Multi-view confounder detection for biomedical studies

Markus Wallner\textsuperscript{1, 2, 3}, Nicolas Kersten\textsuperscript{1, 3, 4}, and Nico Pfeifer\textsuperscript{1, 3*}

\textsuperscript{1}Interfaculty Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany
\textsuperscript{2}Institute for Translational Bioinformatics, University Hospital Tübingen, Germany
\textsuperscript{3}Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Germany
\textsuperscript{4}FZI Research Center for Information Technology, Karlsruhe, Germany
\textsuperscript{*}nico.pfeifer@uni-tuebingen.de

ABSTRACT

In many biomedical studies an important first step is checking for confounding factors. For association studies, confounding can for example be caused by ethnic differences in the case and control groups. In many other settings there might be confounding factors like batch effects or founder effects that also need to be detected and controlled for\textsuperscript{1}. Detecting confounding for data from one data source is well established (e.g., genomics data). Since more and more studies are now based on data from multiple data modalities (e.g., multi-omics), we evaluated whether multi-view confounder detection can benefit from state-of-the-art methods for multi-view data integration. Especially for clustering of multi-omics data, it has been shown that these methods can perform better than methods that treat the data modalities separately\textsuperscript{2}. Our results show that multi-view confounder analysis is possible and that building on multi-view data integration methods is better than treating the different data modalities separately.

When analyzing the genetic basis for common and complex diseases, confounding factors have to be detected and controlled for to decrease the likelihood of finding false positive associations\textsuperscript{3}. If one also has additional data modalities like methylation data, then these data have to be controlled for confounding, too\textsuperscript{4}. If the signal for confounding is weak and distributed among several data modalities (e.g., genetic data and methylation data), methods for joint detection of confounding might be more sensitive compared to methods that check for confounding separately. Similar advantages have been shown for cancer subtype detection based on multi-omics data\textsuperscript{2} as well as a synthetic benchmark data set\textsuperscript{5}. To the best of our knowledge, multi-view confounder detection has never been evaluated systematically. Therefore, we devised a similar synthetic benchmark data set to the one introduced for multi-omics clustering by Rappoport and Shamir (2019)\textsuperscript{5} and evaluated the performance of different ways to detect confounding. In addition to simulated data sets, we also performed similar analyses on real cancer data.

In order to compare the different approaches, we devised a setting in which all methods can be used in a similar way to detect the confounding. The main hypothesis is that if samples cluster into two groups that are not independent of the potential confounder, then this confounder could influence any analysis of the data and should therefore be accounted for. Therefore, we tested whether there was a significant correlation between the resulting cluster labels of each method and the labels of the potential confounding factor by performing a permutation test based on chi-squared tests. The only difference in the evaluated methods is how to devise the clusters or how to handle the results from the tests based on the clusters.

As shown in a benchmark by Rappoport and Shamir in 2018\textsuperscript{2}, there are some key differences in performance between available multi-view clustering algorithms. In order to evaluate the ability to detect multi-view confounding factors, we developed multi-view confounding tests (Multi-Contest) based on several state-of-the-art methods like rMKL-LPP\textsuperscript{6} (Multi-Contest\textsuperscript{rMKL-LPP}), NEMO\textsuperscript{7} (Multi-Contest\textsuperscript{NEMO}), COCA\textsuperscript{7} (Multi-Contest\textsuperscript{COCA}), as well as two simple logical operations (Multi-Contest\textsuperscript{and-P}, Multi-Contest\textsuperscript{or-P}). The clustering by COCA as well as the logical operations are based on single omics K-means clustering. For Multi-Contest\textsuperscript{or-P} one and for Multi-Contest\textsuperscript{and-P} all omics have to show a significant p-value to detect the potential confounding factor in question. Based on these detections and the ground truth we calculated Matthews correlation coefficient (MCC) to compare the performance in the different simulation settings.

