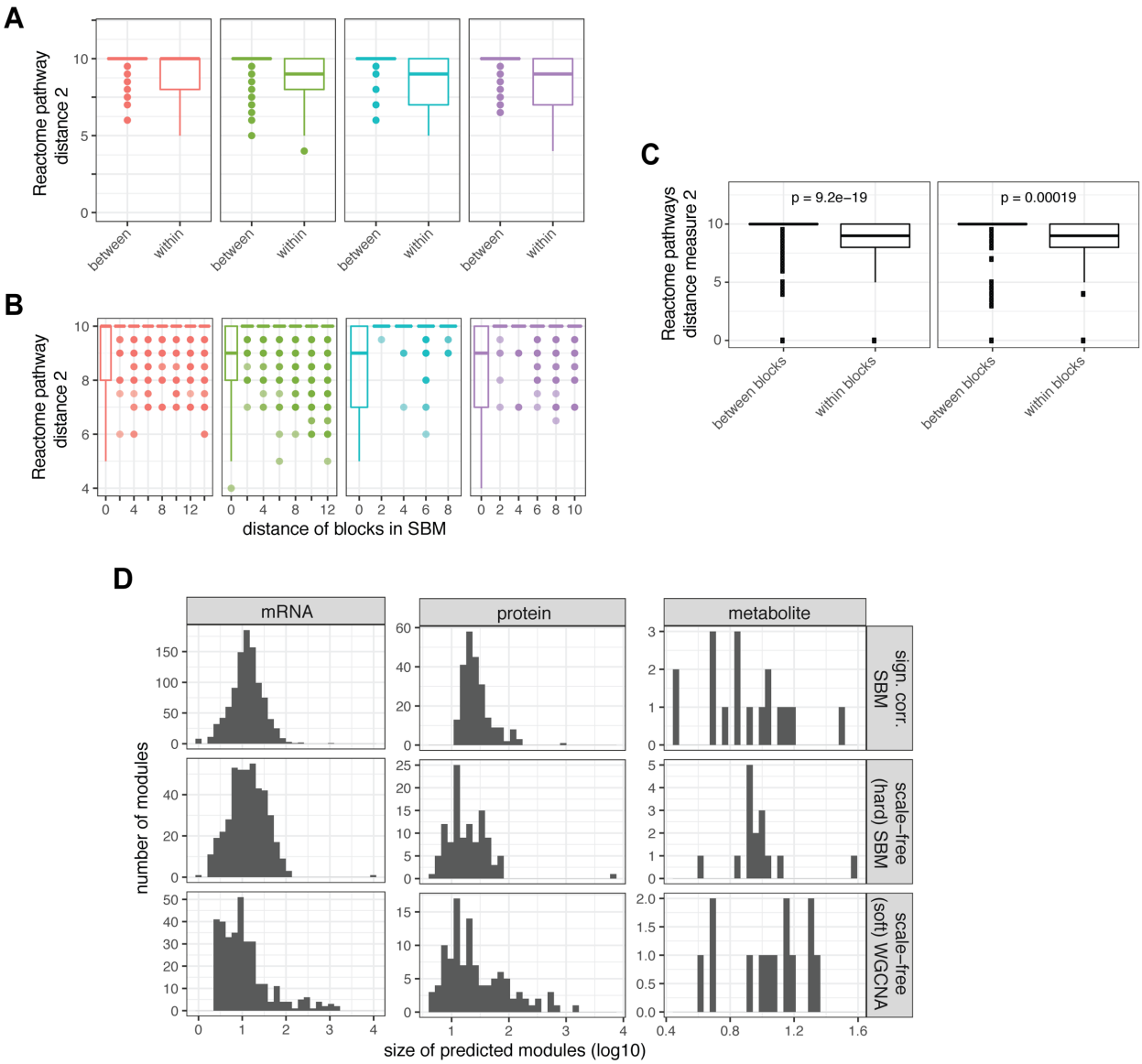
[*] Corresponding authors email: asjagath@ntu.edu.sg, francisco.azuaje@lih.lu



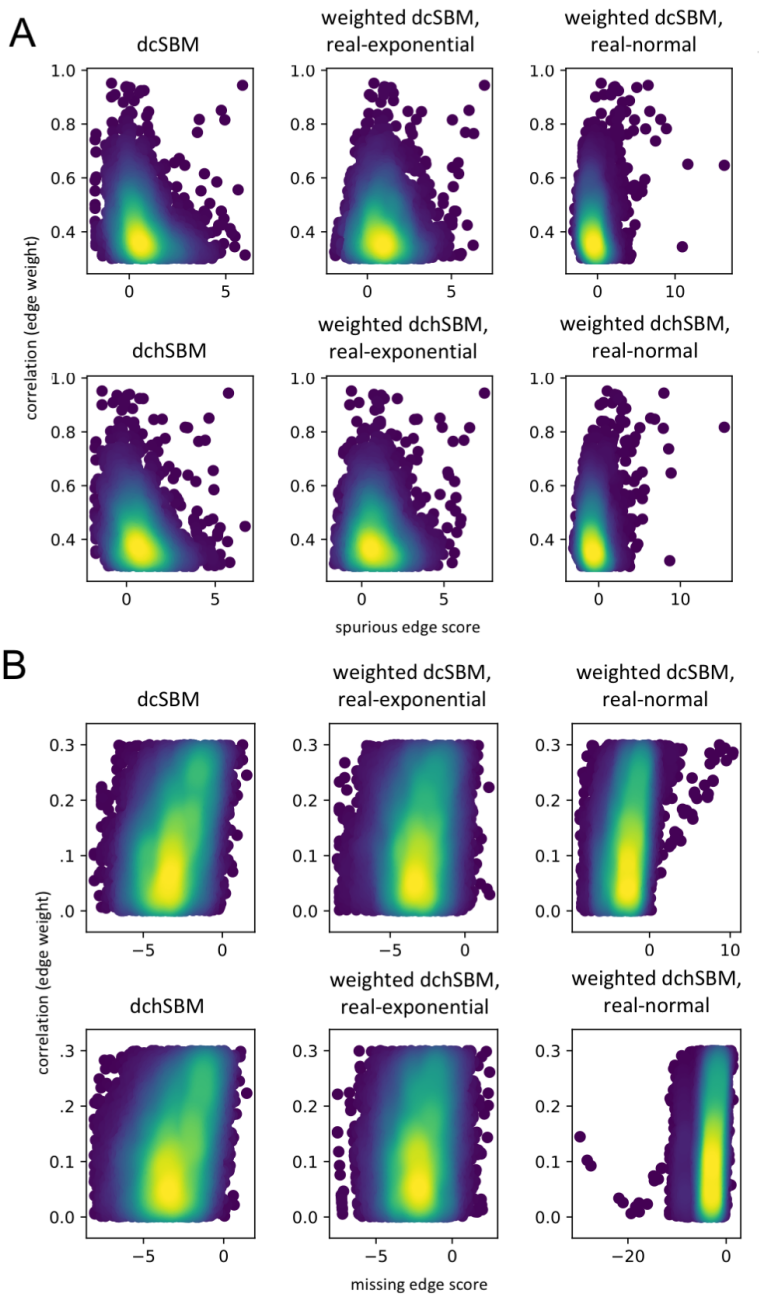
Supplementary Figure S1. The relationship between network reduction by significance of correlation or by scale-freeness. (A) Uncorrected p-values for the correlation being different from zero vs. absolute correlation value for a random subset of 2000 edges from each of the three networks. In contrast to the protein layer in which sample sizes can differ between edges, metabolite and mRNA p-values depend monotonously on the absolute correlation values. For the protein layer, the correlation value for some interactions will be backed by fewer data points only leading to increased p-value despite the same correlation value. **(B)** Scale-free fit indices R^2 for networks reduced by different correlation thresholds as computed using the WGCNA R package function pickHardThreshold (Langfelder&Horvath, 2008) as in Fig. 1D of the main text together with the scale-free fit indices for the mRNA (black squares) or metabolite (black triangles) networks reduced by different thresholds on significance of correlation: Benjamin-Hochberg (BH) or Bonferroni (BF) multiple testing correction, significance thresholds 0.01 or 0.05. **(C)** Scale-free fit indices R^2 for scale-free fitting by thresholding according to uncorrected p-values of correlation being significantly different from zero. The four significance thresholds (Benjamin-Hochberg (BH) or Bonferroni (BF) multiple testing correction, significance thresholds 0.01 or 0.05) considered in the reduction by significance of correlation are marked by black symbols. Indeed, the finally employed threshold (BH, 0.01) also results in an approximately scale-free network ($R^2 > 0.85$). **(D)** Node degrees (i.e. number of adjacent edges) versus median abundance over all samples for the six reduced networks, and Pearson's correlation (R) between these two quantities. Also nodes with very low average abundance can show a high connectivity within the reduced networks.



