Using machine learning approaches for multi-omics data analysis

Reel, Parminder S.; Reel, Smarti; Pearson, Ewan; Trucco, Emanuele; Jefferson, Emily

Published in:
Biotechnology Advances

DOI:
10.1016/j.biotechadv.2021.107739

Publication date:
2021

Using machine learning approaches for multi-omics data analysis: A review

Parminder S. Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, Emily Jefferson

PII: S0734-9750(21)00045-8

DOI: https://doi.org/10.1016/j.biotechadv.2021.107739

Reference: JBA 107739

To appear in: *Biotechnology Advances*

Received date: 15 December 2020

Revised date: 1 March 2021

Accepted date: 25 March 2021

Please cite this article as: P.S. Reel, S. Reel, E. Pearson, et al., Using machine learning approaches for multi-omics data analysis: A review, *Biotechnology Advances* (2021), https://doi.org/10.1016/j.biotechadv.2021.107739

Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review

Parminder S. Reel[1,1], Smarti Reel[1,1], Ewan Pearson[1], Emanuele Trucco[2], Emily Jefferson[1,*]
e.r.jefferson@dundee.ac.uk
[1]Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, United Kingdom
[2]VAMPIRE project, Computing, School of Science and Engineering, University of Dundee, Dundee, United Kingdom
[*]Corresponding author at: Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK.

Abstract

With the development of modern high-throughput omic measurement platforms, it has become essential for biomedical studies to undertake an *integrative* (combined) approach to fully utilise these data to gain insights into biological systems. Data from various omics sources such as genetics, proteomics, and metabolomics can be integrated to unravel the intricate working of systems biology using machine learning-based predictive algorithms. Machine learning methods offer novel techniques to integrate and analyse the various omics data enabling the discovery of new biomarkers. These biomarkers have the potential to help in accurate disease prediction, patient stratification and delivering of precision medicine. This review paper explores different integrative machine learning methods which have been used to provide an in-depth understanding of biological systems during normal physiological functioning and in the presence of a disease. It provides insight and recommendations for interdisciplinary professionals who envisage employing machine learning skills in multi-omics studies.

Keywords: Multi-omics, Machine Learning, Predictive Modelling, Supervised Learning, Unsupervised Learning, Systems Biology

List of Abbreviations

ATHENA	Analysis Tool for Heritable and Environmental Network Associations
BCC	Bayesian consensus clustering
BN	Bayesian Network
CS	Concatenation-based Supervised Learning
CU	Concatenation-based Unsupervised Learning
DNA	Deoxyribo-Nucleic Acid
FCA	Formal Concept Analysis
FDA	Food and Drug Administration
fMKL-DR	fast multiple kernel learning for dimensionality reduction
FSMKL	Multiple Kernel Learning with Feature Selection
HI-DFNForest	Hierarchical integration deep flexible neural forest
JBF	Joint Bayes Factor
JIVE	Joint and Individual Variation Explained
KNN	k-nearest neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
lncRNAs	long non-coding RNAs
MDI	Multiple Dataset Integration
MDS	Multi-Dimensional Scaling
Meta-SVM	Meta-analytic SVM

[1]These authors contributed equally to this study (shared first authorship).

miRNA	microRNA
ML	Machine Learning
MOFA	Multi-Omics Factor Analysis
MOLI	Multi-omics late integration
MORONET	Multi-Omics gRaph cOnvolutional NETworks
MOSAE	Multi-omics Supervised Autoencoder
mRNA	messenger Ribo-Nucleic Acid
MS	Model-based Supervised Learning
MU	Model-based Unsupervised Learning
NEMO	NEighborhood based Multi-Omics clustering
NMF	Non-negative Matrix Factorisation
PCA	Principal Component Analysis
PINS	Perturbation clustering for data integration and disease subtyping
PSDF	Patient-Specific Data Fusion
RF	Random Forest
rMKL-LPP	regularised multiple kernel learning for Locality Preserving Projections
RVM	Relevance Vector Machine
SDP-SVM	Semi-Definite Programming SVM
SmSPK	smoothed shortest path graph kernel
SNF	Similarity Network Fusion
SSL	Semi-supervised learning
SVM	Support Vector Machine
SVR	Support vector regression
TS	Transformation-based Supervised Learning
TU	Transformation-based Unsupervised Learning

1 INTRODUCTION

Digital information is growing rapidly, in terms of five V's (volume, velocity, veracity, variety and value), and hence this is hailed as the *big data* era (BCS, 2014; Bellazzi, 2014; Lee and Yoon, 2017). Health-based big data including linked information for patients, such as their clinical data (for example gender, age, pathological and physiological history) and omics data (such as genetics, proteomics and metabolomics) has now become more widely available (Canuel et al., 2015; Singhal et al., 2016). Recently, such data has been used for precision (also called personalised or stratified) medicine to provide customised healthcare, i.e. providing a bespoke treatment for individuals (Gibson et al., 2015; Kalaitzopoulos, 2016; Malod-Dognin et al., 2017). There has been unprecedented growth in the development of precision medicine supported by ML (machine learning) approaches (Delavan et al., 2017; Peterson et al., 2013; Zou et al., 2017) and data mining tools (Chawla and Davis, 2013; Cheng et al., 2015; Margolies et al., 2016). These techniques have also helped to discover novel omics biological markers which can identify the molecular cause of a disease.

A biomarker is a substance, structure, or process that can be measured in the human body or its products and can provide surrogate information about the presence of a disease/condition (Strimbu and Tavel, 2010). Molecular biomarkers are discovered by analysing the cascade of information provided by different omics (Debnath et al., 2010). For example, the high-sensitivity C-Reactive protein test provides an accurate and quantitative risk assessment for cardiovascular disease (Pfützner and Forst, 2006; Shrivastava et al., 2015). Biomarkers play a significant role in planning preventive measures and decisions for patients (Nielsen, 2017) and can be classified as either *diagnostic*, *prognostic* or *predictive* (Le et al., 2016; Shaw et al., 2015). *Diagnostic biomarkers* are used for determining the presence of disease in a patient, while *prognostic biomarkers* provide information on the overall outcome with or without the standard treatment (Carlomagno et al., 2017). *Predictive biomarkers* are used to identify who is at risk of an outcome (Nalejska et al.,

2014). All of these biomarkers can also be used to identify which treatment will be most suitable for a given patient. For example, the ADNI (Alzheimer's Disease Neuroimaging Initiative) study used a combination of neuroimaging, biochemical and genetic biomarkers to discriminate early Alzheimer's patients from healthy volunteers with an accuracy of 98% (Gupta et al., 2019). Similarly, different forms of Parkinson's syndromes have been investigated by developing an automated tool which fuses multi-site diffusion-weighted MRI imaging biomarkers and disease rating score (MDS-UPDRS III) (Archer et al., 2019). Biomarkers can help identify high-risk individuals before their physiological symptoms are evident. Moreover, they also help in measuring disease progression (Mandel et al., 2010).

In the context of precision medicine, ML has been used to develop diagnostic, prognostic and predictive tools from single omics data (Dias-Audibert et al., 2020; Mamoshina et al., 2018; Sonsare and Gunavathi, 2019). However, ML may have deteriorated performance for certain single omics such as gene data due to inherent characteristics (Kim et al., 2020). ML methods are now also being applied to with multi-omics data (Bersanelli et al., 2016), to investigate and interpret the relationships between data and phenotypes (Kim and Tagkopoulos, 2018). Although ML analysis of multi-omics is still in its embryonic stage, it has already been explored for a wide range of applications, as reported in recent reviews on brain diseases (Garali et al., 2018; Young et al., 2013), diabetes (Kavakiotis et al., 2017), cancers (Borad and LoRusso, 2017; Chaudhary et al., 2017; Wong et al., 2016) cardiovascular disease (Weng et al., 2017), medical imaging (Erickson et al., 2017), single-cell analysis in humans (Cao et al., 2020; A. Ma et al., 2020) and plant science studies (Acharjee et al., 2011). Currently, many of the multi-omic reviews are focussed on individual sub-topics. For example, designing studies (Haas et al., 2017; Hasin et al., 2017), setting up workflows (Kohl et al., 2014), choosing software tools (Misra et al., 2019) and evaluating overfitted performance (McCabe et al., 2020).

In contrast, this review aims at a broader focus, presenting an interdisciplinary perspective to new readers in this domain by providing a background on multi-omics and ML. It takes forward the integration terminologies introduced by Ritchie (Ritchie et al., 2015) and summarises the recent integrative state-of-the-art approaches. We aim to cover various integration methods concisely and include a recommendation flowchart enabling interdisciplinary scientists to have a quick head start in this domain (Bersanelli et al., 2016; Nguyen and Wang, 2020).

Scope of this review: This review investigates the two primary learning strategies in ML, i.e. supervised and unsupervised, which are commonly used within the context of multi-omics integration. This review considers *multi-omics integration* as a process of combining different single omics. Although various ML specialisations such as reinforcement (Coronato et al., 2020), hybrid (Zhou et al., 2019), multi-view (Zhao et al., 2017) and self-supervised learning (Chen et al., 2019) are now emerging in generic health-care applications, they have not yet gained enough momentum in multi-omics analysis, hence they remain beyond the scope of this review.

This paper is organised as follows. Section 2 provides a short background related to multi-omics and ML. Section 3 describes how ML is employed for multi-omics analysis and what are the various real-world challenges of it. In Section 4, details of different multi-omics integration approaches are presented. Section 5, published multi-omics studies using ML methods are discussed. Section 6 describes a recommendation flowchart for choosing an appropriate method for multi-omics integration. Conclusions are provided in Section 7.

2 BACKGROUND

2.1 Multi-omics

In living beings, genetic information in the cells flows from DNA (deoxyribo-nucleic acid) to the mRNA (messenger ribo-nucleic acid) to protein and is dictated by the central dogma of molecular biology (Lodish et al., 2000). This flow of information is often considered analogous to a computer system which has facilitated the understanding of biological information processing (Wang and Gribskov, 2005; D'Onofrio and An, 2010).

The study of DNA, mRNA and proteins is broadly denoted as genomics, transcriptomics, and proteomics respectively. The genetic blueprint of a cell is explored using genomics, which looks at the DNA of individuals and helps us to investigate the presence or absence of certain genes (Gibson, 2015; Vogel and Motulsky, 1997). Transcriptomics studies the transcribed genetic material and examines the genes which are actively expressed and provides information about what is happening at the cellular level (Milward et al., 2016). Proteomics helps in characterising the information flow happening within the cell and the organism in the form of protein pathways and their networks (Wu et al., 2014).

Although metabolomics, lipidomics, and glycomics do not form part of the central dogma analysis (Cobb, 2017), they still provide an invaluable amount of information regarding the metabolites, lipids and glycans (synthesised by the proteome via biosynthetic pathways) (Barh et al., 2011). These substances are the intermediate products of a cell's information flow and therefore are considered to be excellent indicators of the cell's activity. Similar to single-genome studies, metagenomics is used to sequence genetic information from environmental samples without the requirement of isolating individual species (Hugenholtz and Tyson, 2008).