The first simulation setting contains two cluster centers. The simulated samples are represented by points normally distributed around the respective cluster centers. The omics are then generated by adding normal distributed noise (Gaussian noise) to these simulated sample points. We increased the spread of samples around the cluster center as well as the spread of omics around their sample point to test how the performance is influenced. For all these settings we tested different numbers of omics and different strengths of the confounding. Fig. 1 summarizes the performance of the five tested methods for the first simulation setting (sensitivity and specificity plots in Supplementary Fig. 1). Overall, a higher spread of omics/samples leads
to lower MCC values. The spread of samples around the cluster centers seems to have a stronger impact than the spread of omics around the samples. For the four different settings of sample and omics spread in this setup, Multi-Contest\textsuperscript{MKL-LPP} and Multi-Contest\textsuperscript{NEMO} outperform the other methods. The two logical operations perform worst, with Multi-Contest\textsuperscript{and-P} showing significantly lower MCC values.

To evaluate the different methods in a more complex setup, we extended the simulation setting by introducing a third cluster in which all clusters have the same distance to each other. The results for this setting are shown in Fig. 2 (sensitivity and specificity plots in Supplementary Fig. 2). Here Multi-Contest\textsuperscript{MKL-LPP} and Multi-Contest\textsuperscript{NEMO} show the highest MCC values as well. The distance to the next best method (Multi-Contest\textsuperscript{COCA}) is larger than in the previous simulation setting. Multi-Contest\textsuperscript{or-P} shows notably worse performance in this setting, while Multi-Contest\textsuperscript{NEMO} and Multi-Contest\textsuperscript{MKL-LPP} achieve similar MCC values compared to the first setting.

Up to this point all simulated omics of a data set were based on the same underlying structure without omics that were not affected by the confounding factor. In order to create a more refined simulation, we changed the simulation setting such that only a portion of the omics is influenced by a confounding signal with a similar strength as before (now called signal omics / SO). For the remaining omics we reduced the distance between cluster centers and increased the spread for samples around these centers as well as the spread of omics around the sample points. These omics are now called noise omics (NO). The total number of omics per sample was fixed to ten. We tested this setting with 2, 3, and 5 SO and two different cluster distances for the NO each. The resulting MCC values are summarized in Fig. 3 (sensitivity and specificity plots in Supplementary Fig. 3). In this setting the performance of Multi-Contest\textsuperscript{NEMO} (MCC) seems to depend strongly on the number of signal omics. The other methods (except for Multi-Contest\textsuperscript{and-P}) show a trend of higher MCC values with more signal omics as well, but the order of the methods does not change. Only Multi-Contest\textsuperscript{NEMO} improves from second worst with a high amount of noise omics to best/shared best MCC values with an equal amount of signal omics and noise omics. Considering all variations of this setting Multi-Contest\textsuperscript{MKL-LPP} shows the most robust performance.

Finally we evaluated the methods on three TCGA data sets with the following cancer types: Glioblastoma Multiforme (GBM), Liver Hepatocellular Carcinoma (LIHC), and Lung Squamous Cell Carcinoma (LUSC). For each data set gene expression, methylation, and miRNA expression data was included. Additionally we used the gender information as potential confounding factor and compared confounding tests for individual omics clusterings to the previously described Multi-Contest settings. The resulting p-values for Multi-Contest\textsuperscript{or-P} (where a significant test result in one of the omics is sufficient) were adjusted for multiple testing using Bonferroni correction. All resulting p-values are summarized in Table 1.

For the LIHC data set the clustering of the baseline methods as well as Multi-Contest\textsuperscript{COCA}, Multi-Contest\textsuperscript{NEMO}, and Multi-Contest\textsuperscript{MKL-LPP} lead to a positive test result. In this data set the confounding effect is strong enough to be even detectable based on the individual omics data. For this cancer type previously a male to female ratio of 2.5 was observed indicating some kind of gender bias\textsuperscript{8}. With ongoing research, a number of molecular mechanisms that are leading to the observed sex-bias have been discovered\textsuperscript{9–12}. For the GBM data set only the test of confounding based on gene expression, Multi-Contest\textsuperscript{or-P}, Multi-Contest\textsuperscript{NEMO}, and Multi-Contest\textsuperscript{MKL-LPP} resulted in a significant p-value and therefore indicated gender as a confounding factor. Since significant confounding could in this case already be detected by only analyzing gene expression data, an additional evaluation was performed on the remaining two data types that did not show significant confounding in the single-omics setting (GBM-red.). In the GBM-red. data set the more sophisticated methods Multi-Contest\textsuperscript{NEMO} and Multi-Contest\textsuperscript{MKL-LPP} still lead to a significant test result. Previous studies showed that for some GBM subtypes a sex-bias is present\textsuperscript{13}. A recent study also reported sex-specific molecular subtypes of glioblastoma multiforme\textsuperscript{14}. This matches with the positive test result based on gene expression, Multi-Contest\textsuperscript{or-P}, Multi-Contest\textsuperscript{NEMO}, and Multi-Contest\textsuperscript{MKL-LPP}. In the GBM-red. example the confounding was still detectable based on Multi-Contest\textsuperscript{NEMO} and Multi-Contest\textsuperscript{MKL-LPP} even without the expression data. This is an example for a data set where the representation from the integration of multiple omics using consensus clustering (COCA) does not indicate the presence of confounding, while the more sophisticated representations based on NEMO and rMKL-LPP enable a positive test result.

