Dealing with dimensionality

the application of machine learning to multi-omics data

Feldner-Busztin, Dylan; Nisantzis, Panos Firbas; Edmunds, Shelley Jane; Boza, Gergely; Racimo, Fernando; Gopalakrishnan, Shyam; Limborg, Morten Tønsberg; Lahti, Leo; de Polavieja, Gonzalo G.

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btad021

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Dealing with dimensionality: the application of machine learning to multi-omics data

Dylan Feldner-Busztin 1,*,†, Panos Firbas Nisantzis 1,*,†, Shelley Jane Edmunds 2, Gergely Boza 3, Fernando Racimo 4, Shyam Gopalakrishnan 2, Morten Tønsberg Limborg 2, Leo Lahti 5 and Gonzalo G. de Polavieja 1,*

1Champalimaud Centre for the Unknown, Champalimaud Foundation, 1400-038 Lisbon, Portugal, 2Center for Evolutionary Hologenomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark, 3Centre for Ecological Research, 1113 Budapest, Hungary, 4Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark and 5Department of Computing, University of Turku, 20014 Turku, Finland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on July 20, 2022; revised on December 2, 2022; editorial decision on January 6, 2023; accepted on January 11, 2023

Abstract

Motivation: Machine learning (ML) methods are motivated by the need to automate information extraction from large datasets in order to support human users in data-driven tasks. This is an attractive approach for integrative joint analysis of vast amounts of omics data produced in next generation sequencing and other -omics assays. A systematic assessment of the current literature can help to identify key trends and potential gaps in methodology and applications. We surveyed the literature on ML multi-omic data integration and quantitatively explored the goals, techniques and data involved in this field. We were particularly interested in examining how researchers use ML to deal with the volume and complexity of these datasets.

Results: Our main finding is that the methods used are those that address the challenges of datasets with few samples and many features. Dimensionality reduction methods are used to reduce the feature count alongside models that can also appropriately handle relatively few samples. Popular techniques include autoencoders, random forests and support vector machines. We also found that the field is heavily influenced by the use of The Cancer Genome Atlas dataset, which is accessible and contains many diverse experiments.

Availability and implementation: All data and processing scripts are available at this GitLab repository: https://gitlab.com/polavieja_lab/ml_multi-omics_review/ or in Zenodo: https://doi.org/10.5281/zenodo.7361807.

Contact: gonzalo.polavieja@neuro.fchampalimaud.org or panos.firbas@research.fchampalimaud.org or dylan.feldner@research.fchampalimaud.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Algorithmic and hardware developments, including graphics processing unit computation, have spurred a revolution in machine learning (ML) (Dally et al., 2021). At the same time, the amount of data generated by -omic (genomic, transcriptomic, etc.) high-throughput-sequencing and other techniques has been growing exponentially (Lightbody et al., 2019). This growing body of information needs statistical models that can extract accurate and explainable predictions from it.

The concept of ‘training’ defines ML techniques. A subset of the original data is used to train, or change the model’s parameters, so that the model can then make the best possible predictions or decisions. ML models typically work better with large training datasets and as such should be particularly well tailored for multi-omic data integration. The ML field is an exciting frontier and several reviews have been published in this area recently. Reel et al. (2021) rated ML algorithms based on their data-hungri ness, prediction accuracy and other characteristics. The review by Cai et al. (2022) highlighted the use of The Cancer Genome Atlas (TCGA) in multi-omics research, including further independent benchmarking of ML techniques on another dataset—the Cancer Cell Line Encyclopedia (Ghandi et al., 2019; Nusinow et al., 2020). Marcos-Zambrano (2021) and Moreno-Indias (2021) explored ML multi-omics in the context of microbiome research. While these reviews have illustrated that ML applications can handle and thrive on large volumes of data
from multi-omics datasets there is one caveat: the modelling of these datasets often suffers from the low sample size compared to the vast dimensionality.

Ideally, ML techniques would use more samples than features (Bellman, 1961). In practice, however, a single -omic dataset can contain tens of thousands of features (e.g. the result of RNAseq on a tumour sample can include measurements for over 20 000 human genes). This makes multi-omics datasets high-dimensional. On the other hand, most datasets contain at most only a few hundred samples, i.e. one per subject.

Aware of this dimensionality issue, we set out to explore how prevalent it is in the field of ML data integration in multi-omics, what strategies are used to overcome it, and what can be achieved with ML in multi-omics. We took a quantitative approach to gathering characteristics of papers (e.g. techniques used, goals for using ML) from papers and used the PRISMA-ScR framework (Tricco et al., 2018) as a guideline for performing a more structured review.