All measured omics data can be used as a biomarker which helps us to understand and analyse the underlying characteristics and complexities of biological systems (Alberts et al., 2008). Table 1 shows some of the important omics used to study biological systems (Handelsman et al., 1998; Kuska, 1998; Lindon et al., 2011; "Proteomics, transcriptomics," 1999; Vasta and Ahmed, 2008, Wang et al., 2016; Wilkins and Appel, 2007). All of them are part of the same pipeline of biological information, whose output depends on the different inputs and regulation. As shown in Table 1, each of these omics can be measured using specialised high-throughput technologies (for example microarray (Bumgarner, 2013) and mass spectrometry (Glaves et al., 2014) for genomics and metabolomics respectively). The table also includes a list of recent reviews on each of these omics. High-throughput generated omics data (Lightbody et al., 2019) has played a pivotal role in developing precision medicine biomarkers for diseases such as Alzheimer's (Hampel et al., 2017, 2016; Kovacs, 2016), diabetes (Capobianco, 2017; McCarthy, 2017; Mutie et al., 2017), cancer (Borad and LoRusso, 2017; Senft et al., 2017), hypertension (Barnes et al., 2016; Dominiczak et al., 2017), cardiovascular (Costantino et al., 2017) and chronic respiratory diseases (Agache and Rogozea, 2017; Hanania and Diamant, 2017). Recently, these omics have also been integrated for COVID-19 studies (Barh et al., 2020; Overmyer et al., 2020; Zhou et al., 2020). Many other specialised omics have also emerged such as pharmacogenomics (Wang, 2010), methylomics (Liu et al., 2013), interactomics (Luck et al., 2017) and radiomics (Lambin et al., 2017; Wong et al., 2016).

Overall, these omics provide a complete picture of cell biology and related cellular function (Cox, 2009). This provided the impetus for the development of various software

mechanisms which can offer a prediction of a particular phenotype while using the available next-generation multi-omics data (Ritchie et al., 2015). Furthermore, they can be utilised to develop materials and devices which be used for the diagnostic and preventive purpose at the molecular level while targeting molecules with greater accuracy (Giovanni Martinelli et al., 2015).

Table 1: The omics technologies which help us draw a complete picture of the cell biology and related function.

S. No	Omic name	Term coined in	Data extracted	Commonly used High-throughput technologies	Common Reference databases	Recent reviews
1	Genomics	1986 (Kuska, 1998)	Single nucleotide polymorphisms, Rare variants and Copy number variations.	DNA-Sequencing (Sanger (Sanger et al., 1977), Whole genome (J. Huang et al., 2017), Whole-exome (Weisz Hubshman et al., 2018), Single-Cell DNA (L. Zhang et al., 2019) and targeted sequencing (Bewicke-Copley et al., 2019)), Microarray (Bumgarner, 2013).	DDBJ (Tateno et al., 2002), GenBank (Benson et al., 2011), ENA (Leinonen et al., 2011)	(Reuter et al., 2015)
2	Transcriptomics	1999 ("Proteomics, transcriptomics," 1999)	Messenger, Micro and Long non-coding RNA expression.	RNA-Sequencing (Sanger (Alidjinou et al., 2017), Single-Cell RNA (Huang et al., 2018) and targeted sequencing (Mercer et al., 2012)), Microarray (Zhao et al., 2014).	miRBase (Kozomara et al., 2019), Rfam (Kalvari et al., 2018)	(Lowe et al., 2017)
3	Proteomics	1994 (Wilkins and Appel, 2007)	Protein expression	Reverse Phase Protein Array (Boellner and Becker, 2015), Liquid Chromatography - Mass Spectrometry (Karpievitch et al., 2010) and Mass Spectrometry (Timp and Timp, 2020)	HPA (Uhlen et al., 2010), PDB (Burley et al., 2019), Pfam (Finn et al., 2010), UniProt (The UniProt Consortium, 2019)	(Aslam et al., 2017)
4	Metabolomics	2001 (Lindon et al., 2011)	Metabolite expression	Mass Spectrometry (Glaves et al., 2014), Liquid Chromatography - Mass Spectrometry (Zhou et al., 2012), Gas Chromatography - Mass Spectrometry (Fiehn, 2016).	HMDB (Wishart et al., 2018), KEGG (Kanehisa and Goto, 2000)	(Zampieri et al., 2017)
5	Lipidomics	2003 (Wang et al., 2016)	Lipids	Liquid Chromatography - Mass Spectrometry (Li et al., 2020), High-performance Liquid Chromatography - Mass Spectrometry (Knittelfelder et al., 2014) and Direct-Infusion/Shotgun - Mass Spectrometry (Köfeler et al., 2012).	LMSD (Sud et al., 2007), LipiDAT (Caffrey and Hogan, 1992), LipidBank (Watanabe et al., 2000), LipidHome (Foster et al., 2013), LipidPedia (Kuo and Tseng, 2018),	(Yang and Han, 2016)
6	Glycomics	1990 (Vasta and Ahmed, 2008)	Glycomes	Matrix-Assisted Laser Desorption/Ionization Time-of-Flight - Mass Spectrometry (Y. Zhang et al., 2019).	GlyTouCan (Tiemeyer et al., 2017), UniCarb-DB (Campbell et al., 2014)	(Rojas-Macias et al., 2019)
7	Metagenomics	1998 (Handelsman et al., 1998)	Genetic data from environmental (soil, water) samples.	Target Gene Sequencing, Shotgun Metagenome Sequencing, Metatranscriptome Sequencing (Zhou et al., 2015)	MG-RAST (Meyer et al., 2008), SRA (Kodama et al., 2012), MGnify (Mitchell	(Pérez-Cobas et al., 2020)

et al., 2020)

2.2 Machine learning

Classical statistical modelling has always been the *de facto* standard choice for health data analysis and its interpretation. In recent years, with the increasing availability of affordable computing power and high-throughput omics data and the success of artificial intelligence technology in various fields, the use of ML has become popular in health sciences (Lee and Yoon, 2017; Clifton et al., 2015; Hung, 2019; Barnett-Itzhaki et al., 2020; Kirchebner et al., 2020). ML can be used to mine information hidden in the experimental data. In contrast, a conventional statistics-based model is usually developed using the statistical assumptions and draws an inference about a population from a given dataset (Bzdok, 2017).

The objective of ML methods is to acquire knowledge from historic or present-day data and utilise that understanding to make forecasts or choices for unidentified forthcoming data measures (Gammerman, 2010; Obermeyer and Emanuel, 2016). To assist beginners in the ML domain, a glossary of learning approaches covered in this review (Table 2), standard ML terminology (Table 3) and commonly used ML algorithms (Table 4) are provided. The basic foundations of ML and its uses have been extensively covered in literature (Bishop, 2006).

ML is employed in a wide range of scenarios, where designing and programming of explicit algorithms with optimal results is challenging, such as email filtering (Dada et al., 2019), hand-written optical character recognition (Memon et al., 2020), and computer vision (O'Mahony et al., 2020). Also, it has been deployed for self-driving cars (Badue et al., 2021), cyber-security (Handa et al., 2019), automated assistants such as 'Siri', websites that recommend items based on the purchasing decisions of other people and novel solutions to some of the challenging problems of the real world (Watt et al., 2020).

Deep learning has emerged in recent years as the leading class of ML algorithms. It uses neural networks composed of hidden layers performing different operations to find complex representations of data. It has pushed the performance of classifiers beyond that of traditional ML algorithms, especially in scenarios involving large-scale datasets with high dimensionality. On the other hand, it is very computationally intensive, requiring high-throughput or high-performance hardware, and lacks explainability (transparency) in feature selection (*black-box* approach), in the sense that it is difficult to extract from the network the features that the network has found as mainly responsible for the task, e.g. classification (LeCun et al., 2015). However, in the context of multi-omic integration, deep learning offers an exciting opportunity.

Table 2: The different ML learning approaches reviewed for multi-omics integration.

Learning approach	Goal	Description
Supervised	Predict new data	Supervised learning involves fitting a model with labelled training data and then use it for prediction. It can be classed either as a regression (predicted variable is numeric) or classification (predicted variable is categorical) problems (Jiang et al., 2020). The three steps in supervised learning are: (1) fitting a model from the sample input observations (2) evaluating the model and then extensively tuning the hyper-parameters of the model (3) setting up the model for the

		production stage and using it for prediction (Foster et al., 2014).
Unsupervised	Identify clusters	Unsupervised learning is used to find the underlying patterns in unlabelled data using input feature variables without the target/output variable (Badillo et al., 2020). It can be used for clustering (Xu and Tian, 2015), anomaly detection (Thudumu et al., 2020) and dimensionality reduction (X. Xu et al., 2019).

Table 3: The standard ML terminology and related terms.

Term	Definition
Accuracy	It is a ratio of correctly predicted outcomes of a given class to the total outcomes. Accuracy is a measure of the performance of an ML model. It ranges from 0% to 100%.
Classification	It is a supervised learning method which provides predicted output as a discrete class. Classification can be binary, multi-class or multi-label.
Clustering	It is an unsupervised learning method which can group data based on the attributes of the input features.
Cross-Validation	It is a technique which allocates a given set of samples from the dataset which are not used for model training but set aside for testing (to evaluate model performance). K-fold and Leave one out are commonly used cross-validation methods.
Curse of dimensionality	It refers to a set of problems which arise when using datasets with high dimensionality. In the context of ML, it can impact the predictive performance of an ML model (Duda et al., 2001).
Dataset	It is a collection of structured data which comprises of input feature variables and sometimes a corresponding target/output variable.
Ensemble Learning	It is a paradigm where different models are trained for solving the same problem and then combined to get better performance. Bagging, boosting and stacking are commonly used in ensemble learning methods.
Explainability	Supervised learning models can be classed as 'white' or 'black-box' based on their explanation (or lack thereof) of how a decision is reached. This is a growing and important domain of research in deep learning.
Feature Selection	It is a process for selecting the most discriminating features without impacting the classification performance.
Hyper-parameter	It is an empirically tuned internal parameter of an ML model.
Imputation	It is a process of replacing missing values in a dataset with a corresponding statistical estimate. Imputation can be done using mean, median values or employing methods such as KNN (Crookston and Finley, 2008) or MICE (Azur et al., 2011)
Outlier	It is an extremely low or high value of a feature in a dataset (based on the range and distribution). Performance of ML algorithms is sensitive to outliers, hence their detection and exclusion are crucial (Domingues et al., 2018)
Performance Metric	It is a method to evaluate and compare the performance of ML models. For example, precision, recall/sensitivity, specificity, F1 score, Kappa and mean absolute error.
Regression	It is a supervised learning method which provides predicted output as a continuous value.
Training	It is a first step in the learning process which uses training dataset to fit the parameters of a supervised ML model.
Testing	It is a second step in the learning process which uses testing dataset (independent of the training dataset) to assess the predictive performance of a trained supervised ML model.
Bias-variance trade-off	In order to achieve optimal prediction performance, a supervised model should ideally have low bias and low variance. A model is over or underfitted when trade-off is not achieved.