Lung squamous cell carcinoma is a type of non-small cell lung carcinoma (NSCLC). For NSCLC male gender was confirmed to be an independent unfavorable prognostic indicator for survival\textsuperscript{15}. Other studies also reported the influence of gender on targeted treatments\textsuperscript{16,17} as well as gender-specific somatic alterations that influence prognostic biomarkers\textsuperscript{10}. The sex-bias that is reported in the literature was not observed with any of the tested methods for this data set, while some of the advanced methods showed a trend towards significance. Since a ground truth is not available here, it is not possible to decide if the confounding was not detected or if it was not present in this specific data set.

All results from the simulated data sets show, that the confounder detection based on the more sophisticated multi-omics clustering methods NEMO and rMKL-LPP outperforms the approaches based on the individual omics. For the TCGA data
sets we did observe that the confounding in the GBM-red. data set was still detectable by using NEMO and rMKL-LPP representations even though the omic that was most influenced by the confounding factor was removed. This was not possible using the other tested methods. This indicates that when the confounding effect is less pronounced, then methods that consider the omics data jointly are better suited to detect the confounding than methods which evaluate the different data modalities separately.

Figure 1. MCC values for the simulation settings with two cluster centers. Results are shown as means ± SEM of Matthews Correlation Coefficient (MCC) values from different number of omics and strength of simulated confounding (triplicates for each combination, total n = 36). Different combinations of sample spread and omics spread were used to test how the methods perform on a less clear cluster structure.
Figure 2. MCC values for the simulation settings with three cluster centers. Results are shown as means ± SEM of Matthews Correlation Coefficient (MCC) values from different strength of simulated confounding with ten omics (triplicates for each strength, total n = 24). Different combinations of sample spread and omics spread were used to test how the methods perform on less a less clear cluster structure.

Table 1. Resulting p-values from testing gender labels as potential confounding factor for the three cancer data sets Liver Hepatocellular Carcinoma (LIHC), Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LUSC), and a reduced GBM data set (GBM-red.) obtained from GBM by removing the expression data. The expression, methylation, and miRNA results are based on single omics K-means clusterings. For Multi-Contest^{and-P} the highest p-value of the individual omics is reported (because all have to give a significant result). For Multi-Contest^{or-P} the lowest p-value of the single omics is reported (with adjustment for multiple testing) because a significant result in one of the tested omics is sufficient. Multi-Contest^{COCA} combines the information of these three single omics clusterings, while Multi-Contest^{NEMO} and Multi-Contest^{rMKL-LPP} do not rely on those single omics clusterings. Values below the significance threshold of $\alpha = 0.05$ are shown in bold face.

