# Supplementary Information - Identification of drug mode of action based on gene expression data: Application in Drug Induced Lung Injury

## Investigation of the statistical significance of the ILP predictions

### Comparison with an independent statistical method.

Here, an independent statistical method is applied to validate the DILD network constructed by the ILP algorithm. To this end, the GUIDE algorithm is used [1, 2]. GUIDE is an algorithm that builds a classification and regression tree model to predict the values of one or more response variables $(Y_1, Y_2, ...)$ from the values of the predictor variables $(X_1, X_2, ...)$. It can also produce an importance score for each $X_i$. Classification and regression trees were also shown to predict oral absorption in humans based on predictors of chemical substructures [3].

Here, the drug targets from STITCH were used as predictor variables $(X_i)$ and the differential gene expressions for each drug were used as response variables $(Y_i)$. GUIDE was used to construct regression trees, modeling how drug targets correlate statistically with the differential gene expressions. Since GUIDE is agnostic to the protein connectivity in the PKN, it cannot construct functional mechanistic pathways in similar fashion to the ILP, but can produce scores for the drug targets that represent their importance in predicting the observed gene expression signatures. Drug targets with importance score greater than 1.0 are considered significant and their overlap with the nodes in the DILD network is computed using the hypergeometric cdf.

Of the 4478 drug targets in total present in the PKN, the GUIDE algorithm identified 78 to be predictive, with 71 being present in the optimized network. The ILP algorithm conserved in the solution 1056 drug targets (of the original 4478). Note that not all 1056 drug targets are used to generate signal, many of them are included as regular signaling nodes. Thus, an enrichment score can be calculated as follows: number of hits: 71; out of 78 nodes obtained from GUIDE; number of tries: 1056; total number of targets: 4478. If GUIDE and ILP are orthogonal, the probability of such a result is the same as the probability (2.0630e-37) of randomly drawing 1056 balls from 4478, of which 78 are black and 4400 are white, and getting 71 or more black balls. Therefore the p-value is 2.0630e-37. The significant overlap between the ILP and GUIDE predictions, further establishes the statistical significance of the ILP results.

## Sensitivity analysis with respect to model parameters.

The ILP formulation incorporates two user defined parameters that determine the contribution of the measurement-prediction mismatch (parameter $\alpha$) and the solution size (parameter $\beta$) in the objective function. Minimization the measurement-prediction mismatch implies connecting the drug targets with as many of the over-/under-expressed genes as possible based on the structure of the PKN. However, this leads to the rapid increase of the solution size, which is penalized by the $\beta$ parameter. Thus, the two objectives are conflicting. See also the Methods section. High values of the $\alpha/\beta$ ratio imply the algorithm prioritizes the goodness of fit to the data over the size of the solution. Low values of the $\alpha/\beta$ ratio prioritize the small size of the solution over the goodness of fit. To illustrate how the algorithms performance is affected by the $\alpha$ and $\beta$ parameters, the pathway construction procedure is repeated for 12 values of the $\alpha/\beta$ ratio while monitoring the solution size, the goodness of fit to the data, and model predictions with respect to the consistently up- and down-regulated signaling proteins. Results are shown in the Supplementary Figures 1, 2 and 3.

In Supplementary Figure 1A we demonstrate how the average solution size changes for different values of the $\alpha/\beta$ ratio and in Supplementary Figure 1B the corresponding trends for the average fitness error. We tested 12 $\alpha/\beta$ ratios: 20, 10, 6.67, 5, 4, 3.33, 2.85, 2.5, 2.22, 2, 1.82, and 1.67. A ratio of e.g. 20 implies that up to 20 nodes may be added in the solution for fitting one more measured gene. High values offer the advantage of fitting as many of the expressed genes as possible, but at the cost of including a large number of intermediate nodes, which may lack biological meaning. For example, if a signaling cascade of 20 nodes/interactions is required to fit the expression of a single gene, and given the inherent noise of gene expression data, then there is very little evidence supporting the functionality of this cascade. On the other hand, low values of the $\alpha/\beta$ ratio guarantee the functionality of the included cascades, but do not allow the solution to branch out to a big part of the expressed genes. All the analysis presented in this paper has been performed with $\alpha/\beta$ ratio equal to 5. In the Supplementary Figures 1A and 1B note that the long error bars are attributed to the variability of the drug specific pathways in terms of solution size and differentially expressed genes. Moreover, the residual error of around 45%, even for very high ratios, is attributed to the fact that the transcription regulation of many genes is not known (i.e. there are no TFs known to express them), thus the corresponding genes cannot be incorporated in the solution.