Table 4: The commonly used ML algorithms and their attributes. The rank [1 – Low, 2 – Medium, 3 – High, 4 – Very High] denoted to attributes is pragmatically assigned based on available literature (Amancio et al., 2014; Barredo Arrieta et al., 2020; de Andrade et al., 2020; Lorena et al., 2011; Rashidi et al., 2019; Sakr et al., 2017).

Family	Models	Comparative Accuracy	Overfitting Risk	Samples needed	Explainability	Hyper-parameter Tuning	Complexity	Implementation Time	Computation Cost

Probability-based (Bayesian)	Bayesian Network	2	2	2	2	3	3	2	3
	Naive Bayes	2	2	2	2	2	3	2	3
Information based (Tree)	Decision Tree	2	3	2	3	2	2	1	2
	Random Forest	3	2	1	3	3	2	1	2
	Gradient Boosting	3	3	2	1	4	4	2	3
Error based (Linear)	Linear Regression	1	3	2	3	1	2	1	2
	Logistic Regression	1	3	2	3	1	2	1	2
	Partial Linear Regression	2	1	3	3	2	2	1	2
Similarity-based (Instance)	K nearest neighbour	2	3	2	2	2	3	1	1
	Self-Organising Maps	2	3	2	2	3	3	1	1
Support Vectors	Linear SVM	3	3	3	1	3	2	2	2
	Non-linear (Kernel) SVM	3	3	3	1	3	3	3	3
Neural Network-based	Artificial Neural Network	3	3	2	1	3	3	3	3
	Deep Learning (Neural Network)	4	1	4	1	4	4	4	4

Figure 1 shows the number of publications indexed on the *Web of Science* website (Clarivate Analytics, 2020) with different key topics. This information was collected from the Web of Science by entering a different keyword and searching across all databases[2]. Although the use of ML in medical science can be dated back to 1970s (Davenport and Kalakota, 2019), more rapid growth is evident in the past 10 years. Moreover, publications based on 'multi-omics integration' and 'multi-omics and machine learning' have started to emerge in the last 5 years and have gained popularity in precision and computational medicine domain. Although deep learning is widely popular in other related domains (such as medical imaging (Erickson et al., 2017) and clinical natural language processing (Wu et al., 2020)), the interest has been more limited for multi-omics analysis (X. Tan et al., 2020). This is because multi-omics studies are challenging to deploy as they require specialised high-throughput omic infrastructure (as highlighted earlier in section 2). This fact is reinforced by the evidence that most of the current literature employs deep learning on large-scale multi-omics datasets from open sources such as TCGA[3], CCLE[4] and GDSC[5] for cancer prognosis (Poirion et al., 2018; Seal et al., 2020; Tong et al., 2020; Lee et al., 2020; Zhu et al., 2020) and anti-cancer drug response (Sharifi-Noghabi et al., 2019; Li et al., 2019; Deng et al., 2020).

3 CHALLENGES IN MULTI-OMICS ANALYSIS USING MACHINE LEARNING

The use of ML to analyse high-throughput generated multi-omic data poses key unique challenges. They can be summarised as follows.

3.1 Heterogeneity, sparsity and outliers

[2] All Web of Science databases included: Web of Science Core Collection, BIOSIS Citation Index, BIOSIS Previews, Current Contents Connect, Data Citation Index, Derwent Innovations Index, KCI-Korean Journal Database, MEDLINE®, Russian Science Citation Index, SciELO Citation Index and Zoological Record.
[3] TCGA: The Cancer Genome Atlas
[4] CCLE: Cancer Cell Line Encyclopaedia
[5] GDSC: Genomics of Drug Sensitivity in Cancer

Multi-omic data from different high-throughput sources are usually heterogeneous (Bersanelli et al., 2016). For example, transcriptomics and proteomics use different normalisation and scaling techniques before omics analysis. This leads to different dynamic ranges and data distribution. Also, some omics are more prone to generating sparse data (e.g. in the case of metabolomics, some values might be below the limit of detection and hence assigned null value (Antonelli et al., 2019)) than others. Therefore, imputation (Liew et al., 2011) and outlier detection (Vivian et al., 2020) should be considered for each omic separately, before planning their integration.

3.2 Class imbalance and overfitting

In disease classification, certain disease classes are rarer than others which can cause a class imbalance in the multi-omics dataset (Haas et al., 2017). For example, primary hypertension is the most common form of hypertension with 95% prevalence while endocrine hypertension occurs in only 5% (Rimoldi et al., 2014). The ML model trained using an imbalanced dataset may be overfitted i.e. high accuracy for training data but underperformance for unseen test data. Therefore, to classify these two types of hypertension one of the following approaches can be used: 1) Collect more data if possible, or 2) consider using weighted or normalised metrics to measure the ML performance (such as F1-Score or Kappa (Jeni et al., 2013)), or 3) consider over or under-sampling the under or over-represented class respectively, or 4) consider synthetic sample generation (such as SMOTE (Chawla et al., 2002) or ADASYN (Haibo He et al., 2008)) for the under-represented class. Similarly, techniques such as regularisation, bagging, hyperparameter tuning and cross-validation can used to balance bias-variance trade-off (Lee, 2010). Any of the above approaches can be used, depending on data and problem, to overcome the class imbalance and overfitting problems.

3.3 More features than data (p >> n)

Most multi-omics datasets suffer from the classical '*curse of dimensionality'* problem, i.e. having much fewer observation samples (*n*) than multi-omics features (*p*) (Misra et al., 2019). The resulting high-dimensional space often contains correlated features which are redundant and can mislead the algorithm training (James et al., 2017). The dimensional space of the data can be reduced by employing dimensionality reduction techniques such as *feature extraction* and *feature selection*. Feature extraction refers here to techniques computing a subset of representative features which summarise the original dataset and its dimensions[6]. These features are *functions of the original ones*, for instance, PCA (principal component analysis) (Jolliffe, 2002), LDA (linear discriminant analysis) (Martinez and Kak, 2001) and MDS (multidimensional scaling) (Young and Hamer, 1987). On the other hand, feature selection finds a subset of the *original* features that maximise the accuracy of a predictive model (Guyon and Elisseeff, 2003). It can be based on a prior knowledge i.e. evident from known literature or based on a database such as a Biofilter (Bush et al., 2009). Formally, feature selection methods can be classed as *filter* (Information gain (Roobaert et al., 2006), ReliefF (Beretta and Santaniello, 2011), Chi-square statistics (Lee et al., 2011)), *wrapper* (Recursive feature elimination (Guyon et al., 2002), Sequential feature selection (Pudil et al., 1994)) and *embedded* (such as LASSO (Least Absolute Shrinkage and

[6] We note that "feature extraction" has a different meaning in image processing and computer vision.

Selection Operator) (Zou, 2006)) techniques. Xu et al.(X. Xu et al., 2019) and Stańczyk (Stańczyk and Jain, 2015) provide an excellent resource for understanding and exploring the use of different dimensionality reduction techniques in the generic ML domain. Meng (Meng et al., 2016b) offers a review of these methods from the perspective of multi-omics data analysis.

3.4 Computation and storage cost

The use of ML for multi-omics analysis comes with computational and data storage cost (Herrmann et al., 2020). Most ML algorithms require high computation power and large volumes of storage capacity to save the logs, results and analysis. In recent years, ML models can be deployed on dedicated graphics processing units (Schmidhuber, 2015) and cloud computing platforms (Armbrust et al., 2010) such as Amazon EC2 ("Amazon EC2," n.d.), Microsoft Azure ("Cloud Computing Services | Microsoft Azure," n.d.) and Google Cloud Platform ("Cloud Computing Services," n.d.). The related costs should be considered well in advance before planning an ML-based multi-omics workflow.

3.5 What algorithm works best for what conditions?

The commonly used ML algorithms have different attributes (Table 4) and therefore it is crucial to choose an appropriate algorithm for the multi-omics analysis. In the literature, many reviews cover the key strengths and weaknesses of different ML algorithms using single omics (Amancio et al., 2014; López Pineda et al., 2015; Sakr et al., 2017; Uddin et al., 2019) and multi-omics (Ma et al., 2016; Francescatto et al., 2018; J. Xu et al., 2019; Sathyanarayanan et al., 2020) datasets. Most of them use a systematic workflow which involves simultaneous performance evaluation of different algorithms using a common dataset. Since each multi-omics dataset is unique, using a similar workflow could allow the selection of the best-suited algorithm. Later, in Section 6 a recommendation flowchart is proposed which can help the inter-disciplinary user to choose from available methods.

Recently, various artificial intelligence-driven automated ML platforms and tools (Feurer et al., 2015; Olson et al., 2018; Waring et al., 2020) have also emerged which can be utilised to exhaustively search for best ML model and corresponding parameter tuning, however, they are computationally expensive.

3.6 Translating ML: bench to bedside

Various ML based multi-omics publications have emerged in the past 5 years (see Figure 1) and some use performance metrics such as decision curve (Vickers and Elkin, 2006) and calibration (Dankers et al., 2019) analytics to evaluate their diagnostic utility. Still, only very few have been translated into clinical practice, for example, *Idx* (diabetic retinopathy detection), *FerriSmart* (measure liver iron concentration) and *SubtleMR* (image processing software for radiology) (Benjamens et al., 2020; Hamamoto et al., 2020).

One of the key issue which hinders the clinical deployment of ML methods is *transparency* and *explainability* (*Black box medicine and transparency*, 2020). A transparent and explainable ML algorithm seems essential to build trust for clinical decision making (Gunning et al., 2019). Recently, the U.S. Food and Drug Administration (FDA) have issued the "*Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device* Action Plan" to ensure deployment of ML based products is safe for patients to better assist the health care providers (Health, 2021). A recent in-depth analysis by Muehlematter showed most of the FDA approved and "*Conformité Européenne"* marked ML products are

in the field of radiology. It also highlighted the key differences between U.S. and European policy implications around the approval of AI/ML-based devices (Muehlematter et al., 2021).

All the above challenges directly impact the use of ML for multi-omics analysis. However, there are few other challenges related to multi-omics studies which are not ML related such as study design (Haas et al., 2017), multi-site sample collection & management (Pinu et al., 2019), multi-site data sharing and governance (Saulnier et al., 2019), visualisation (Mougin et al., 2018), ethical standards (Lévesque et al., 2018), and finally making the research reproducible (Conesa and Beck, 2019) and translational (Schumacher et al., 2014). A broader checklist of criteria is investigated by McShane while focussing on various aspects ranging from specimen requirements, predictive model development to clinical trial designing and related regulatory approvals (McShane et al., 2013b, 2013a).

4 DATA INTEGRATION METHODS FOR MULTI-OMICS

In recent years, various new data integration methods have been introduced from the modern developments in mathematical, statistical and computational sciences. For the benefit of the readers, Table 5 includes a summary of a few reviews which cover the breadth of multi-omics integration for generic as well as specialised domains such as oncology (Buescher and Driggers, 2016; Nicora et al., 2020) and toxicology (Canzler et al., 2020). Most of these reviews have strived to introduce different categorical terminologies (for example: ‘’early’, ‘late’ and ‘intermediate’ in (Gligorijević and Pržulj, 2015) or ‘bottom-up’ and ‘top-down’ in (Yu and Zeng, 2018)) which enable them to group the integration methods based on different factors/parameters.