<table>
<thead>
<tr>
<th>Method</th>
<th>LIHC</th>
<th>GBM</th>
<th>GBM-red.</th>
<th>LUSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>expression</td>
<td>$&lt;10^{-3}$</td>
<td>0.013</td>
<td>-</td>
<td>0.073</td>
</tr>
<tr>
<td>methylation</td>
<td>0.031</td>
<td>0.517</td>
<td>0.517</td>
<td>0.261</td>
</tr>
<tr>
<td>miRNA</td>
<td>0.041</td>
<td>1.000</td>
<td>1.000</td>
<td>0.727</td>
</tr>
<tr>
<td>Multi-Contest^{and-P}</td>
<td>0.041</td>
<td>1.000</td>
<td>1.000</td>
<td>0.727</td>
</tr>
<tr>
<td>Multi-Contest^{or-P}</td>
<td>$&lt;3\cdot10^{-3}$</td>
<td>0.039</td>
<td>1.000</td>
<td>0.219</td>
</tr>
<tr>
<td>Multi-Contest^{COCA}</td>
<td>$&lt;10^{-3}$</td>
<td>0.104</td>
<td>1.000</td>
<td>0.073</td>
</tr>
<tr>
<td>Multi-Contest^{NEMO}</td>
<td>$&lt;10^{-3}$</td>
<td>$&lt;10^{-3}$</td>
<td>$&lt;10^{-3}$</td>
<td>0.060</td>
</tr>
<tr>
<td>Multi-Contest^{rMKL-LPP}</td>
<td>0.040</td>
<td>0.010</td>
<td>0.042</td>
<td>0.097</td>
</tr>
</tbody>
</table>
Figure 3. MCC values for simulated data sets containing signal omics and noise omics. Results are shown as means ± SEM of Matthews Correlation Coefficient (MCC) values from different strength of simulated confounding with ten omics (triplicates for each strength, total n = 12). The signal omics (SO) to noise omics (NO) ratio (SO:NO ratio) as well as the noise level used for the simulation are indicated at the bottom of the figure.
Methods

Data simulation
In order to evaluate the performance of different approaches to reveal the presence of confounding factors, different data sets were simulated. With simulated data sets the ground truth about cluster affiliation of samples and the presence of confounding is known. The basis for the simulation settings is a simulation used by Rappoport and Shamir to test the performance of multi-omics clustering methods (including NEMO)\(^5\).

All simulated data sets contain 300 samples and for each setting 300 data sets were simulated. Each sample gets labeled with a gender label (‘m’ or ‘f’) as confounding factor, 150 samples per label. With no confounding present in these data sets no difference in gender label proportions in the clusters would be expected. To introduce confounding, the probability for samples with a specific gender label to be assigned to a certain cluster was changed in the simulation. All cluster centers (CC1 - CC5 are shown in Table 2. The most basic simulations contain two cluster centers in \(\mathbb{R}^{10}\) at CC1 and CC2. The samples were then randomly assigned to the two cluster centers. For the samples labeled with ‘f’ both clusters had a probability of 0,5, for the ones labeled with ‘m’ the probability for the first cluster center was increased to different levels in the range 0,55 - 0,70 (step size 0,05) to simulate different levels of confounding. Each sample gets assigned a sample point based on a multivariate normal distribution with the identity matrix \(I\) as covariance matrix around the cluster centers. Individual omics profiles are now generated by adding normal noise \(\mathcal{N} (\mu = 0, \Sigma = 2I)\) to the sample points. Data sets containing 2, 5, and 10 omics were created to model different sizes of data sets.

To test the influence of the spread of samples around the cluster centers and omics around sample points we changed the covariance matrices used for omics generation to \(3I\) and \(4I\) and for the final setting the covariance matrix for the sample point generation to \(2I\) while keeping the omics generation with \(2I\).

In the next step the simulation setting was extended to three cluster centers by adding a third cluster center (CC3). In the process the probabilities for the cluster assignment were adjusted as well. The samples labeled with ‘f’ have equal probability for all three clusters. For the samples labeled with ‘m’ the probability to get assigned to the first cluster was set in the range 0,15 - 0,5 (step size 0,05) while the remaining probability was split evenly between the remaining two clusters.

With this setting only data sets containing ten omics were simulated. For the spread of samples around the cluster centers or omics around the sample points, three settings were tested. First the default covariance matrices as in the first setting was used, secondly \(2I\) for the samples around the cluster centers, and lastly \(3I\) was used for the noise generating the omics from the sample points.

Up to this point, all omics of one data set were based on the same underlying structure (position of the sample points around the cluster centers). To better simulate different information content and structure in different omics a third setting was created. Here data sets with two cluster centers and 10 omics were created with a number of omics \((2, 3, 5)\) containing a rather clear cluster separation (signal omics / SO) and the remaining omics where the cluster centers are closer together and the spread is increased (noise omics / NO). With this general approach two noise levels were used. The signal omics were created with the same settings as used in the first two cluster setting. For the first noise level (low) the cluster centers are CC4 and CC2 with covariance matrix for sample points equal to \(2I\) and for omics generation equal to \(3I\).