In Supplementary Figures 2 and 3 we show the predicted signaling activity of the consistently up- and down-regulated nodes of Figures 5,6 across all the selected $\alpha/\beta$ ratios. We observe that almost all of the consistently up-/down-regulated nodes demonstrate the same trend for all ratio values, even though these nodes were computed using $\alpha/\beta = 5$, demonstrating the robustness and statistical significance of model predictions. However, a few nodes that cross the $y = 0$ dashed line are observed, such as GSK3B and AKT1 of the up-regulated proteins and NFE2L2, NR4A1 of the down-regulated proteins. This occurs because as the $\alpha/\beta$ ratio changes, genes that were out of reach are now reachable and a node that could be used to connect to over-expressed genes, it is now used to connect to under-expressed genes, indicating uncertainty with respect to its signaling activity.

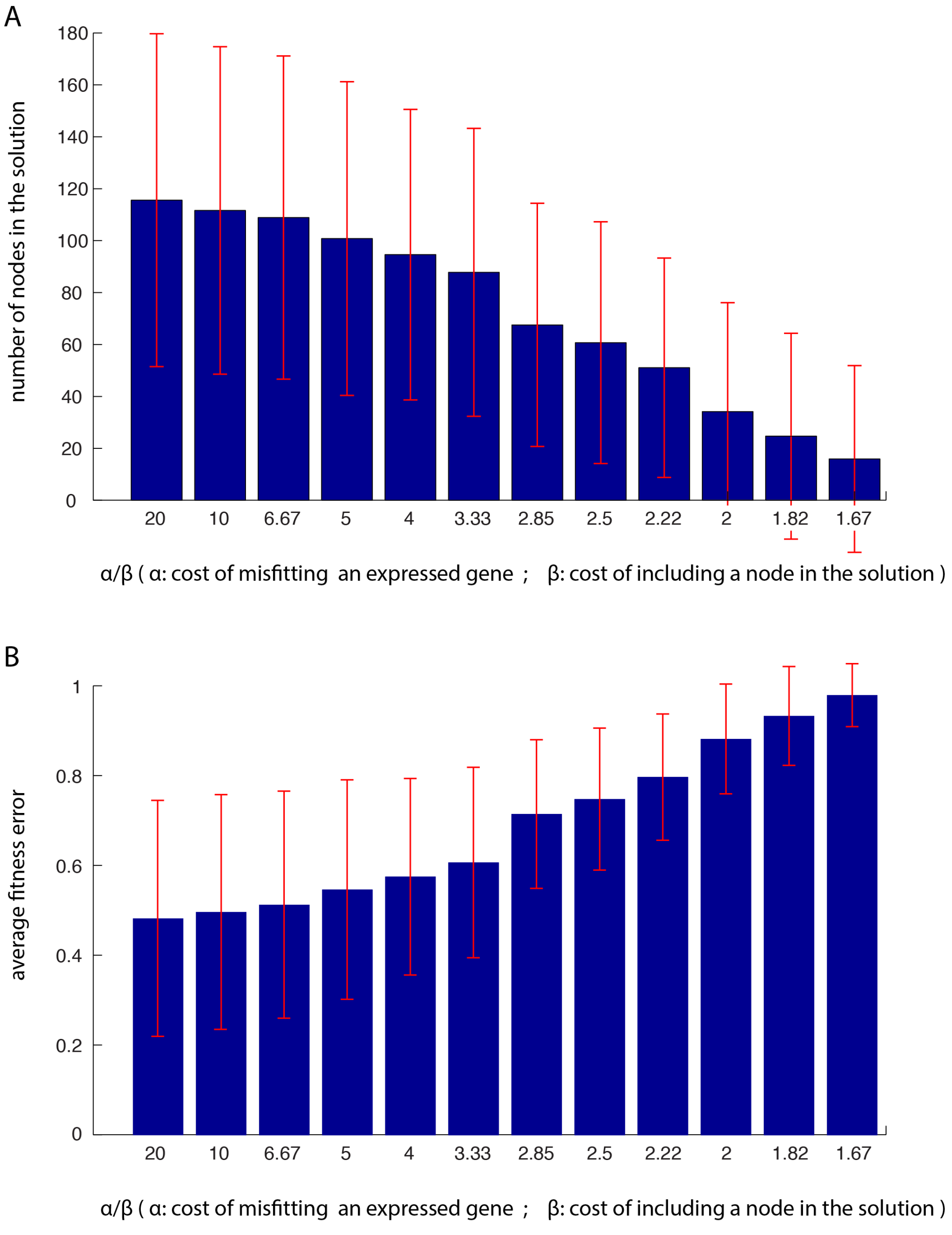

Figure 1: Dependence of (A) solution size and (B) fitness error from model parameters. Lower values of the $\alpha/\beta$ ratio prioritize the solution size over the fitness error, while higher values lead to better fit to the data but at the cost of increased solution size, compromising the biological significance of the resulting topology. The bars refer to average values across all drug specific pathways.

