## Simulated datasets

### Generating multi-omics data

Three datasets were simulated each with 200 observations (n) and 260 variables (p). The 200 observations were split equally over two groups (G1 and G2), whereas the 260 variables were generated by varying the degree of correlation and fold-change (δ) between G1 and G2: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables, and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables. The resulting dataset was of the form:



The discriminatory variables (corDis and unCorDis) were generated using the following model:



where the loadings, **w**1, **w**2, and **w**3 were 30-vectors, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3]. For G1, the outer components **u**1, **u**2, **u**3 were 3-vectors drawn from a multivariate normal distribution with a mean value of -δ/2 and a mean value of δ/2 for G2. For corDis variables, cor(**u**1, **u**2) = 1, cor(**u**1, **u**3) = 1, cor(**u**2, **u**3) = 1, whereas for unCorDis variables, cor(**u**1, **u**2) = 0, cor(**u**1, **u**3) = 0, cor(**u**2, **u**3) = 0.

The nondiscriminatory variables (corNonDis and unCorNonDis) were generated by drawing 100-vectors each with 200 elements, from a multivariate normal distribution with a mean of 0. For corNonDis variables, cor(**u**1, **u**2) = 1, cor(**u**1, **u**3) = 1, cor(**u**2, **u**3) = 1, whereas for unCorNonDis variables, cor(**u**1, **u**2) = 0, cor(**u**1, **u**3) = 0, cor(**u**2, **u**3) = 0.

***E****j* is a 200 x 260 residual matrix where each element is drawn from a normal distribution with zero mean and variance according to the grid [0.1, 0.2, 0.6, 1]. The following grid of values were used for the fold-change: [0.1, 0.5, 1, 2].

### Simulation analysis

Using fold-change values of [0.5, 1, 2] and noise values of [0.2, 0.5, 1, 2], 16 (4x4) sets of three datasets were generated, and DIABLO was applied, either with the full or null design (DIABLO\_full and DIABLO\_null). The full design, connects all blocks in the design matrix (C), such that cij=1, i=1,2,3 and j=1,2,3, whereas the null design does not connect any datasets in the design matrix (C), such that cij=0, i=1,2,3 and j=1,2,3. One component was retained in the DIABLO model, selecting 30 variables from each dataset for a total of 90 variables (across all datasets). In addition, other integrative schemes such as concatenation and ensemble-based classifiers were also tested using the sPLSDA classifier. For the concatenation-based scheme, all datasets were concatenated into one matrix containing 3x260=880 variables and sPLSDA was applied, retaining 1 component and 90 variables. For the ensemble-based scheme, a sPLSDA classifier was applied to each dataset separately retaining one component and 30 variables per dataset. The consensus predictions were determined using a majority vote scheme. A 10-fold cross-validation averaged over 50 simulations was used to evaluate the performance of each method/scheme and the number of each type of variable selected in each model was recorded.