iTOP: Inferring the Topology of Omics Data

Nanne Aben[1,2], Johan A. Westerhuis[3], Yipeng Song[3], Henk A.L. Kiers[4], Magali Michaut[1], Age K. Smilde[3,*], Lodewyk F.A. Wessels[1,2,5,*]

1 Division of Molecular Carcinogenesis, Oncode Institute, Netherlands Cancer Institute, Amsterdam 1066CX, The Netherlands.
2 Faculty of EEMCS, Delft University of Technology, Delft 2628CD, The Netherlands.
3 Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam 1098XH, The Netherlands.
4 Heymans Institute, University of Groningen, Groningen 9712CP, The Netherlands.
5 Cancer Genomics Netherlands, Utrecht 3584CT, The Netherlands.

*a.k.smilde@uva.nl and l.wessels@nki.nl

1 Supplementary Materials

1.1 The modified RV coefficient

For data matrices $\mathbf{X}$ where the number of variables is much greater than the number of objects (i.e. $p \gg n$), the RV coefficient is known to be biased upwards [5, 4]. To account for this bias, we remove the diagonal of the configuration matrix, as in the modified RV coefficient [5].

$$\tilde{\mathbf{S}}_i = \mathbf{S}_i - diag(\mathbf{S}_i)$$
$$\tilde{\mathbf{S}}_j = \mathbf{S}_j - diag(\mathbf{S}_j)$$
$$RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j) = \frac{Vec(\tilde{\mathbf{S}}_i)^T Vec(\tilde{\mathbf{S}}_j)}{\sqrt{Vec(\tilde{\mathbf{S}}_i)^T Vec(\tilde{\mathbf{S}}_i) \times Vec(\tilde{\mathbf{S}}_j)^T Vec(\tilde{\mathbf{S}}_j)}}$$

We note that for the modified RV coefficient, the average of $Vec(\tilde{\mathbf{S}})$ is not zero. This means that $RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j)$ is actually not equal to the correlation (but rather to the congruence) between $Vec(\tilde{\mathbf{S}}_i)$ and $Vec(\tilde{\mathbf{S}}_j)$. Regardless, for simplicity, we do describe the RV coefficient in terms of the correlation between $Vec(\tilde{\mathbf{S}}_i)$ and $Vec(\tilde{\mathbf{S}}_j)$ in the introduction and the first results subsection.

Mayer et al. (2011) [4] have reported that the modified RV coefficient does not correct all of the abovementioned $p \gg n$ bias. They propose the adjusted RV coefficient, based on the adjusted r^2 measure. However, the adjusted RV coefficient requires the data to be column-wise centered and autoscaled (i.e. scaled such that each column has a standard deviation of one). As we have shown in the Methods and Materials of the main text, binary datasets can be centered by kernel centering the configuration matrix (essentially using a set

of linear transformation to center the kernel space (corresponding to $\mathbf{S}$) rather than the input space ($\mathbf{X}$)). However, a similar approach cannot be taken with autoscaling, because determining the standard deviation (by which each column needs to be scaled) is a non-linear operation and hence cannot be performed in kernel space. Similarly, the adjusted RV coefficient requires one to take the adjusted r^2 between columns in the input space, which is also a non-linear operation that hence cannot be performed in kernel space. Finally, the benefit of the adjusted RV coefficient over de modified RV coefficient is extremely small when using a sufficient number of objects (e.g. $n > 50$) [4]. Therefore, we prefer to use the modified RV coefficient, which does not have the aforementioned limitations, while practically correcting the same amount of bias.

1.2 Partial Mantel Test

The concept of partial matrix correlations has been explored previously by Smouse et al. (1986) [6], who based their measure on the Mantel Test [3]. The Mantel test essentially measures the correlation on the vectorized form of the distance matrices (rather than configuration matrices) corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$. We prefer to base the partial matrix correlation on the RV coefficient instead because of two disadvantages of the Mantel Test. First, the Mantel Test does not necessarily result in a correlation close to zero for orthogonal data, while the RV coefficient does. Second, the Mantel Test always results in high matrix correlations when applied to high-dimensional matrices. While the original RV coefficient also suffers from the second limitation, the modified RV coefficient [5] alleviates this problem. Notably, this modification does not alleviate the problem for the Mantel Test. While both issues do not affect significance estimates resulting from a permutation test, they greatly affect the interpretation of the coefficients. Hence, we prefer to base our work on the RV coefficient rather than the Mantel Test.