For the second noise level (high) the cluster centers CC5 and CC2 were used with the same covariance matrices.

| Table 2. Cluster centers (CC) used for simulation of synthetic data sets. |
|------------------|------------------|
| Cluster center   | Coordinates in \(\mathbb{R}^{10}\) |
| CC1              | (2,2,2,2,0,0,0,0,0,0) |
| CC2              | (0,0,0,0,0,0,0,0,0,0) |
| CC3              | (1,1,1,1,1,1,732,1,732,1,732) |
| CC4              | (1,1,1,1,1,0,0,0,0,0) |
| CC5              | (0,5,0,5,0,5,0,5,0,0,0) |

Real omics preprocessing
In order to test the performance of the different methods on real omics data we used the Glioblastoma Multiforme (GBM), Liver Hepatocellular Carcinoma (LIHC), Lung Squamous Cell Carcinoma (LUSC) preprocessed data sets that were previously used to benchmark multi-omics clustering algorithms\(^2\) and are based on TCGA data sets (downloaded from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html). On these data sets the method-specific preprocessings were applied as described in the benchmark\(^2\). Each data set contains gene expression, DNA methylation and miRNA expression data. All samples that were not labeled as primary tumor were dropped. For each data set all patients that are not present in all
Table 3. Covariance matrices used in simulation (sample/omics spread) of synthetic data sets based on labels used in result graphs.

<table>
<thead>
<tr>
<th>Label in graph</th>
<th>sample spread</th>
<th>omics spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>1I</td>
<td>2I</td>
</tr>
<tr>
<td>medium</td>
<td>2I</td>
<td>3I</td>
</tr>
<tr>
<td>high</td>
<td>-</td>
<td>4I</td>
</tr>
</tbody>
</table>

three omics were dropped (resulting number of samples: GBM:274, LIHC:369, LUSC:341). For K-means clustering, rMKL-LPP and Nemo, the data was filtered to exclude features with zero variance and all features based on sequencing data were log-transformed. The data was further reduced for K-means clustering such that only the 2000 highest variance gene expression and methylation features were selected. Finally all selected features were normalized to have a mean of zero and a standard deviation of 1. In addition to the complete real omics data sets, we created a reduced version of the GBM data (GBM-red.) by removing the gene expression data modality.

For these preprocessing steps we used the R functions from the NEMO GitHub page (https://github.com/Shamir-Lab/NEMO).

K-means

K-means clusterings were performed on the single omics data sets. Here the K-means function of the scikit-learn Python package (sklearn.cluster.KMeans) was used. The best value for the number of clusters k was selected by average silhouette score for a predefined range of k (2 <= k <= 15) (sklearn.metrics.silhouette_score).

Cluster-Of-Cluster Analysis (COCA)

In order to integrate the clusterings based on the single omics types we used Cluster Of Clusters Analysis (COCA). This method is widely used in cancer studies since its first introduction. The goal of this method is to combine the information of the single omics clusterings into one consensus clustering. The R package "coca" was used to perform the consensus clustering step based on a matrix of clusters created form the single omics k-means clusterings. The maximum number of iterations (B) was set to 2000. All other parameters were used with their default values.

Neighborhood-based multi-omics clustering (NEMO)

NEMO is a simple similarity-based multi-omics clustering approach. The clustering is based on inter-patient similarity matrices for the different omics that get integrated into one matrix before the final clustering step. We used the R package as described in the publication. All parameters were kept at default values.

rMKL-LPP

Regularized Multiple Kernel Learning with Locality Preserving Projections (rMKL-LPP) is an unsupervised joint dimensionality reduction method that was first introduced in 2015. rMKL-LPP performs a non-linear dimensionality reduction by using a kernelized representation of the sample-similarities for each data type. By learning a weighted linear combination of these kernel matrices, multi-omics data is integrated and projected to the n-dimensional space. Sample clusters are subsequently computed by applying the K-means algorithm to the dimensionality-reduced projection of the samples. To determine the number of clusters k, the average silhouette score metric is used for a predefined range of k (2 <= k <= 15).