Following the PRISMA-ScR guidelines, we developed an explicit statement of the questions being addressed in this scoping review (Table 1).

Table 1. Review questions

<table>
<thead>
<tr>
<th>Review questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is multi-omic data in practice?</td>
</tr>
<tr>
<td>1.1. Which types of -omics features made up the ‘data dimensions’?</td>
</tr>
<tr>
<td>1.2. How is the data structured: how many features versus samples?</td>
</tr>
<tr>
<td>2. What analysis has been done on multi-omics data?</td>
</tr>
<tr>
<td>2.1. Which ML techniques were used?</td>
</tr>
<tr>
<td>2.2. What were the goals of the ML application?</td>
</tr>
<tr>
<td>2.3 What were the targets/labels of classification tasks?</td>
</tr>
<tr>
<td>3. Can we explain trends using an analysis of the citation of papers since 2015?</td>
</tr>
</tbody>
</table>

3 Results

3.1 What is multi-omic data in practice?

3.1.1 Which types of -omics features made up the ‘data dimensions’?

Using the papers from the ‘ML AND multi-omics AND integration’ search, we compiled the types of -omics data that were used by researchers. Transcriptomics were by far the most popular measurement and were used 152 times (many works use multiple types of transcriptomics such as mRNA and miRNA). This accounted for 42% of all the -omics data uses (Fig. 2a). Epigenomics and genomics data follow with 79 (22%) and 77 uses (21%) respectively. After that were proteomics (21 uses, 6%), metabolomics (6 uses, 2%), metagenomics (2 uses, 1%) and others (24 uses, 7%). This distribution is heavily influenced by the TCGA (Tomczak, 2015) database, but the trend remains when excluding this database (see Supplementary Fig. S1).

Most papers we surveyed used three or more different -omics types (Fig. 2c). The -omics types that most often appeared together were transcriptomics and epigenomics, followed by transcriptomics and genomics. The top two combinations remained unchanged when we ignored the papers that relied on TCGA (see Supplementary Fig. S1).

3.1.2 How is the data structured: how many features versus samples?

There were far more features than samples in most cases (Fig. 3). The median number of features used in the surveyed publications was 33 415 while the median number of samples was 447. Due to outliers, these are different from the mean number of features, 73 966 and the mean number of samples, 1767. Most multi-omics ML method development research relies on existing data. TCGA was used in 73% of the surveyed papers. Creating multi-omics datasets has high costs in terms of money and time, but also requires a broad range of expertise not often found in a single research group. Besides this, the recent push for FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson, 2016) and the open science movement is helping to make data more easily accessible. Now researchers can tap into the community’s shared resources in order to supplement their own data, or simply to experiment with new iterations of ML techniques. Databases where biological data are uniformly processed make this process even easier and allow many different techniques to be tested, facilitating progress.

After splitting publications into those that relied on TCGA and those that relied on other datasets, we found no significant difference in the number of samples used, but we did observe a difference in the number of features (P < 0.01 using the Mann-Whitney U test).

3.2 What analysis has been done on multi-omics data?

3.2.1 Which ML techniques were used?

To build a perspective of the role of ML in multi-omics integration, we investigated how it differed from ML’s role in data integration in general. In this view, we gathered 100 more papers, this time searching for ‘ML AND Integration’, omitting the multi-omics term. With these and our initial set, we recorded which ML techniques were used in each paper.

In Figure 4a, we show the number of appearances of different ML techniques in the surveyed literature, omitting those that only
Using Fisher’s exact test, we tested for a statistically significant difference in the number of appearances of each technique in the multi-omics group versus the general ML group. We found such differences in autoencoders and Cox PH where these were more common in the multi-omics group (see Supplementary Fig. S3).

3.2.2 What were the goals of the ML application?
Classification (e.g. separating diseases into subtypes) was the most common goal in the reviewed papers (Fig. 4b). In some instances, this was used to discretize a regression problem. For example, several papers focused on survival prediction. Instead of predicting the number of years that a patient would survive (i.e. a regression task), the labels were ‘survival time $> 5$ years’ versus ‘survival time $< 5$ years’ (i.e. a classification task). Dimensionality reduction was the second most common goal, often applied before classification. Regression, network inference and denoising followed.

3.2.3 What were the targets/labels of classification tasks?
The dominant category of labels was survival prediction (Fig. 4c), followed by disease/patient/organism subtyping and response to intervention. Other labels included disease progression and classification.