1.3 PC algorithm

We used the order-independent PC algorithm proposed by Colombo and Maathuis (2014) [2], that was implemented in the R package pcalg. This algorithm uses partial correlations to infer a topology between variables (or in our work: partial matrix correlations to infer a topology between datasets). After inferring the topology, the PC algorithm can also attempt to infer causality between nodes in the topology, using two additional assumptions: 1) the causality graph underlying the data is a DAG (Directed Acyclic Graph); and 2) all variables are observed (or in our work: there are no hidden / unobserved datasets). It is important to keep these assumptions in mind when interpreting causality inferred by the PC algorithm.

1.4 Elastic Net regression

We used Elastic Net regression [7] as implemented in the R package glmnet, with λ set to λ_{min} and α set to 0.5. Predictive performance was assessed by using nested cross-validation, as implemented in the R package TANDEM, where the inner cross-validation loop was used to optimize the λ parameters for each stage, and the outer cross-validation loop was used to determine the predictive performance.

1.5 TANDEM

TANDEM [1] is a variable selection method that prioritizes variables selection from certain datasets over others. Consider a response vector $\mathbf{y}$ (e.g. drug response of a single drug) and two datasets $\mathbf{X}_1$ and $\mathbf{X}_2$. TANDEM performs the variable selection in two stages. In the first stage, Elastic Net regression [7] is used to explain as much of $\mathbf{y}$ as possible using $\mathbf{X}_1$. In the second stage, Elastic Net regression is used to explain the residuals from the first stage (i.e. the part of $\mathbf{y}$ that could not be explained using $\mathbf{X}_1$) using $\mathbf{X}_2$.

We used the implementation from the R package TANDEM, with λ set to λ_{min} for both stages and α set to 0.5. Predictive performance was assessed by using nested cross-validation, where the inner cross-validation loop was used to optimize the λ parameters for each stage, and the outer cross-validation loop was used to determine the predictive performance.

The relative contribution of a dataset was determined by dividing the sum-of-squares of the prediction from one dataset divided by the sum-of-squares of the overall prediction. For more information, we refer to Aben et al. (2016) [1].

We determined the variable importance VI of variable j in the same way as in our previous work on TANDEM [1], using:

$$VI = \frac{\|\mathbf{x}_j \boldsymbol{\beta}\|_2^2}{\|\mathbf{X}\boldsymbol{\beta}\|_2^2}$$

Where $\mathbf{X}$ is the input matrix for TANDEM, defined as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$; $\mathbf{x}_j$ is the j'th variable of $\mathbf{X}$; and $\boldsymbol{\beta}$ is the regression coefficients estimated by TANDEM.

References

[1] Nanne Aben, Daniel J Vis, Magali Michaut, and Lodewyk FA Wessels. Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, 2016.

[2] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[3] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.

[4] Claus-Dieter Mayer, Julie Lorent, and Graham W Horgan. Exploratory analysis of multiple omics datasets using the adjusted rv coefficient. *Statistical applications in genetics and molecular biology*, 10(1), 2011.

[5] Age K Smilde, Henk AL Kiers, S Bijlsma, CM Rubingh, and MJ Van Erk. Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*, 25(3):401–405, 2008.

[6] Peter E Smouse, Jeffrey C Long, and Robert R Sokal. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic zoology*, 35(4):627–632, 1986.