The input for rMKL-LPP are similarity matrices, here computed using the radial basis function (RBF) kernel.

\[ K_{RBF}(x, y) = \exp(-\gamma \|x - y\|^2) \]  

For all data sets (simulated and real cancer data) the parameter \(\gamma\) of the RBF kernel was set to the general rule of thumb

\[ \gamma_{\text{default}} = \frac{1}{2p}, \]  

with \(p\) being the number of features.

For the rMKL calculation, two parameters were selected, \(k_N\) the number of neighbors and \(n\) the number of dimensions to which the samples will be projected. For \(k_N\) the default value of nine was used for all computations. For the simulated data sets the number of dimensions \(n\) was set to ten. For the cancer data sets, the number of dimensions was set to five (the default value for web-rMKL).
Test for presence of confounding factors

The test for the presence of confounding is based on cluster labels as well as labels for potential confounding factors. We used the chi-squared test for independence with cluster assignments and potential confounder labels (‘f’/’m’ in the simulated data sets). Previous studies reported the possibility of inaccurate approximation of the statistic and a resulting overestimation of significance\(^2\). Because of this we decided to perform permutation tests as they were suggested previously\(^2\).

For the permutation test we randomly permuted the cluster labels of the samples. We performed \(1e3\) permutations at a time until the 95% confidence interval did not cross 0.05 or the maximum number of \(1e5\) iterations was reached. The confidence intervals were computed based on the number of permutations resulting in a greater or equal chi-squared statistic compared to the original labeling. For this the binomtest of scipy (scipy.stats.binomtest) was used. For the chi-squared test we used a function from python scipy (scipy.stats.chi2_contingency). If the resulting p-value is significant with an alpha cutoff of 0.05 the tested factor is counted as a confounding factor.

For the most basic evaluation we used the K-means clusterings of the single omics and required either one (or-P) or all omics (and-P) to give a significant result to count the tested factor as confounder. This results in multiple tests per data set and therefore bonferroni multiple testing correction was applied (using statsmodels.stats.multipletests) for the or-P setting. For the more advanced methods (COCA, NEMO, rMKL-LPP) the result is only one clustering so no multiple testing correction was required. We called the tests based on these clusterings Multi-Contest\(^{COCA}\), Multi-Contest\(^{NEMO}\), and Multi-Contest\(^{rMKL-LPP}\).

For the simulated data sets the ground truth for the presence of confounding is calculated based on the known true cluster assignments used in the simulation.

Matthews correlation coefficient

To compare the confounder detection performance of the different methods over the different simulation settings we calculated Matthews correlation coefficient (MCC) (calculated with sklearn.metrics.matthews_corrcoef). The MCC values are calculated for each data set and then combined into plots. One plot for the simple settings containing two clusters, one for the settings with three clusters, and one for the settings with different strength of confounding in the omics. For the simple two cluster settings the different covariance matrix combinations (sample/omics spread) are plotted as separate points while the different number of omics and the different strength of confounding is combined into mean and standard error. For the settings with three clusters an equivalent plot is created (here no combination of different numbers of omics in the data set were simulated). For the final simulation setting the different combinations of covariance matrices and weakness level get plotted as separate points.
References


**Acknowledgements**

The results shown here are in whole or part based upon data generated by the TCGA Research Network: [https://www.cancer.gov/tcga](https://www.cancer.gov/tcga). M.W. and N.P. acknowledge funding from ZIV Consortium (Center for Innovative Care, digital@bw, 42-04HV.MED(18)/25/1). N.K. acknowledges funding from the Ministry of Economic Affairs, Labour and Housing of Baden-Württemberg (WM BW) within the ’PRIMO’ project (reference number 3-4332.62-HSG/84).

**Author contributions statement**

Conceptualization, N.P. and M.W.; data simulation and processing, M.W.; implementation and statistical analysis, M.W. and N.K.; project supervision, N.P. and N.K.; All authors contributed equally to result interpretation, manuscript preparation and manuscript revision.

**Additional information**

The authors declare no conflicts of interest.