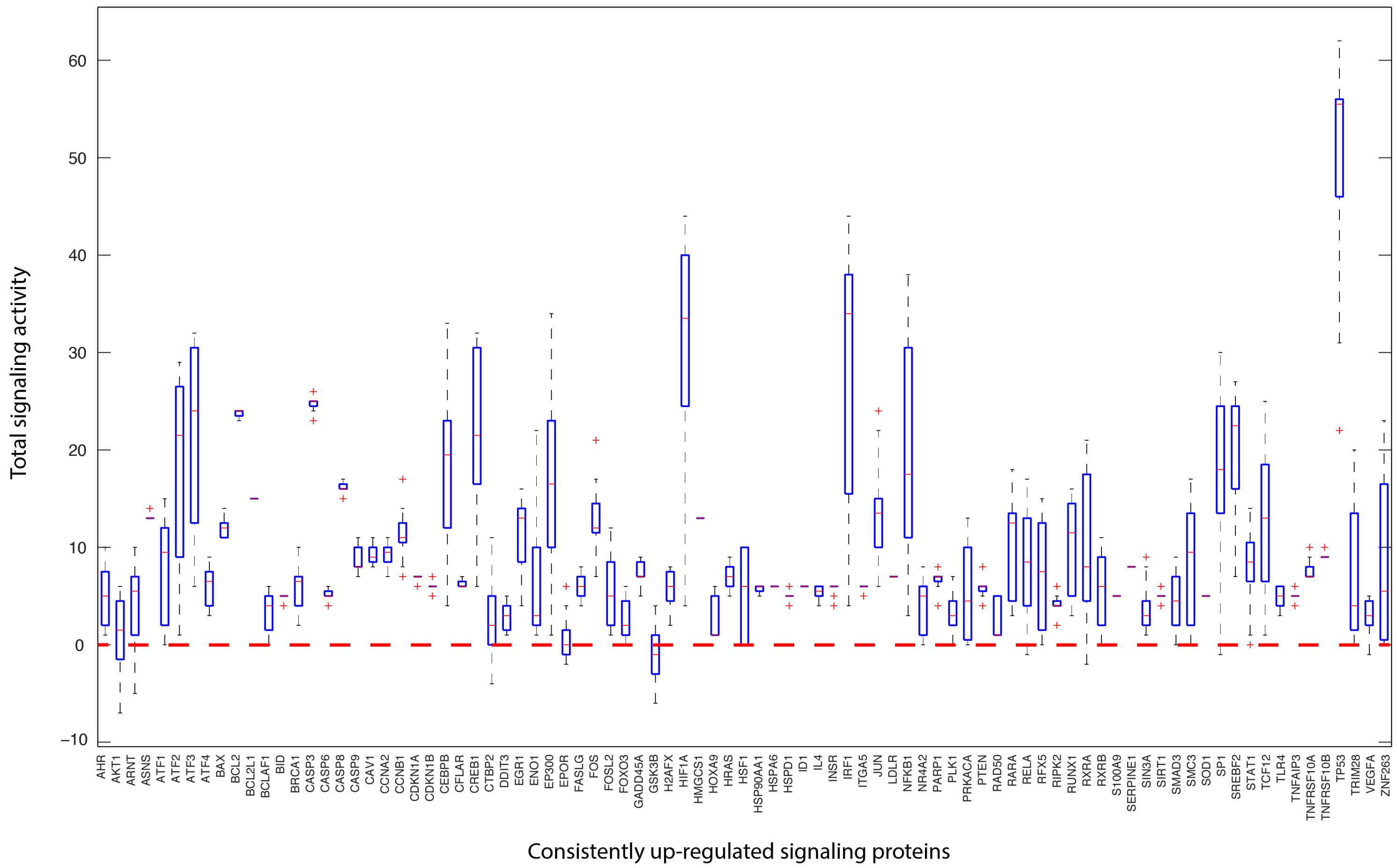

Figure 2: Dependence of the signaling activity of the consistently up-regulated proteins from model parameters. We observe that most of the nodes demonstrate the same trend across all $\alpha/\beta$ ratio values.