[7] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

2 Supplementary Figures

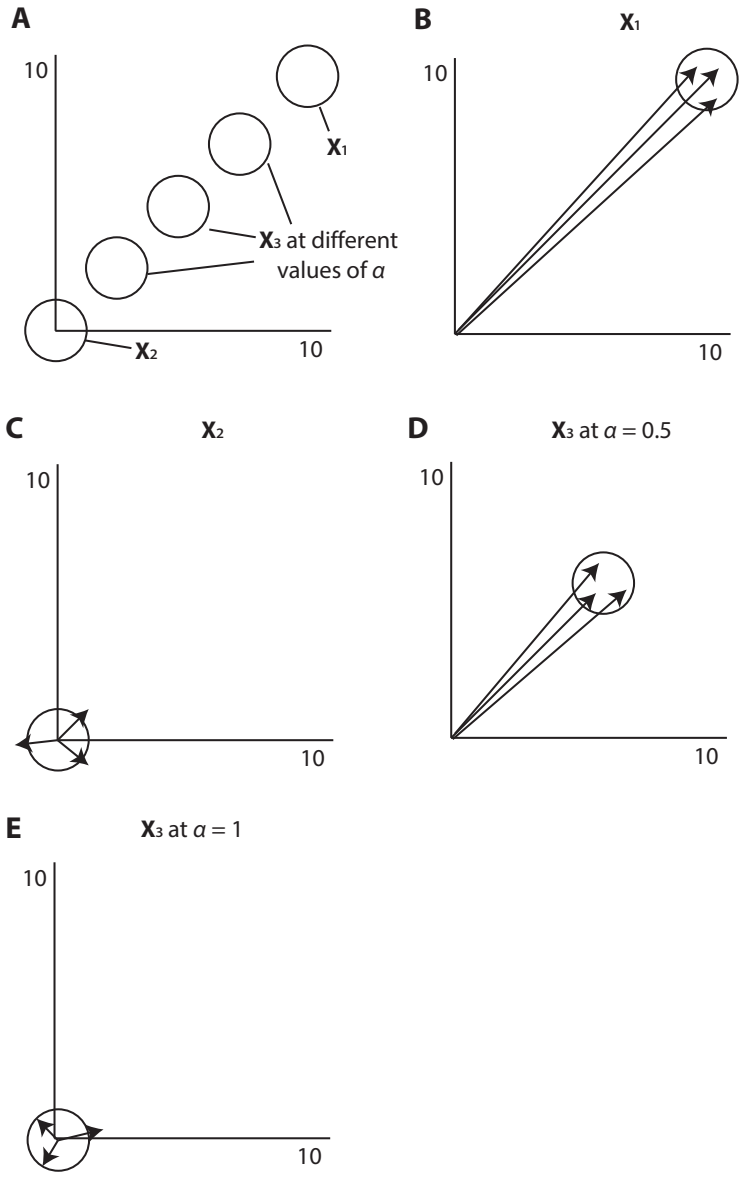
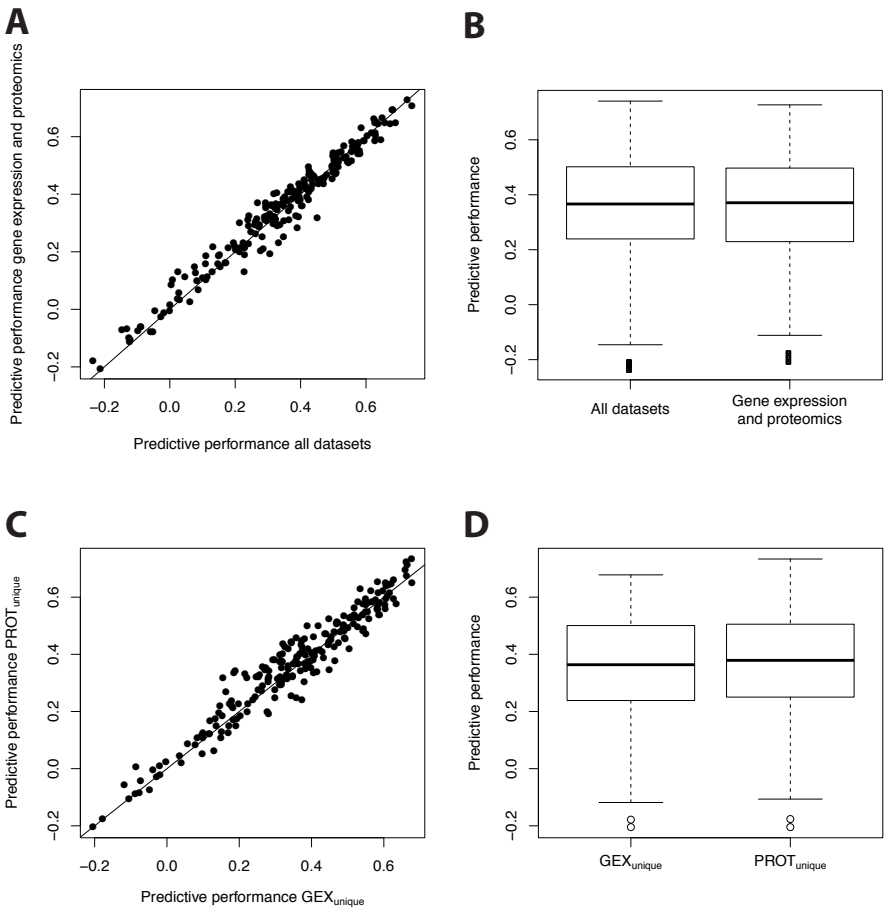


Figure 1: Illustration accompanying Figure 4A. (A) Cartoon of the densities of $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ in a two-dimensional space. (B-E) Cartoon of the directions of the inner products between objects from (B) $\mathbf{X}_1$, (C) $\mathbf{X}_2$, (D) $\mathbf{X}_3$ at $\alpha = 0.5$, and (E) $\mathbf{X}_3$ at $\alpha = 1$.