Table 5: Summary table of few reviews in multi-omics integration

Year of review	Review Reference	Terminology introduced for classifying various integration methods	Omics reviewed									Application domain covered
			Genomics	Transcriptomics	Metabolomics	Proteomics	Epigenomics	Interactomics	Metagenomics	Lipidomics	Phosphoproteomics	
2009	(Van Deun et al., 2009)	matrix decomposition			✓							Micro-organism (Escherichia coli)
2009	(Ebbels and Cavill, 2009)	‘conceptual’, ‘statistical’ & ‘model’			✓	✓						Generic
2012	(Lussier and Li, 2012)	‘cross-scale’ & ‘multi-scale’	✓	✓								Prediction of clinical outcomes.
2015	(Ritchie et al., 2015)	‘concatenation’, ‘transformation’ & ‘model’	✓	✓								Generic
2015	(Gligorijević and Pržulj, 2015)	‘early’, ‘late’ & ‘intermediate’	✓	✓								Generic
2016	(Bersanelli et al., 2016)	‘sequential’, ‘simultaneous’, ‘network-based versus network-free’ & ‘Bayesian vs non	✓	✓		✓						Generic
2016	(Gligorijević et al., 2016)	-	✓	✓			✓					Disease subtyping, biomarkers discovery & drug repurposing.
2016	(Buescher and Driggers, 2016)	-	✓	✓	✓	✓						Cancer biology

2017	(Lin and Lane, 2017)	-	✓	✓	✓	✓						Investigated (Ritchie et al., 2015) from ML perspective
2017	(S. Huang et al., 2017)	-	✓	✓		✓						Patient survival prediction
2017	(Hasin et al., 2017)	'genome', 'phenotype' & 'environment-' first approach	✓	✓	✓	✓	✓					Generic
2018	(Yu and Zeng, 2018)	'bottom-up' & 'top-down' mode	✓	✓	✓	✓						Generic
2018	(Kim and Tagkopoulos, 2018)	'data-to-data', 'data-to-knowledge' & 'knowledge-to-knowledge'	✓	✓	✓	✓	✓	✓				Generic
2018	(Rappoport and Shamir, 2018)	-	✓	✓		✓						Cancer benchmarking
2019	(Tini et al., 2019)	-		✓	✓	✓	✓					Multiple (Mitochondrial metabolism, Platelet reactivity & Breast cancer)
2019	(Mirza et al., 2019)	-	✓	✓	✓	✓	✓					Generic (more focussed on ML)
2019	(López de Maturana et al., 2019)	OnO (omics & non-omics)	✓	✓	✓	✓	✓		✓			Generic
2019	(Wu et al., 2019)	'vertical', 'horizontal', 'parallel' & 'hierarchical'	✓	✓	✓	✓						Generic
2020	(Canzler et al., 2020)	-	✓	✓	✓	✓	✓				✓	Toxicological research
2020	(Eicher et al., 2020)	-	✓	✓	✓	✓						Generic (more focussed on ML)
2020	(Nicora et al., 2020)	-	✓	✓	✓	✓						Oncology
2020	(Jamil et al., 2020)	'element', 'pathway' and 'mathematical' based approach.		✓	✓	✓						Plant systems biology.
2020	(Nguyen and Wang, 2020)	'single-view' & 'multi-view'	✓	✓	✓	✓			✓	✓		Generic

As mentioned earlier, this section adopts the categorical terminologies from Ritchie (Ritchie et al., 2015) and builds upon it to summarise a complete spectrum of recent integration methods. It concisely covers them giving a clear perspective to a new interdisciplinary user. The various integration methods are classed as either 'concatenation-', 'model-' or 'transformation'-based and described below in detail.

4.1 Concatenation-based integration methods

Concatenation-based integration methods consider developing a model using a joint data matrix which is formed by combining multiple omics datasets. Figure 2 shows the stages of concatenation-based integration. *Stage 1* includes the raw data from three individual omics (e.g. genomics, proteomics, and metabolomics) along with the corresponding phenotypic information. Commonly, concatenation-based integration does not require any pre-processing and hence does not have a *Stage 2*. In *Stage 3*, the data from the individual omics is concatenated to form a single large matrix of multi-omics data. Finally, in *Stage 4* the joint matrix is used for supervised or unsupervised analysis. The main advantage of using concatenation-based methods is the simplicity of employing ML for analysing continuous or categorical data, once the concatenation of all individual omics is completed. These methods use all the concatenated features equally and can select the most discriminating features for a given phenotype.

The different concatenation-based integration methods can be further classed as:

4.1.1 Supervised learning concatenation-based methods

Different concatenation-based supervised learning methods have been used for phenotypic prediction. In scenarios where the number of features in the joint matrix are higher, different feature selection methods described in Section 3 can be employed during concatenation (Sorzano et al., 2014).

The concatenated multi-omics data (in the form of a joint matrix) is provided as input to different classical ML methods such as DT (decision tree) (Quinlan, 1993), NB (naive Bayes) (Domingos and Pazzani, 1997), ANN (artificial neural networks) (Bishop, 1995), SVM (support vector machine) (Vapnik, 1995), KNN (k-nearest neighbors) (Altman, 1992), RF (random forest) (Breiman, 2001) and K-Star (Cleary and Trigg, 1995) in the literature (Kim and Tagkopoulos, 2018; Lin and Lane, 2017; Auslander et al., 2016; Acharjee et al., 2016; Zhang et al., 2018; Ding et al., 2018; Wang et al., 2020). For example, a joint matrix of multi-omics features (which included gene expression, copy number variation and mutation) was used with classical RF and SVM to predict anti-cancer drug response (Stetson et al., 2014).

Similarly, multivariate LASSO models (Zou, 2006; Nicolai and Bühlmann Peter, 2010; Mankoo et al., 2011) have been investigated. Also, Boosted trees (Elith et al., 2008) and SVR (support vector regression) (Awad and Khanna, 2015) have been investigated for finding the longitudinal predictors of glycaemic health (Li et al., 2018).

Other than classical ML algorithms, deep neural networks (Tang et al., 2019) have also been widely used to analyse concatenated multi-omics data. They have been studied to identify robust survival subgroups of liver cancer using RNA, miRNA and methylation data (Chaudhary et al., 2017).

4.1.2 Unsupervised learning concatenation-based methods

Various concatenation-based unsupervised methods have been used for clustering and association analysis. Different matrix factorisation-based methods have evolved in recent years. Joint NMF (non-negative matrix factorisation) (Zhang et al., 2012) was proposed to integrate multi-omics data with non-negative values. It involved decomposing the joint matrix into loadings and factors, bringing the different omics into a common basis matrix. Joint NMF is computationally slow and needs large memory allocation.

Similarly, Shen (Shen et al., 2009) proposed iCluster framework which used principles similar to NMF but allows integration of datasets having negative values. They showed the functioning of the framework by using copy number, mRNA expression and methylation data to conduct a cancer subtype discovery in glioblastoma. This framework was also employed for a landmark study which used genomic and transcriptomic data from 2,000 breast tumours and discovered novel subgroups amongst them (Curtis et al., 2012).

Later, the iCluster+ framework by Mo (Mo et al., 2013), offered a significant enhancement over iCluster framework. The iCluster+ framework can discover patterns and combine a range of omics having binary, categorical and continuous values and was demonstrated by combining genomic data from the colorectal cancer datasets.

Another adaptation of NMF was evaluated as JIVE (Joint and Individual Variation Explained) which captures joint variation across integrating data types and structural variation of each data type along with the residual noise (Lock et al., 2013). It was used to investigate gene expression and miRNA data on brain tumour samples. The sparsity problem in JIVE was improved by JBF (Joint Bayes Factor) (Ray et al., 2014). JBF used

joint factor analysis to evaluate the feature space and converted it into shared and datatype-specific components.

The MoCluster proposed by Meng (Meng et al., 2016a), used multi-block multivariate analysis for highlighting the patterns across different input omics data and then finds the joint clusters amongst them. MoCluster was validated by integrating proteomic and transcriptomic data and shows a noticeably higher clustering accuracy and lower computation cost in comparison to both Cluster and iCluster+.

Fridley (Fridley et al., 2012) has studied the genomic effects due to the gemcitabine drug using high-throughput data from mRNA expression and SNPs. They integrated these two datasets into one large input matrix and developed a Bayesian pathway analysis which uses a stochastic search variable selection. Their proposed that the Bayesian integrative model offers better performance in detecting the genomic effects in comparison to using conventional single-omics analysis. Similarly, Zhu (Zhu et al., 2012) has also explored BN (Bayesian network) to understand cell regulation in yeast using metabolomics and transcriptomics data.

Also, LRAcluster (Wu et al., 2015) was developed to integrate high-dimensional multi-omics data and find low-dimensional manifold to identify molecular subtypes of cancer.

Recently, iClusterBayes was introduced by Mo (Mo et al., 2018), which is a fully Bayesian latent variable model. It overcomes the limitations of iCluster+, in terms of statistical inference and computational speed. iClusterBayes includes a binary indicator prior for selection of variable and generalises for binary data and count data. Also, Argelaguet (Argelaguet et al., 2018) have developed MOFA (Multi-Omics Factor Analysis) which disentangles the heterogeneity shared across different omics to discover the principal source of variability. It can integrate partially overlapping datasets.

4.2 Model-based integration methods

Model-based integration methods create multiple intermediate models for the different omics data and then build a final model from various intermediate models (Figure 2). *Stage 1* sets up the raw data from the three individual omics along with the corresponding phenotypic information. In *Stage 2*, individual models are developed for each of the omics which are later integrated into a joint model in *Stage 3*. Finally, in *Stage 4* the joint model is analysed. The major advantage of model-based integration methods is that they can be used for merging models based on different omic types, where each model is developed from a different patient group having the same disease information (He et al., 2016; Ritchie et al., 2015).

Model-based integration approaches facilitate the understanding of interactions amongst different omics for a certain phenotype (for example, survival in pancreatic cancer). The final multi-dimensional joint model in *Stage 4* can be built using an ML algorithm (such as neural networks) which uses the most relevant variables from each omics models (from *Stage 3*). This approach allows the analysis of the improvement in the predictive power for individual models and also finds the best discriminating features.

The different model-based integration methods can be further classed as follows.

4.2.1 Supervised learning model-based methods

Model-based supervised learning methods include a variety of frameworks for developing a model, such as majority-based voting (Drăghici and Potter, 2003)**,** hierarchical classifiers (Bavafaye Haghighi et al., 2019) and ensemble-based approaches (such as XGBoost (B. Ma et al., 2020) and KNN (Shen and Chou, 2006)).

Deep learning methods have also been adopted for model-based supervised learning (Poirion et al., 2020). MOLI (multi-omics late integration) (Sharifi-Noghabi et al., 2019) method used type-specific encoding sub-networks to learn features from somatic mutation, CNA and gene expression data independently and then later concatenated them for predicting the response to a given drug. Lee (Lee et al., 2020) has proposed a deep learning-based auto-encoding approach for integrating four omics to create a survival prediction model. Also, HI-DFNForest (hierarchical integration deep flexible neural forest) framework (J. Xu et al., 2019) was developed which uses stacked auto-encoder (Vincent et al., 2010) to learn high-level representations from three omic datasets. Later, these representations are integrated to predict cancer subtype classification. Similarly, Chaudhary (Chaudhary et al., 2017) has used autoencoders along with SVM for survival prediction in subgroups of hepatocellular carcinoma.