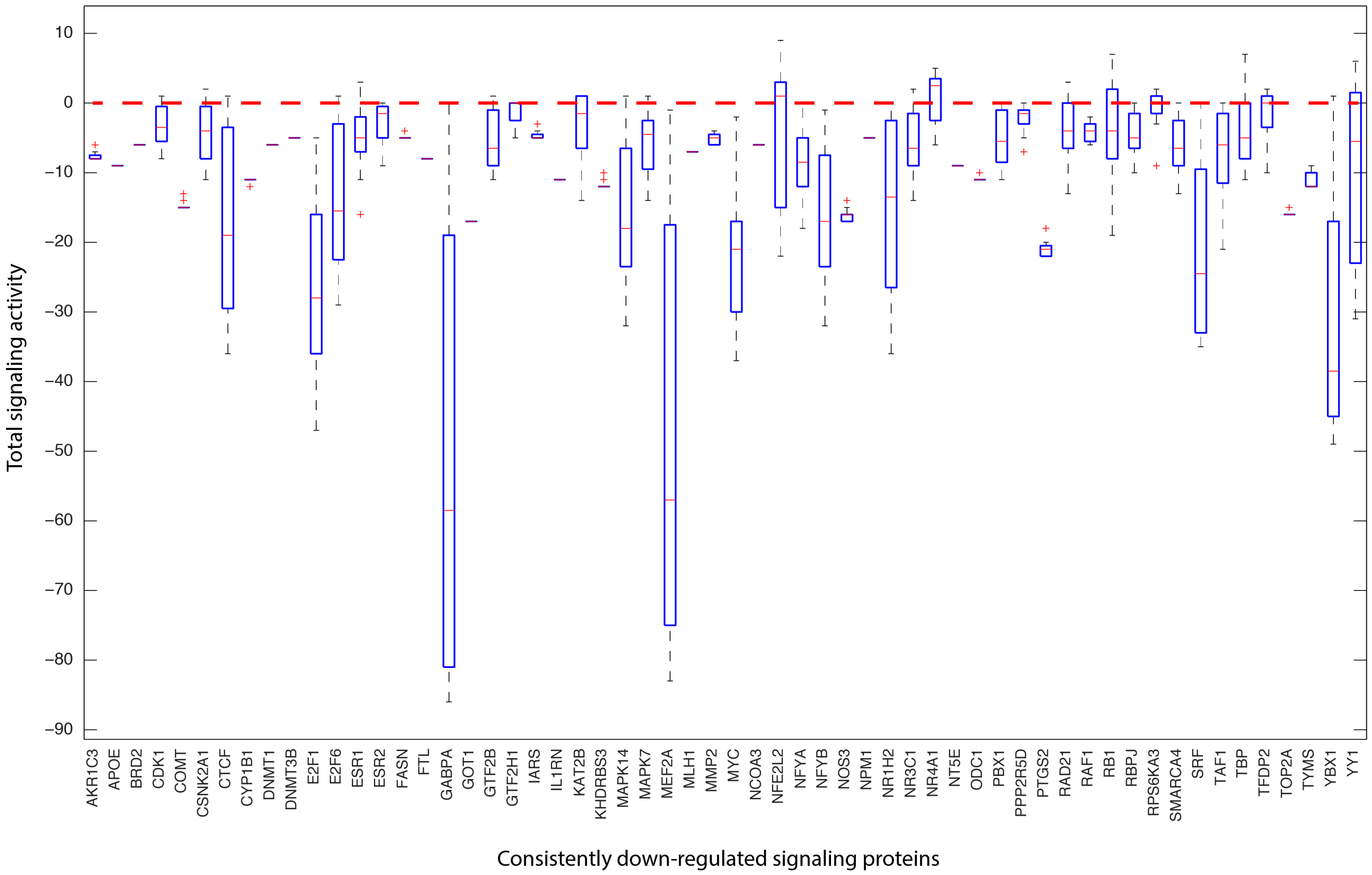

Figure 3: Dependence of the signaling activity of the consistently down-regulated proteins from model parameters. We observe that most of the nodes demonstrate the same trend across all $\alpha/\beta$ ratio values.

## ILP predictions on randomized data

To further explore the statistical significance of ILP predictions and subsequently, the significance of the DILD network and candidate drugs, we randomize the gene expression data, drug targets and PKN connectivity, repeat the pathway construction procedure, and compare our findings with the protein activities calculated based on the original PKN and data.

In more detail regarding the randomization of the gene expression data, we replaced the over- and under-expressed genes upon perturbation with the toxic compounds (as extracted from the cMAP), with arbitrary genes of the same number. The ILP was then implemented on the randomized dataset, constructing compound specific signaling pathways originating at the drug targets and terminating at the selected genes. The predicted protein activities were extracted and averaged across all drugs to identify peristent trends in similar fashion to what was performed with the original data and described in section 2.3. Results are shown in Supplementary Figure 4B. In addition to randomizing the gene expression data, the drug targets were also randomized. Drug targets, as obtained from the STITCH database, were replaced with arbitrary proteins of the same number and the pathway construction procedure was repeated. Results are shown in Supplementary Figure 4C. Finally, in Supplementary Figure 4D the gene expression data and drug targets were randomized simultaneously.

The effect of the PKN connectivity on model predictions was also interrogated. More specifically, the Switching Algorithm (as it is implemented in the BiRewire R package) [4] was applied for randomizing the network structure while conserving the node degrees, and then the pathway construction procedure was repeated to predict protein activities and investigate how these change depend on the network structure. The PKN was first converted into an unsigned network and then the Switching algorithm was applied, performing a total of 35,879,700 pairwise switching steps (by sampling a randomised version of the PKN every 358,797 switching steps, which corresponds to the empirical lower bound proposed in [4]). A total number 100 random networks were build in this way. Subsequently, in each of these random networks, 2% of the interactions (randomly selected) were labeled negative in accordance to the original PKN, and 50% of the interactions (randomly selected) changed directionality. The ILP was implemented to calculate compound specific signaling pathways for each of the 100 random PKNs. The predicted protein activities were averaged across all solutions and plotted in the Supplementary Figure 4E. Finally, the gene expression data, drug targets and PKN connectivity were all randomized simultaneously and results were plotted in Supplementary Figure 4F.