3.3 Can we explain trends using an analysis of the citation of papers since 2015?
Figure 5 suggests that 2018 sparked an interest in this field, especially for autoencoders. In that year Chaudhary et al. (2018) was published and went on to become the most cited paper in this field. More details on this paper are included in the discussion.
Discussion

Multi-omics datasets often contain large numbers of features \((P)\) for a relatively small number of samples \((n)\). This has been described as the \(n \ll P\) problem (see Fig. 6b). This is opposite of the ideal situation for many ML applications, where a dataset with \(n \gg P\) is considered ideal (Fig. 6a).

Mathematical analysis becomes more challenging in high-dimensional spaces. A typical problem with high dimensionality (large number of features) is data ‘sparsity’. Considering each data-point as a coordinate, as the number of features increases (and therefore the number of data dimensions increases), the volume of the dimensional space spanned by the data points increases rapidly, such that the available data becomes sparse. This, in turn, makes inference and prediction particularly difficult unless large amounts of data points are available for analysis. This phenomenon is known as the ‘curse of dimensionality’, and the high heterogeneity of biological data amplifies this challenge for researchers.

We found that a vast share of ML multi-omic data integration approaches revolve around overcoming this ‘curse’. Researchers have to take steps to maximize the number of available samples and minimize the number of features. After that, they still tend to rely on models that are not too sensitive to having relatively few samples.

4.1 Reducing the number of features \((P)\)

To minimize \(P\), or the number of features, one can select a subset of more relevant features. Alternatively, one can use an algorithm that merges and transforms features in a smaller number of new ones, or one can apply a combination of the previous two approaches.

4.2 Feature selection

In computational modelling, any ability to leverage prior knowledge of the system into the model can be beneficial since appropriate prior assumptions can help to find an optimal model. This principle is often applied implicitly. For example, raw sequencing reads of RNAseq are not analysed as-is, but rather mapped onto gene transcripts as those have been defined in the latest genome assembly of a model organism. Similarly, the genome-wide signal of DNA methylation assays is typically discretized into a number of active/inactive methylation sites (Yuan et al., 2019), since such sites can be considered functional units of gene regulation.

Beyond this discretization of the raw data, biological expertise can offer critical advantages through feature selection (see Fig. 6c)
by discarding features with low probability to be relevant. For instance, Athreya et al. (2018) reduced over 7 million features consisting of SNPs and metabolites, to 65 predictor variables. They used a variety of reduction criteria including keeping SNPs that had a strong association with metabolite concentrations. Generally, some researchers will keep the most variable genes and discard those with consistently low activity levels. Epigenomic data can similarly be reduced by only considering the loci that are found near relevant genes or by looking at regions encompassing multiple methylation sites rather than looking at individual sites separately.

Feature selection can also be done using computational algorithms rather than biological expertise. One notably common application in multi-omics studies is to combine feature selection with survival analysis with the Cox PH model. The standard Cox PH model utilizes linear regression techniques for selecting informative features. This was the second most popular ML technique for multi-omics, but not so in the broader non-multi-omic field (Fig. 4a). The dominance of TCGA data, where survival analysis can often be applied, and the need of feature selection in the multi-omic field likely contributed to this disparity in popularity. Favourable characteristics of feature selection with Cox PH model applications include that the outcomes provide interpretable values representing the magnitude and direction of the effect of multiple features on survival. Furthermore, a degree of explainability is lost from the features as it is not always easy to explain what a principal component exactly is.

Another example of feature extraction are autoencoders. These neural networks run input data through a series of layers of a neural network, with one of these being a smaller ‘bottleneck’ layer. As an example, Chaudhary et al. (2018) used a network with three layers of 500, 100 and 500 units, respectively. This middle bottleneck layer produces the extracted features which can then be used for further analysis. An advantage of autoencoders is that by adding several layers with non-linear activation functions, the technique can model complex nonlinear functions. Autoencoders were more popular for multi-omic ML data integration than for general ML data integration.

Autoencoders take in the original inputs, compress them into a lower dimensional representation and then reconstruct the inputs minimizing the difference between original and reconstructed inputs (i.e. ‘reconstruction loss’). Denoising is particularly pertinent to multi-omics. Here ‘noise’ can be thought of as something that randomly changes pixel values of an image, so that the resulting image is corrupted. Denoising autoencoders (Vincent et al., 2008) add to
the autoencoder framework by corrupting (adding noise to) the original inputs. An example is Seal et al. (2020), which added a randomly drawn number to each input. The aim of this is that the model learns to create a lower-dimensional representation that retains the most important aspects of the original information.