In the past years, ATHENA (Analysis Tool for Heritable and Environmental Network Associations) was developed for analysing multi-omics data (Chung and Kang, 2019; Holzinger et al., 2014). It uses grammatical evolution neural networks along with Biofilter (Bush et al., 2009) and Random Jungle (Schwarz et al., 2010) to investigate different categorical and quantitative variables and develop prediction models.

Recently, MOSAE (Multi-omics Supervised Autoencoder) (K. Tan et al., 2020) was developed for pan-cancer analysis and compared with conventional ML methods such as SVM, DT, naïve Bayes, KNN, RF and AdaBoost. Similarly, Denoising autoencoder has been incorporated along with L1-penalized logistic regression for identifying ovarian cancer subtypes (Guo et al., 2020).

4.2.2 Unsupervised learning model-based methods

Various model-based unsupervised learning methods have been implemented in the past. PSDF (Patient-Specific Data Fusion) (Yuan et al., 2011) is a non-parametric Bayesian model for clustering prognostic cancer subtypes by combining gene expression and copy number variation data. It uses a two-step process and limits the integration to only two datatypes. Similarly, CONEXIC (Akavia et al., 2010) also uses a BN to integrate gene expression and copy number variation from tumour samples to identify driver mutations. On the other hand, clustering methods such as FCA (Formal Concept Analysis) consensus clustering (Hristoskova et al., 2014), MDI (Multiple Dataset Integration) (Kirk et al., 2012), PINS (Perturbation clustering for data integration and disease subtyping) (Nguyen et al., 2017), PINS+ (Nguyen et al., 2019) and BCC (Bayesian consensus clustering) (Lock and Dunson, 2013) are more flexible and allow late-stage integration of clusters.

Different network-based methods are also available for association analysis. Lemon-Tree (Bonnet et al., 2015) implemented ensemble methods for reconstructing module networks which used somatic copy number alterations and gene expression in brain tumour samples. Furthermore, SNF (Similarity Network Fusion) (Wang et al., 2014) constructs networks of samples for respective data type and then effectively fuse them into a joint network which denotes the complete range of original data. It combines mRNA expression, DNA methylation and microRNA (miRNA) expression data from cancer datasets.

4.3 Transformation-based integration methods

Transformation-based integration methods transform each of the omics datasets firstly into graphs or kernel matrices and then combines all of them into one before constructing a model.

Figure 2 shows the various stages of transformation-based integration. *Stage 1* sets up the raw data from the three individual omics along with the corresponding phenotypic information. In *Stage 2*, the individual transformations (in the form of graph or kernel relationship) are developed for each of the omics which are later integrated into a joint transformation in *Stage 3*. Finally, in *Stage 4* it is analysed. The primary advantage of the transformation-based integration methods is that they can be used to combine a wide range of omics if unique information (such as patient ID) is available.

Graphs provide a formal means to transform and portray relationships between different omics samples where the nodes and edges of a graph represent the subjects and their relationships, respectively. Similarly, Kernel methods enable the transformation of data from its original space into a higher dimensional feature space. These methods then explore linear decision functions in the feature space which were non-linear in the original space.

The transformation-based integrative methods can be classed as follows.

4.3.1 Supervised learning transformation-based methods

In the past, various *transformation-based supervised learning methods* have been presented. Most of them are kernel and graph-based algorithms (Yan et al., 2017). The kernel-based integration approaches include SDP-SVM (Semi-Definite Programming SVM) (Lanckriet et al., 2004), FSMKL (Multiple Kernel Learning with Feature Selection) (Seoane et al., 2014), RVM (Relevance Vector Machine) (Bowd et al., 2005; Tipping, 2001) and Ada-boost RVM (Wu et al., 2010). Moreover, fMKL-DR (fast multiple kernel learning for dimensionality reduction) (Giang et al., 2020) has been used along with SVM for combining gene expression, miRNA expression, and DNA methylation data. Similarly, the graph-based integration approaches consist of graph-based SSL (semi-supervised learning[7]) (Tsuda et al., 2005; Culp and Michailidis, 2008; Kim et al., 2015; Yue et al., 2017; Bhardwaj and Van Steen, 2020), graph sharpening (Shin et al., 2010, 2007), composite network (Mostafavi and Morris, 2010) and BN (Rhodes et al., 2005).

Overall, it is evident from the literature that kernel-based algorithms have superior performance to graph-based approaches, but they usually need more time for the training phase. In contrast, graph-based approaches can disclose the relations between samples while taking less computation time. Yan (Yan et al., 2017) provide an extensive comparison between different graph- and kernel-based integration approaches in a supervised learning context using various standardised test datasets. It highlights the better classification performance of RVM, Ada-boost RVM and SDP-SVM in comparison to SSL, graph sharpening, composite network and BN.

[7] For the sake of simplicity, the semi-supervised integration methods (graph-based) are grouped under supervised learning.

Recently, MORONET (Multi-Omics gRaph cOnvolutional NETworks) (Wang et al., 2020) is introduced, which use graph convolutional networks taking benefit of the omics features and the associations among patients (as defined by the patient similarity networks) for better classification results.

4.3.2 Unsupervised learning transformation-based methods

Different *transformation-based unsupervised methods* have been introduced. Some of them are kernel- and graph-based methods. Lately, rMKL-LPP (regularised multiple kernel learning for Locality Preserving Projections) (Speicher and Pfeifer, 2015) was implemented for clustering analysis. It used an individual kernel for each omics along with graph embedding framework to identify biologically meaningful subgroups for five different cancer types. Similarly, PAMOGK (Tepeli et al., 2019) is developed for integrating multi-omics data with pathways using graph kernel, SmSPK (smoothed shortest path graph kernel). It used somatic mutations, transcriptomics and proteomics data to find subgroups of kidney cancer.

Meta-SVM (Meta-analytic SVM) is proposed by Kim (Kim et al., 2017), which integrates multiple omics data and able to detect consensus genes associated with diseases across studies such as breast cancer and idiopathic pulmonary fibrosis. Recently, NEMO (NEighborhood based Multi-Omics clustering) (Rappoport and Shamir, 2019) is introduced which uses an inter-patient similarity matrix–based distance metric for evaluating the input omic datasets individually. These omics matrices are then combined into one matrix and then analysed using spectral-based clustering. It can work on partial data sets (no imputation needed), where measurements are only available for a subset of omics data.

Table 6 highlights the advantages and disadvantages of various integration methods. Table 7 summarises various multi-omics integration methods based on learning type.

Table 6: The advantages and disadvantages of using different integrative methods.

Integrative Method	Advantages	Disadvantages
Concatenation-Based	• Easy and straightforward. • Enables the use of classical supervised and unsupervised methods.	• Ideally, requires all omics data for all patients. • Need proper normalisation before concatenation. • Does not consider the unique distribution of each omics. • Memory and computation-intensive when the concatenated matrix is large.
Model-Based	• Facilitates the understanding of interactions amongst different omics. • Omics data can be from a different set of patients with a similar phenotype. • Does not increase dimensional complexity.	• Not effective if omics data is extremely heterogeneous. • Could lead to an overfitted solution • Weak signals could be lost.
Transformation-Based	• Graph representation easy to understand and computationally less intensive. • Kernel methods provide superior performance. • Multi-omics data for the same patient can be used for their disease subgroup analysis.	• Kernel methods are computationally more intensive than graph methods. • Transformation can be sometimes challenging.

Table 7: The summary of multi-omics integration methods based on learning type. For abbreviations please refer to *List of Abbreviations*.

		Multi-omics Integration Methods		
		Concatenation-based	**Model-based**	**Transformation-based**
Learning Type	**Supervised**	• Classical ML (DT (Quinlan, 1993), NB (Domingos and Pazzani, 1997), ANN (Bishop, 1995), SVM (Vapnik, 1995), KNN (Altman, 1992), K-Star (Cleary and Trigg, 1995) • LASSO (Zou, 2006; Nicolai and Bühlmann Peter, 2010; Mankoo et al., 2011) • BT (Elith et al., 2008) • SVR (Awad and Khanna, 2015) • DNN (Tang et al., 2019)	• Majority-based voting (Drăghici and Potter, 2003) • Hierarchical Classifiers (Bavafaye Haghighi et al., 2019) • Ensemble-based classifiers (XGBoost (B. Ma et al., 2020) and KNN (Shen and Chou, 2006)) • MOLI (Sharifi-Noghabi et al., 2019) • HI-DFNForest (J. Xu et al., 2019) • ATHENA (Chung and Kang, 2019; Holzinger et al., 2014)	• SDP-SVM (Lanckriet et al., 2004) • FSMKL (Seoane et al., 2014) • RVM (Bowd et al., 2005; Tipping, 2001) • Ada-boost RVM (Wu et al., 2010) • fMKL-DR (Giang et al., 2020) • SSL (Tsuda et al., 2005; Culp and Michailidis, 2008; Kim et al., 2015; Yue et al., 2017; Bhardwaj and Van Steen, 2020), • Graph sharpening (Shin et al., 2010, 2007) • Composite network (Mostafavi and Morris, 2010) • BN (Rhodes et al., 2005) • MORONET (Wang et al., 2020)
	Unsupervised	• Joint NMF (Zhang et al., 2012) • iCluster (Shen et al., 2009) • iCluster+ (Mo et al., 2013) • JIVE (Lock et al., 2013) • JBF (Ray et al., 2014) • BN (Fridley et al., 2012; Zhu et al., 2012) • MoCluster (Meng et al., 2016a) • iClusterBayes (Mo et al., 2018) • MOFA (Argelaguet et al., 2018)	• PSDF (Yuan et al., 2011) • FCA consensus clustering (Hristoskova et al., 2014) • MDI (Kirk et al., 2012) • BCC (Lock and Dunson, 2013) • Lemon-Tree (Bonnet et al., 2015) • SNF (Wang et al., 2014)	• rMKL-LPP (Speicher and Pfeifer, 2015) • PAMOGK (Tepeli et al., 2019) • Meta-SVM (Kim et al., 2017) • NEMO (Rappoport and Shamir, 2019)

5 APPLICATION OF INTEGRATIVE METHODS IN MULTI-OMICS STUDIES

The availability of high-throughput omics provides a unique opportunity to explore the complex relationships between different omics and phenotypic targets instead of mono-omics evaluation. This section describes various multi-omics studies which deployed methods investigated in the previous section. Table 8 summarises different phenotypic target-based, multi-omics studies published and tabulates them across the span of 7 main omics namely, genomics, transcriptomics, metabolomics, proteomics, glycomics, lipidomics and epigenomics. Genomics is further divided into gene expression, DNA methylation, somatic point mutation and copy number alteration. Similarly, transcriptomics is further classed into lncRNAs (long non-coding RNAs) and microRNAs (mRNA and miRNA). The various multi-omics studies are broadly grouped based on the target and the corresponding ML method used.