Overall, in all randomization setups we observe that the predicted protein activities are significantly different than the ones obtained from the original data. However, there are peristent trends present in the different setups. More specifically, we observe that in Supplementary Figure 4B and 4C, the proteins on the left end of the x-axis tend to be up-regulated and the proteins on the right tend to be down-regulated, in similar fashion to the protein activities of Supplementary Figure 4A. This is expected since protein activities are determined by a combination of factors, and randomizing either the gene expression data or drug targets alone is not enough to completely randomize the ILP results. For example, many of the consistently up- and down- regulated proteins are transcription factors connected to the differentially expressed genes. Thus, even if drug targets are randomized (4C), the signal has to, eventually, go through the transcription level to reach the differentially expressed genes, despite the fact it originated at different drug targets. In similar fashion, many the consistently up- and down-regulated proteins are drug targets. Thus, even if the gene expression data are randomized (4B), the signal has to originate at the same drug targets, despite the fact it terminates at different genes. On the other hand, if gene expression data and drug targets are randomized simultaneously (4D), the ILP results seem completely random. Similar observations can be made for the Supplementary Figure 4E. Randomizing the structure of the PKN significantly affects the ILP predictions, apart from a clear trend on the far left of the figure, where proteins are consistently up-regulated. This is attributed to the fact that many of these proteins are drug targets and signal transduction has to originate there (by design) regardless of the the network structure or gene expression data. Finally, all trends seem to have disappeared upon randomization of gene expression data, drug targets and PKN connectivity simultaneously (4F). The above make clear that the construction of compound specific signaling pathways leverages equally the gene expression data, prior knowledge of drug targets, and network connectivity and the contribution of every one of these factors can be isolated and quantified.

To quantify the randomization of protein activities (model predictions) upon randomization of the input data, we calculate how many of the 640 nodes that are predicted to be up-regulated in the final solution are also up-regulated after the randomization of the input data. And also how many of the 397 nodes that are down-regulated in the final solution are also down-regulated after the randomization of the input data.

With regards to Supplementary Figure 4B, 596 nodes are up-regulated in total, 457 of them are overlapping with the final solution. The statistical significance of the overlap is calculated using the hypergeometric cdf: 457 hits, out of 596 nodes predicted to be up-regulated with the randomized data, 640 nodes up-regulated in total in the final solution, and total number of nodes in the pool equal to 1150. This yields p-value=1.7757e-52. With regards to the down-regulated nodes, the overlap is 204 out of 299 nodes down-regulated in total, yielding p-value=1.5686e-45. These p-values validate what was previously observed, that nodes at the two ends of Figure 4B are showing a clear trend to either up- or down-regulation, in the same manner as with the original data of Figure 4A. And these trends are statistically significant. Similarly, statistically significant overlap with the original data of Figure 4A is observed for Figure 4C ($p-value = |O|^{-26}$) and Figure 4E ($p-value = |O|^{-48}$). The significance of the overlap is lower for Figure 4D, where with respect to the up-regulated nodes p-value=0.83 (not significant) and with respect to the down-regulated nodes p-value=1.02e-05. The overlap is lower for Figure 4F. With respect to the up-regulated nodes p-value=0.023 (borderline significant) and with respect to the down-regulated nodes p-value=0.3118 (not-significant). The borderline significant p-value for Figure 4F is more likely attributed to the node degrees of the PKN. Nodes with higher degrees are more likely to be used in the solution since they facilitate the connection of drug targets and gene expressions.

# References

[1] W.-Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386, 2002.

[2] J Hur, A Y Guo, W Y Loh, E L Feldman, and J P F Bai. Integrated systems pharmacology analysis of clinical drug-induced peripheral neuropathy. *CPT Pharmacometrics Syst Pharmacol*, 3:e114, 2014.

[3] Jane P F Bai, Andrey Utis, Gordon Crippen, Han-Dan He, Volker Fischer, Robert Tullman, He-Qun Yin, Cheng-Pang Hsu, Lan Jiang, and Kin-Kai Hwang. Use of classification regression tree in predicting oral absorption in humans. *J Chem Inf Comput Sci*, 44(6):2061–2069, Nov-Dec 2004.

[4] Andrea Gobbi, Francesco Iorio, Kevin J Dawson, David C Wedge, David Tamborero, Ludmil B Alexandrov, Nuria Lopez-Bigas, Mathew J Garnett, Giuseppe Jurman, and Julio Saez-Rodriguez. Fast randomization of large genomic datasets while preserving alteration counts. *Bioinformatics*, 30(17):i617–23, Sep 2014.