A
Predictive performance gene expression and proteomics
Predictive performance all datasets
−0.2
0.0
0.2
0.4
0.6
B
Predictive performance
All datasets
Gene expression and proteomics
C
Predictive performance PROT$_{unique}$
Predictive performance GEX$_{unique}$
D
Predictive performance
GEX$_{unique}$
PROT$_{unique}$

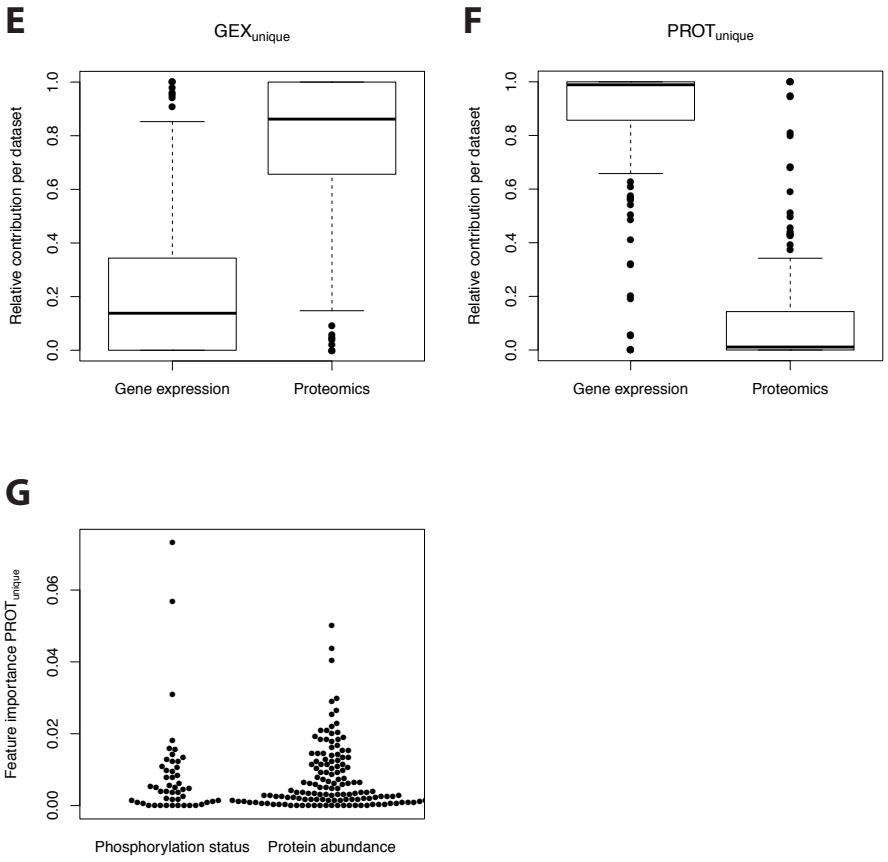


Figure 2: Drug response prediction models. (A) Predictive performance (Pearson correlation between observed and predicted drug response) of either a model trained on all datasets except drug response (i.e. mutation, CNA, methylation, cancer type, gene expression and proteomics), or a model trained on on gene expression and proteomics only, for each of the 217 drugs. (B) Predictive performance (Pearson correlation between observed and predicted drug response) of $\text{GEX}_{\text{unique}}$ vs. $\text{PROT}_{\text{unique}}$ models for each of the 217 drugs. (C&D) Distribution of relative contributions of gene expression and proteomics in $\text{GEX}_{\text{unique}}$ and $\text{PROT}_{\text{unique}}$ models respectively, across all 217 drugs. (E) variable importance for $\text{PROT}_{\text{unique}}$ models (averaged across drugs) for two classes of variables in the proteomics data: phosphorylation status and protein abundance.