It is evident from Table 8 that most of the multi-omics studies focus on different forms of cancer. In particular, the presence of many multi-omics studies related to breast (Chen et al., 2017; Lee et al., 2017; List et al., 2014; Ma et al., 2016; Nam et al., 2009) and ovarian (Anděl et al., 2015; Mankoo et al., 2011; Paik et al., 2017; Zhang et al., 2014) cancer highlights the research thrust by the scientific community in these domains.

Many intra-omics studies have successfully explored the integration of gene expression and DNA methylation. LASSO methods have been used for this particular integration by Taskesen (Taskesen et al., 2015) and Lee (Lee et al., 2017) for acute myeloid leukaemia and breast cancer respectively. LASSO has also been employed for cancer prognosis

(Zhao et al., 2015). Similarly, mRNA – miRNA integration was investigated using Neural Fuzzy Network for colorectal cancer (Vineetha et al., 2013), SVM for pancreatic cancer (Kwon et al., 2015), and RF for cardiac tissue ageing (Dimitrakopoulos et al., 2014) and ovarian cancer (Anděl et al., 2015) respectively. SVM has also been used for oral squamous cell carcinoma study by integrating different transcriptomics namely mRNA, miRNA and IncRNA (Li et al., 2017).

Metabolomics and proteomics have been integrated using RF for analysis of prostate cancer (Fan et al., 2011) and thyroid functioning (Pietzner et al., 2017). Similarly, metabolomics is integrated with mRNA for studying ulcerative colitis (Bjerrum et al., 2014) and cancer survival (Kim et al., 2014). On the other hand, glycomics and epigenomics have only appeared once in the multi-omics context (along with mRNA and metabolomics) and used by Zierer (Zierer et al., 2016) for the study of age-related comorbidities using a graphical variant of RF.

Recently, metabolomics and proteomics have also been integrated with lipidomics to evaluate COVID-19 patients using PLS-DA (Partial Least Squares Discriminant Analysis) and Extra Trees (Overmyer et al., 2020; Thomas et al., 2020).

Multi-omics studies have also been successfully conducted in plants (potato (Acharjee et al., 2016, 2011)) and animals (such as canine heart disease (Li et al., 2015)).

Table 8: Multi-omics studies using different MI methods. For abbreviations please refer to *List of Abbreviations*.

OMICS ▸ / Target ▾	Genomics				Transcriptomics			Metabolomics	Proteomics	Glycomics	Lipidomics	Epigenomics	Method Used	Method Type	Reference
	Gene expression	DNA methylation	Somatic point mutation	Copy number alteration	mRNA	miRNA	lncRNA								
Humans															
Age-related					✓			✓		✓		✓	Graphical RF	CU	(Zierer et al., 2016)
Acute myeloid leukaemia	✓	✓											LASSO	CS	(Taskesen et al., 2015)
Anti-cancer therapeutic response	✓			✓									RF & SVM	CS	(Stetson et al., 2014)
Biomedical data classification		✓			✓	✓							MORONET	TS	(Wang et al., 2020)
Brain cancer	✓	✓	✓	✓		✓							LASSO	CS	(Lu et al., 2016)
	✓					✓							JIVE	CU	(Lock et al., 2013)
	✓		✓	✓	✓								iClusterBayes	CU	(Mo et al., 2018)
	✓			✓									Lemon-Tree	MU	(Bonnet et al., 2015)
		✓			✓	✓							SNF	M	(Wang et al.,

												U	2014)
Breast cancer	✓	✓						✓			RF	CS	(List et al., 2014)
	✓	✓									LASSO	CS	(Lee et al., 2017)
	✓							✓			RF & SVM	CS	(Nam et al., 2009)
	✓	✓		✓							LASSO	CS	(Chen et al., 2017)
	✓							✓			SVM	CS	(Auslander et al., 2016)
	✓			✓							iCluster	CU	(Shen et al., 2009)
	✓			✓					✓		SVM, RF, SVM & Multi-Kernel Learning	CS & TS	(Ma et al., 2016)
	✓			✓							iCluster	CU	(Curtis et al., 2012)
		✓			✓	✓			✓		BCC	M U	(Lock and Dunson, 2013)
	✓			✓							FSMKL	TS	(Seoane et al., 2014)
		✓		✓	✓						Meta-SVM	TU	(Kim et al., 2017)
Cancer survival					✓			✓			SVM & RF	CS	(Kim et al., 2014)
Cancer prognosis	✓	✓		✓		✓					LASSO	CS	(Zhao et al., 2015)
	✓			✓							PSDF	M U	(Yuan et al., 2011)
	✓			✓							CONEXIC	M U	(Akavia et al., 2010)
Cancer drug response	✓		✓	✓							MOLI (DL)	MS	(Sharifi-Noghabi et al., 2019)
Cardiac tissue ageing					✓	✓					RF	CS	(Dimitrakopoulos et al., 2014)
Colorectal cancer					✓	✓					Neural Fuzzy Network	CU	(Vineetha et al., 2013)
COVID-19 analysis								✓	✓	✓	PLS-DA	CS	(Thomas et al., 2020)
								✓	✓	✓	Extra Trees	CS	(Overmyer et al., 2020)
Chronic lymphocytic leukaemia		✓	✓		✓						MOFA	CU	(Argelaguet et al., 2018)
Gastric cancer	✓					✓					SVM & RF	CS	(Yan et al., 2012)
Kidney cancer		✓	✓	✓	✓	✓					iClusterBayes	CU	(Mo et al., 2018)
	✓		✓						✓		PAMOGK	TU	(Tepeli et al., 2019)
Liver cancer		✓			✓	✓					Auto-encoder, SVM	MS	(Chaudhary et al., 2017)
Lung cancer	✓			✓							iCluster	CU	(Shen et al., 2009)
		✓		✓	✓	✓					Auto-encoder	MS	(Lee et al., 2020)
Neuroblastoma	✓			✓							Auto-encoders, SVM & NB	CS	(Zhang et al., 2018)
Ovarian cancer					✓	✓					RF	CS	(Anděl et al., 2015)
		✓			✓						RF	CS	(Paik et al., 2017)
	✓	✓		✓		✓					LASSO	CS	(Mankoo et al., 2011)
	✓	✓				✓					Joint NMF	CU	(Zhang et al., 2012)
	✓	✓		✓							JBF	CU	(Ray et al., 2014)
	✓	✓	✓	✓							BN	TS	(Zhang et al.,

												2014)
	✓	✓		✓		✓				Graph SSL	TS	(Kim et al., 2015)
Oral squamous cell carcinoma					✓	✓	✓			SVM	CS	(Li et al., 2017)
	✓		✓	✓						iCluster+	CU	(Mo et al., 2013)
					✓				✓	moCluster	CU	(Meng et al., 2016a)
	✓	✓	✓	✓						LRAcluster	CU	(Wu et al., 2015)
		✓			✓	✓				XGBoost	MS	(B. Ma et al., 2020)
	✓	✓	✓							RF	MS	(Bavafaye Haghighi et al., 2019)
Pan-cancer analysis	✓	✓				✓				HI-DFN Forest (AE)	MS	(J. Xu et al., 2019)
		✓			✓	✓			✓	MOSAE	MS	(K. Tan et al., 2020)
	✓	✓		✓		✓				PINS	M U	(Nguyen et al., 2017)
	✓	✓			✓					fMKL-DR	TS	(Giang et al., 2020)
	✓	✓				✓				rMKL-LPP	TU	(Speicher and Pfeifer, 2015)
	✓	✓				✓				NEMO	TU	(Rappoport and Shamir, 2019)
Pancreatic cancer					✓	✓				SVM	MS	(Kwon et al., 2015)
Prostate cancer								✓	✓	RF	CS	(Fan et al., 2011)
Precision oncology	✓			✓						Auto-encoders, Elastic Net, SVM & Consensus Clustering	CS & M U	(Ding et al., 2018)
Thyroid function								✓	✓	RF	CS	(Pietzner et al., 2017)
Ulcerative colitis					✓			✓		LASSO	CS	(Bjerrum et al., 2014)
Plants												
Potato flesh colour	✓							✓	✓	RF	CS	(Acharjee et al., 2016)
	✓							✓		RF	MS	(Acharjee et al., 2011)
Animals & Micro-organisms												
Dog heart disease					✓			✓		RF	CS	(Li et al., 2015)
Yeast	✓							✓		BN	CU	(Zhu et al., 2012)

Overall, the different recent multi-omics studies highlight the superiority of integration methods in understanding the complexity of different diseases and uncovering the underlying abnormalities from the vastly generated multi-omics data, which is not always possible with individual omics analysis.

6 RECOMMENDATIONS

Today a plethora of multi-omic integration methods are available for both supervised and unsupervised learning as evident in the current review. This information can overwhelm interdisciplinary scientists and would require a time-consuming effort to understand the challenging mathematical and computational concepts behind them. Hence, we suggest

that interdisciplinary teams working on multi-omics always include ML practitioners to assist with the choice of methods, the development of solutions, the interpretation of results and their significance and limits. Such truly interdisciplinary teams offer real opportunities for better mutual understanding of the different fields, practice and expertise necessary, leading ultimately to more robust conclusions. In addition, to facilitate the method selection process, a recommendation flowchart is proposed in Figure 3. It shows the various decision steps required for choosing an appropriate method (or family of methods) for a given scenario. For example, to choose a method for integrating two omics for unsupervised learning one can choose model-based method such as 'PSDF or Lemon-Tree' if the two omics are gene expression and CNV, otherwise 'MDI or SNF' can be used. Similarly, 'NEMO' can be used in scenarios where the datasets are partially overlapping, and transformation approach is required. Hence, it can be used for the purpose of biomedical analysis, including diagnosis, prognosis and biomarker identification, by posing them as supervised or unsupervised learning problems.

Clearly, a '*one-size-fits-all*' approach is not feasible. Also, unfortunately, the existing literature does not provide many direct comparisons between methods using the same publicly available datasets. Hence, to choose the best method which suits a given dataset and question, an empirical approach which investigates the use of different methods, guided by ML practitioners is recommended.

7 CONCLUSIONS

This paper reviewed various ML approaches used for integration of multi-omics data for analysis. A concise background of multi-omics and ML was presented. It examined the concatenation-, model- and transformation based integration methods, employed for multi-omics data along with their advantages and disadvantages. Also, various existing multi-omics studies have been summarised. Finally, a recommendation flowchart is presented for interdisciplinary professionals to choose an appropriate method for a multi-omics dataset. Overall, this work showcases the recent findings in the multi-omics domain and signifies the key role of ML in the future of personalised healthcare.

Disclosure: The authors have nothing to disclose.

Acknowledgement
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 633983. Ewan Pearson and Emanuele Trucco would like to acknowledge the National Institute for Health Research (NIHR) global health research unit on global diabetes outcomes research at the University of Dundee (INSPIRED project, Award number 16/136/102) for useful discussions.