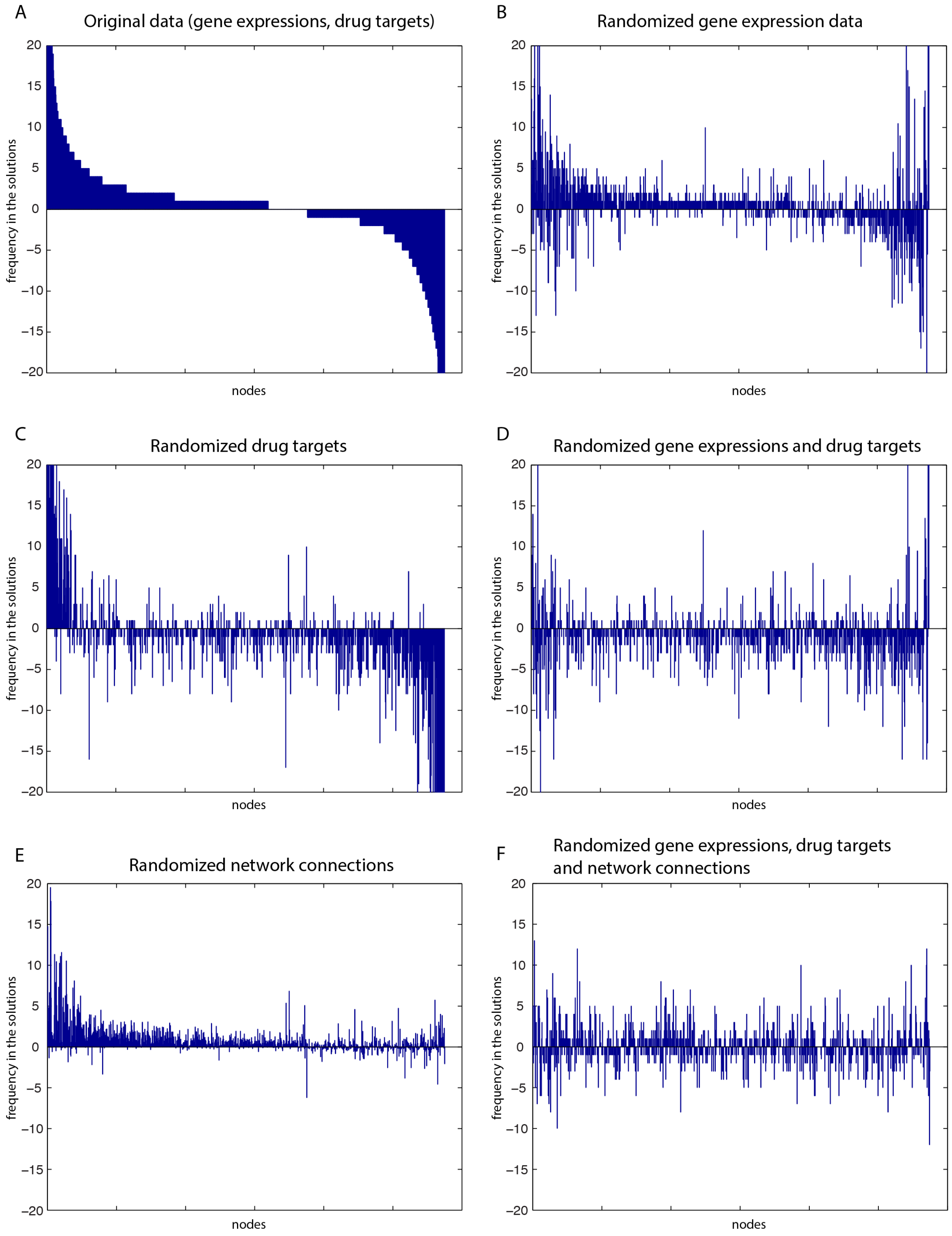

Figure 4: Average protein activities across all solutions with respect to the (A) Original data (gene expressions, drug targets, and PKN), (B) Randomized gene expression data, (C) Randomized drug targets, (D) Randomized gene expression data and drug targets, (E) Randomized PKN connectivity, (F) Randomized gene expression data, drug targets and PKN connectivity. The x-axis in all subfigures corresponds to individual signaling proteins; the y-axis corresponds to the number of solutions where the respective node is up- or down-regulated. The order of the nodes is the same in all subfigures. The nodes on the left end of the x-axis are consistently up-regulated across all solutions (with the original data) and the nodes on the right end are consistently down-regulated across all solutions.