The success of autoencoders may largely be thanks to the application of the backpropagation (Rumelhart et al., 1986) algorithm. This is an optimization algorithm that has proven widely effective in adjusting ML model parameters towards specific goals (e.g., minimizing reconstruction loss). In addition, autoencoders are relatively easy to use.

4.4 Increasing the number of samples (n)
An alternative to producing a problem-specific dataset and maximizing n by deeper sampling, is to turn to publicly available data. Motivations for using public data could be either lack of resources or because of different research interests. Researchers in the papers we surveyed overwhelmingly relied on public databases of multi-omics data, mainly, TCGA.

TCGA contains approximately 10 000 samples across 33 human cancer types (Tomczak, 2015) in a well maintained, sophisticated portal with an application programming interface (API). This high number of datasets combined with its ease of use have made it a favourite source for researchers. A possible limitation of the field’s reliance on this dataset is its focus on cancer. Cancer is characterized by highly dysregulated metabolism and aberrant signalling pathways (Cairns, 2011), which may lead to relatively easily identifiable -omics signals. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.

Multi-omic datasets often include non-overlapping data, i.e. where not all -omics data are obtained for each sample. This can lead to a significant decrease in viable samples, reducing statistical power for the analysis. For example, in Zhang (2018), data from 407 neuroblastoma patients were used. Of those, 380 had copy number alterations. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.

Multi-omic datasets often include non-overlapping data, i.e. where not all -omics data are obtained for each sample. This can lead to a significant decrease in viable samples, reducing statistical power for the analysis. For example, in Zhang (2018), data from 407 neuroblastoma patients were used. Of those, 380 had copy number alterations. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.

Multi-omic datasets often include non-overlapping data, i.e. where not all -omics data are obtained for each sample. This can lead to a significant decrease in viable samples, reducing statistical power for the analysis. For example, in Zhang (2018), data from 407 neuroblastoma patients were used. Of those, 380 had copy number alterations. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.

Multi-omic datasets often include non-overlapping data, i.e. where not all -omics data are obtained for each sample. This can lead to a significant decrease in viable samples, reducing statistical power for the analysis. For example, in Zhang (2018), data from 407 neuroblastoma patients were used. Of those, 380 had copy number alterations. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.

Multi-omic datasets often include non-overlapping data, i.e. where not all -omics data are obtained for each sample. This can lead to a significant decrease in viable samples, reducing statistical power for the analysis. For example, in Zhang (2018), data from 407 neuroblastoma patients were used. Of those, 380 had copy number alterations. This may not be congruent with other fields, such as agriculture or nutrition where small interventions are made (e.g. changing feed), and which may generate more subtle -omics signals (Edmunds et al., 2012). Another disadvantage of the widespread use of this database is that it does not contain certain types of omics, such as ATACseq. If the multi-omics community relied too heavily on TCGA, it may bias future research towards the -omics data types contained in the TCGA dataset. On the other hand, Figure 3 shows that TCGA and other datasets are fairly comparable (although statistically different in number of features), TCGA could be generally representative of other datasets.
their analysis. Typical steps included manual feature selection, followed by algorithmic selection and/or extraction. Autoencoders and Cox PH were commonly used. Causal models, graph neural networks and transformers are some promising approaches for the field.

ML was most often used in classification problems, where tools like random forests and SVM were most commonly used. Their low barrier to entry, adaptability to many kinds of data types, and ability to work with relatively few samples are possible reasons for their popularity. These are both not new techniques, which may suggest that there is a gap for classification techniques that cope well with highly dimensional data and a low number of samples.

The dominance of TCGA as a source of data in the multi-omics ML integration field, highlights the impact that such a database can have. Datasets that are easy to discover and use propel innovation by allowing fast iterations over different techniques. Curating datasets and goals for ML competitions could accelerate this even further. Perhaps a good direction for the field would be to produce sets and goals for ML competitions could accelerate this even further by allowing fast iterations over different techniques. Curating data-sets and goals for ML competitions could accelerate this even further. FindingPheno: D.F.-B., P.F.N. and G.D.P. were also supported by the search and innovation programme [952914] (FindingPheno). In addition to findingPheno, D.F.-B., P.F.N. and G.D.P. were also supported by the search and innovation programme [952914] (FindingPheno).

Funding
This project received funding from the European Union’s Horizon 2020 research and innovation programme [952914] (FindingPheno). In addition to findingPheno: D.F.-B., P.F.N. and G.D.P. were also supported by the Champalimaud Foundation; F.R. was also supported by a VUllum Fonden Young Investigator award [project no. 00025300]; and M.T.L. and S.G. were supported by the Danish National Research Foundation [CEH-DNRF143].

Conflict of Interest: none declared.

References