Supplementary Figure S2. The weighted SBM seems not appropriate for edge prediction from edge confidence scores for fully connected networks. We fitted the fully connected correlation-based metabolite network (162 nodes, 13041 edges, data from Budczies et al., 2013) to a weighted degree-corrected SBM with prior assumption of a real-normal distribution of edge weights between blocks for 50 random initializations using the Python module graph-tool (Peixoto, 2014). Overall, 16 blocks are predicted for the best fitting SBM. **(A)** Edge weight (correlation) vs. spurious edge confidence scores for the best fitting SBM. Against expectations on edge confidence scores, mainly edges with high weight are predicted as spurious. **(B)** Edge confidence scores averaged over the three best fitting SBMs show similar results as the single best fit. Edges with score > 0 in all three SBMs are highlighted in orange. **(C)** Edge weight distributions between the largest 5 blocks of the SBM from (A). Edges with high weight are outliers in the edge weight distributions and consequently are predicted as spurious. **(D)** Edge confidence scores for the best fitting weighted SBM with a real-exponential distribution as prior assumption for the edge weights. In contrast to the real-normal assumption, only 6 blocks are predicted and different edge confidence scores are obtained, which are visibly dominated by their block association. Again, against expectations, edges with high weights tend to be predicted most spurious. This shows the strong influence of the prior on global SBM characteristics and edge prediction for fully connected weighted networks.



Supplementary Figure S3. Pathway characteristics for alternative distance measure, and block size distributions. (A) For the hierarchy level 1 clustering of each of the four mRNA and protein SBMs, we calculated the average distances between every pair of Reactome pathways between blocks and those within blocks, for distance measure (ii) (see Methods) based on the Reactome hierarchy. The lower the distances the more similar are the pathways. As for distance measure (i) (main Fig. 3B), the pathways associated to one single block (within) are significantly more similar than those associated to different blocks (Welch's t-test p-value < 0.01) suggesting that biological functions are consistent within blocks and distinct between blocks. **(B)** Distance of Reactome pathways (as in A) between blocks (or within blocks, for distance of blocks in SBM being zero) versus the distance of the blocks in the SBM hierarchy for blocks on hierarchy level 1. We do not find evidence that the Reactome hierarchy is reflected in the SBM hierarchy. **(C)** Between-module vs. within-module distances of Reactome terms as in (A) for the clusterings predicted from WGCNA-based module detection (Langfelder&Horvath, 2008). **(D)** Distribution of predicted module (block) sizes for the six networks using SBM (lowest hierarchical level only) or the three networks using network generation and module detection from WGCNA (Langfelder&Horvath, 2008).



Supplementary Figure S4. Edge predictions for a reduced weighted network with planar or hierarchical weighted SBMs. We fitted the correlation-based metabolite network reduced by neglecting edges with |correlation| < 0.3 to the unweighted, degree-corrected planar (dcSBM) or hierarchical SBM (dchSBM) or to weighted SBMs (dcwSBM or dchSBM) using a real-exponential distribution (middle) or a real-normal distribution (right) as prior assumptions for the edge weight distribution in the Python module graph-tool. **(A)** Absolute edge weights (correlations) versus the according SBM-derived spurious edge confidence scores. **(B)** Absolute edge weights (correlations) versus the according SBM-derived missing edge confidence scores. Edge predictions are overall in accordance with expectations on edge relevance given by correlations, i.e., edges with large absolute correlation are preferentially predicted as missing and not as spurious.