8 References

Acharjee, A., Kloosterman, B., de Vos, R.C.H., Werij, J.S., Bachem, C.W.B., Visser, R.G.F., Maliepaard, C., 2011. Data integration and network reconstruction with ~omics data using Random Forest regression in potato. Anal. Chim. Acta 705, 56–63. https://doi.org/10.1016/j.aca.2011.03.050

Acharjee, A., Kloosterman, B., Visser, R.G.F., Maliepaard, C., 2016. Integration of multi-omics data for prediction of phenotypic traits using random forest. BMC Bioinformatics 17, 180. https://doi.org/10.1186/s12859-016-1043-4

Agache, I., Rogozea, L., 2017. Asthma Biomarkers: Do They Bring Precision Medicine Closer to the Clinic? Allergy Asthma Immunol. Res. 9, 466–476. https://doi.org/10.4168/aair.2017.9.6.466

Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., Pe'er, D., 2010. An Integrated Approach to Uncover Drivers of Cancer. Cell 143, 1005–1017. https://doi.org/10.1016/j.cell.2010.11.013

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2008. Molecular Biology of the Cell 5E, 5 edition. ed. Garland Science, New York.

Alidjinou, E.K., Deldalle, J., Hallaert, C., Robineau, O., Ajana, F., Choisy, P., Hober, D., Bocket, L., 2017. RNA and DNA Sanger sequencing versus next-generation sequencing for HIV-1 drug resistance testing in treatment-naive patients. J. Antimicrob. Chemother. 72, 2823–2830. https://doi.org/10.1093/jac/dkx232

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am. Stat. 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879

Amancio, D.R., Comin, C.H., Casanova, D., Travieso, G., Bruno, O.M., Rodrigues, F.A., Costa, L. da F., 2014. A Systematic Comparison of Supervised Classifiers. PLOS ONE 9, e94137. https://doi.org/10.1371/journal.pone.0094137

Amazon EC2 [WWW Document], n.d. . Amaz. Web Serv. Inc. URL https://aws.amazon.com/ec2/ (accessed 10.29.20).

Anděl, M., Kléma, J., Krejčík, Z., 2015. Network constrained forest for regularized classification of omics data. Methods, Network-based Approaches to the Analysis of Omics Data 83, 88–97. https://doi.org/10.1016/j.ymeth.2015.04.006

Antonelli, J., Claggett, B.L., Henglin, M., Kim, A., Ovsak, G., Kim, N., Deng, K., Rao, K., Tyagi, O., Watrous, J.D., Lagerborg, K.A., Hushcha, P.V., Demler, O.V., Mora, S., Niiranen, T.J., Pereira, A.C., Jain, M., Cheng, S., 2019. Statistical Workflow for Feature Selection in Human Metabolomics Data. Metabolites 9. https://doi.org/10.3390/metabo9070143

Archer, D.B., Bricker, J.T., Chu, W.T., Burciu, R.G., McCracken, J.L., Lai, S., Coombes, S.A., Fang, R., Barmpoutis, A., Corcos, D.M., Kurani, A.S., Mitchell, T., Black, M.L., Herschel, E., Simuni, T., Parrish, T.B., Comella, C., Xie, T., Seppi, K., Bohnen, N.I., Müller, M.L., Albin, R.L., Krismer, F., Du, G., Lewis, M.M., Huang, X., Li, H., Pasternak, O., McFarland, N.R., Okun, M.S., Vaillancourt, D.E., 2019. Development and validation of the automated imaging differentiation in parkinsonism (AID-P): a multicentre machine learning study. Lancet Digit. Health 1, e222–e231. https://doi.org/10.1016/S2589-7500(19)30105-0

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O., 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. 14. https://doi.org/10.15252/msb.20178124

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M., 2010. A view of cloud computing. Commun. ACM 53, 50–58. https://doi.org/10.1145/1721654.1721672

Aslam, B., Basit, M., Nisar, M.A., Khurshid, M., Rasool, M.H., 2017. Proteomics: Technologies and Their Applications. J. Chromatogr. Sci. 55, 182–196. https://doi.org/10.1093/chromsci/bmw167

Auslander, N., Yizhak, K., Weinstock, A., Budhu, A., Tang, W., Wang, X.W., Ambs, S., Ruppin, E., 2016. A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. Sci. Rep. 6. https://doi.org/10.1038/srep29662

Awad, M., Khanna, R., 2015. Support Vector Regression, in: Awad, M., Khanna, R. (Eds.), Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress, Berkeley, CA, pp. 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple Imputation by Chained Equations: What is it and how does it work? Int. J. Methods Psychiatr. Res. 20, 40–49. https://doi.org/10.1002/mpr.329

Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., Zhang, J.D., 2020. An Introduction to Machine Learning. Clin. Pharmacol. Ther. 107, 871–885. https://doi.org/10.1002/cpt.1796

Badue, C., Guidolini, R., Carneiro, R.V., Azevedo, P., Cardoso, V.B., Forechi, A., Jesus, L., Berriel, R., Paixão, T.M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., De Souza, A.F., 2021. Self-driving cars: A survey. Expert Syst. Appl. 165, 113816. https://doi.org/10.1016/j.eswa.2020.113816

Barh, D., Blum, K., Madigan, M.A. (Eds.), 2011. OMICS: Biomedical Perspectives and Applications, 1 edition. ed. CRC Press, Boca Raton.

Barh, D., Tiwari, S., Weener, M.E., Azevedo, V., Góes-Neto, A., Gromiha, M.M., Ghosh, P., 2020. Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19. Comput. Biol. Med. 126, 104051. https://doi.org/10.1016/j.compbiomed.2020.104051

Barnes, J.W., Tonelli, A.R., Heresi, G.A., Newman, J.E., Mellor, N.E., Grove, D.E., Dweik, R.A., 2016. Novel methods in pulmonary hypertension phenotyping in the age of precision medicine (2015 Grover Conference series). Pulm. Circ. 6, 439–447. https://doi.org/10.1086/688847

Barnett-Itzhaki, Z., Elbaz, M., Butterman, R., Amar, D., Amitay, M., Racowsky, C., Orvieto, R., Hauser, R., Baccarelli, A.A., Machtinger, R., 2020. Machine learning vs. classic statistics for the prediction of IVF outcomes. J. Assist. Reprod. Genet. 37, 2405–2412. https://doi.org/10.1007/s10815-020-01908-1

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bavafaye Haghighi, E., Knudsen, M., Elmedal Laursen, B., Besenbacher, S., 2019. Hierarchical Classification of Cancers of Unknown Primary Using Multi-Omics Data. Cancer Inform. 18. https://doi.org/10.1177/1176935119872163

BCS, T.C.I. for I., 2014. Big Data: Opportunities and challenges. BCS, The Chartered Institute for IT.

Bellazzi, R., 2014. Big data and biomedical informatics: a challenging opportunity. Yearb. Med. Inform. 9, 8–13. https://doi.org/10.15265/IY-2014-0024

Benjamens, S., Dhunnoo, P., Meskó, B., 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. Npj Digit. Med. 3, 1–8. https://doi.org/10.1038/s41746-020-00324-0

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2011. GenBank. Nucleic Acids Res. 39, D32–D37. https://doi.org/10.1093/nar/gkq1079

Beretta, L., Santaniello, A., 2011. Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets. J. Biomed. Inform. 44, 361–369. https://doi.org/10.1016/j.jbi.2010.12.003

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., Milanesi, L., 2016. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 17, 167–177. https://doi.org/10.1186/s12859-015-0857-9

Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., Wang, J., 2019. Applications and analysis of targeted genomic sequencing in cancer studies. Comput. Struct. Biotechnol. J. 17, 1348–1359. https://doi.org/10.1016/j.csbj.2019.10.004

Bhardwaj, A., Van Steen, K., 2020. Multi-omics Data and Analytics Integration in Ovarian Cancer. Artif. Intell. Appl. Innov. 584, 347–357. https://doi.org/10.1007/978-3-030-49186-4_29

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA.

Bjerrum, J.T., Rantalainen, M., Wang, Y., Olsen, J., Nielsen, O.H., 2014. Integration of transcriptomics and metabonomics: improving diagnostics, biomarker identification and phenotyping in ulcerative colitis. Metabolomics Off. J. Metabolomic Soc. 10, 280–290. https://doi.org/10.1007/s11306-013-0580-3

Black box medicine and transparency (Executive Summary), 2020. . PHG Foundation (University of Cambridge), London, UK.

Boellner, S., Becker, K.-F., 2015. Reverse Phase Protein Arrays—Quantitative Assessment of Multiple Biomarkers in Biopsies for Clinical Use. Microarrays 4, 98–114. https://doi.org/10.3390/microarrays4020009

Bonnet, E., Calzone, L., Michoel, T., 2015. Integrative Multi-omics Module Network Inference with Lemon-Tree. PLOS Comput. Biol. 11, e1003983. https://doi.org/10.1371/journal.pcbi.1003983

Borad, M.J., LoRusso, P.M., 2017. Twenty First Century Precision Medicine in Oncology: Genomic Profiling in Patients With Cancer. Mayo Clin. Proc. 92, 1583–1591. https://doi.org/10.1016/j.mayocp.2017.08.002

Bowd, C., Medeiros, F.A., Zhang, Z., Zangwill, L.M., Hao, J., Lee, T.-W., Sejnowski, T.J., Weinreb, R.N., Goldbaum, M.H., 2005. Relevance Vector Machine and Support Vector Machine Classifier Analysis of Scanning Laser Polarimetry Retinal Nerve Fiber Layer Measurements. Invest. Ophthalmol. Vis. Sci. 46, 1322–1329. https://doi.org/10.1167/iovs.04-1122

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Buescher, J.M., Driggers, E.M., 2016. Integration of omics: more than the sum of its parts. Cancer Metab. 4, 4. https://doi.org/10.1186/s40170-016-0143-y

Bumgarner, R., 2013. DNA microarrays: Types, Applications and their future. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 0 22, Unit-22.1. https://doi.org/10.1002/0471142727.mb2201s101

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L.D., Christie, C., Duarte, J.M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D.S., Green, R.K., Guranovic, V., Guzenko, D., Hudson, B.P., Liang, Y., Lowe, R., Peisach, E., Periskova, I., Randle, C., Rose, A., Sekharan, M., Shao, C., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Young, J., Zardecki, C., Zhuravleva, M., Kurisu, G., Nakamura, H., Kengaku, Y., Cho, H., Sato, J., Kim, J.Y., Ikegawa, Y., Nakagawa, A., Yamashita, R., Kudou, T., Bekker, G.-J., Suzuki, H., Iwata, T., Yokochi, M., Kobayashi, N., Fujiwara, T., Velankar, S., Kleywegt, G.J., Anyango, S., Armstrong, D.R., Berrisford, J.M., Conroy, M.J., Dana, J.M., Deshpande, M., Gane, P., Gáborová, R., Gupta, D., Gutmanas, A., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., Nair, S.,

Patwardhan, A., Paysan-Lafosse, T., Pravda, L., Salih, O., Sehnal, D., Varadi, M., Vařeková, R., Markley, J.L., Hoch, J.C., Romero, P.R., Baskaran, K., Maziuk, D., Ulrich, E.L., Wedell, J.R., Yao, H., Livny, M., Ioannidis, Y.E., 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47, D520–D528. https://doi.org/10.1093/nar/gky949

Bush, W.S., Dudek, S.M., Ritchie, M.D., 2009. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 368–379.

Bzdok, D., 2017. Classical Statistics and Statistical Learning in Imaging Neuroscience. Front. Neurosci. 11. https://doi.org/10.3389/fnins.2017.00543

Caffrey, M., Hogan, J., 1992. LIPIDAT: A database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis. Chem. Phys. Lipids 61, 1–109. https://doi.org/10.1016/0009-3084(92)90002-7

Campbell, M.P., Nguyen-Khuong, T., Hayes, C.A., Flowers, S.A., Alagesan, K., Kolarich, D., Packer, N.H., Karlsson, N.G., 2014. Validation of the curation pipeline of UniCarb-DB: Building a global glycan reference MS/MS repository. Biochim. Biophys. Acta BBA - Proteins Proteomics, Computational Proteomics in the Post-Identification Era 1844, 108–116. https://doi.org/10.1016/j.bbapap.2013.04.018

Canuel, V., Rance, B., Avillach, P., Degoulet, P., Burgun, A., 2015. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Brief. Bioinform. 16, 280–290. https://doi.org/10.1093/bib/bbu006

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U.E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., Hackermüller, J., 2020. Prospects and challenges of multi-omics data integration in toxicology. Arch. Toxicol. 94, 371–388. https://doi.org/10.1007/s00204-020-02656-y

Cao, K., Bai, X., Hong, Y., Wan, L., 2020. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics 36, i48–i56. https://doi.org/10.1093/bioinformatics/btaa443

Capobianco, E., 2017. Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective. Clin. Transl. Med. 6, 23. https://doi.org/10.1186/s40169-017-0155-4

Carlomagno, N., Incollingo, P., Tammaro, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M.L., Minieri, G., Pisani, A., Riccio, E., Sabbatini, M., Bracale, U.M., Calogero, A., Lodaro, C.A., Santangelo, M., 2017. Diagnostic, Predictive, Prognostic, and Therapeutic Molecular Biomarkers in Third Millennium: A Breakthrough in Gastric Cancer. BioMed Res. Int. 2017. https://doi.org/10.1155/2017/7869802

Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.X., 2017. Deep Learning based multi-omics integration robustly predicts survival in liver cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-17-0853

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953

Chawla, N.V., Davis, D.A., 2013. Bringing big data to personalized healthcare: a patient-centered framework. J. Gen. Intern. Med. 28 Suppl 3, S660-665. https://doi.org/10.1007/s11606-013-2455-8

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. 58, 101539. https://doi.org/10.1016/j.media.2019.101539

Chen, Y., Wang, X., Wang, G., Li, Z., Wang, J., Huang, L., Qin, Z., Yuan, X., Cheng, Z., Zhang, S., Yin, Y., He, J., 2017. Integrating multiple omics data for the discovery of potential Beclin-1 interactions in breast cancer. Mol. Biosyst. 13, 991–999. https://doi.org/10.1039/c6mb00653a

Cheng, P.F., Dummer, R., Levesque, M.P., 2015. Data mining The Cancer Genome Atlas in the era of precision cancer medicine. Swiss Med. Wkly. 145, w14183. https://doi.org/10.4414/smw.2015.14183

Chung, R.-H., Kang, C.-Y., 2019. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience 8. https://doi.org/10.1093/gigascience/giz045

Clarivate Analytics, 2020. Web of Science [v.5.35] - Web of Science Core Collection Basic Search [WWW Document]. Web Sci. URL https://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch (accessed 10.23.20).

Cleary, J.G., Trigg, L.E., 1995. K*: An Instance-based Learner Using and Entropic Distance Measure, in: Proceedings of the Twelfth International Conference on International Conference on Machine Learning, ICML'95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 108–114.

Clifton, D.A., Niehaus, K.E., Charlton, P., Colopy, G.W., 2015. Health Informatics via Machine Learning for the Clinical Management of Patients. Yearb. Med. Inform. 10, 38–43. https://doi.org/10.15265/IY-2015-014

Cloud Computing Services | Microsoft Azure [WWW Document], n.d. URL https://azure.microsoft.com/en-gb/ (accessed 10.29.20).

Cloud Computing Services [WWW Document], n.d. . Google Cloud. URL https://cloud.google.com/ (accessed 10.29.20).

Cobb, M., 2017. 60 years ago, Francis Crick changed the logic of biology. PLOS Biol. 15, e2003243. https://doi.org/10.1371/journal.pbio.2003243

Conesa, A., Beck, S., 2019. Making multi-omics data accessible to researchers. Sci. Data 6, 251. https://doi.org/10.1038/s41597-019-0258-4

Coronato, A., Naeem, M., Pietro, G.D., Paragliola, G., 2020. Reinforcement Learning for Intelligent Healthcare Applications: A Survey. Artif. Intell. Med. 101964. https://doi.org/10.1016/j.artmed.2020.101964

Costantino, S., Libby, P., Kishore, R., Tardif, J.-C., El-Osta, A., Paneni, F., 2017. Epigenetics and precision medicine in cardiovascular patients: from basic concepts to the clinical arena. Eur. Heart J. https://doi.org/10.1093/eurheartj/ehx568

Cox, B., 2009. Building bridges from "omics" to cell biology. Genome Biol. 10, 305. https://doi.org/10.1186/gb-2009-10-3-305

Crookston, N.L., Finley, A.O., 2008. yaImpute: An R Package for kNN Imputation. J. Stat. Softw. 23, 1–16. https://doi.org/10.18637/jss.v023.i10

Culp, M., Michailidis, G., 2008. Graph-based semisupervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 30, 174–179. https://doi.org/10.1109/TPAMI.2007.70765

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J.D., Tavaré, S., Caldas, C., Aparicio, S., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352. https://doi.org/10.1038/nature10983

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O., Ajibuwa, O.E., 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5. https://doi.org/10.1016/j.heliyon.2019.e01802

Dankers, F.J.W.M., Traverso, A., Wee, L., van Kuijk, S.M.J., 2019. Prediction Modeling Methodology, in: Kubben, P., Dumontier, M., Dekker, A. (Eds.), Fundamentals of Clinical Data Science. Springer, Cham (CH).

Davenport, T., Kalakota, R., 2019. The potential for artificial intelligence in healthcare. Future Healthc. J. 6, 94–98. https://doi.org/10.7861/futurehosp.6-2-94

de Andrade, B.M., de Gois, J.S., Xavier, V.L., Luna, A.S., 2020. Comparison of the performance of multiclass classifiers in chemical data: Addressing the problem of overfitting with the permutation test. Chemom. Intell. Lab. Syst. 201, 104013. https://doi.org/10.1016/j.chemolab.2020.104013

Debnath, M., Prasad, G.B.K.S., Bisen, P.S., Prasad, G.B.K.S., 2010. Molecular diagnostics: promises and possibilities. Springer, Dordrecht.

Delavan, B., Roberts, R., Huang, R., Bao, W., Tong, W., Liu, Z., 2017. Computational drug repositioning for rare diseases in the era of precision medicine. Drug Discov. Today. https://doi.org/10.1016/j.drudis.2017.10.009

Deng, L., Cai, Y., Zhang, W., Yang, W., Gao, B., Liu, H., 2020. Pathway-Guided Deep Neural Network toward Interpretable and Predictive Modeling of Drug Sensitivity. J. Chem. Inf. Model. 60, 4497–4505. https://doi.org/10.1021/acs.jcim.0c00331

Dias-Audibert, F.L., Navarro, L.C., de Oliveira, D.N., Delafiori, J., Melo, C.F.O.R., Guerreiro, T.M., Rosa, F.T., Petenuci, D.L., Watanabe, M.A.E., Velloso, L.A., Rocha, A.R., Catharino, R.R., 2020. Combining Machine Learning and Metabolomics to Identify Weight Gain Biomarkers. Front. Bioeng. Biotechnol. 8. https://doi.org/10.3389/fbioe.2020.00006

Dimitrakopoulos, G.N., Dimitrakopoulou, K., Maraziotis, I.A., Sgarbas, K., Bezerianos, A., 2014. Supervised method for construction of microRNA-mRNA networks: application in cardiac tissue aging dataset. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2014, 318–321. https://doi.org/10.1109/EMBC.2014.6943593

Ding, M.Q., Chen, L., Cooper, G.F., Young, J.D., Lu, X., 2018. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol. Cancer Res. 16, 269–278. https://doi.org/10.1158/1541-7786.MCR-17-0378

Domingos, P., Pazzani, M., 1997. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. Mach Learn 29, 103–130. https://doi.org/10.1023/A:1007413511361

Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J., 2018. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognit. 74, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037

Dominiczak, A., Delles, C., Padmanabhan, S., 2017. Genomics and Precision Medicine for Clinicians and Scientists in Hypertension. Hypertens. Dallas Tex 1979 69, e10–e13. https://doi.org/10.1161/HYPERTENSIONAHA.116.08252

D'Onofrio, D.J., An, G., 2010. A comparative approach for the investigation of biological information processing: An examination of the structure and function of computer hard drives and DNA. Theor. Biol. Med. Model. 7, 3. https://doi.org/10.1186/1742-4682-7-3

Drăghici, S., Potter, R.B., 2003. Predicting HIV drug resistance with neural networks. Bioinforma. Oxf. Engl. 19, 98–107.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern classification, 2nd ed. ed. Wiley, New York.

Ebbels, T.M.D., Cavill, R., 2009. Bioinformatic methods in NMR-based metabolic profiling. Prog. Nucl. Magn. Reson. Spectrosc. 55, 361–374. https://doi.org/10.1016/j.pnmrs.2009.07.003

Eicher, T., Kinnebrew, G., Patt, A., Spencer, K., Ying, K., Ma, Q., Machiraju, R., Mathé, E.A., 2020. Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. Metabolites 10. https://doi.org/10.3390/metabo10050202

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Erickson, B.J., Korfiatis, P., Akkus, Z., Kline, T.L., 2017. Machine Learning for Medical Imaging. Radiogr. Rev. Publ. Radiol. Soc. N. Am. Inc 37, 505–515. https://doi.org/10.1148/rg.2017160130

Fan, Y., Murphy, T.B., Byrne, J.C., Brennan, L., Fitzpatrick, J.M., Watson, R.W.G., 2011. Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. J. Proteome Res. 10, 1361–1373. https://doi.org/10.1021/pr1011069

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., 2015. Efficient and Robust Automated Machine Learning, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.) Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 2962–2970.

Fiehn, O., 2016. Metabolomics by Gas Chromatography-Mass Spectrometry: the combination of targeted and untargeted profiling. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 114, 30.4.1-30.4.32. https://doi.org/10.1002/0471142727.mb3004s114

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., 2010. The Pfam protein families database. Nucleic Acids Res. 38, D211–D222. https://doi.org/10.1093/nar/gkp985

Foster, J.M., Moreno, P., Fabregat, A., Hermjakob, H., Steinbeck, C., Apweiler, R., Wakelam, M.J.O., Vizcaíno, J.A., 2013. LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0061951

Foster, K.R., Koprowski, R., Skufca, J.D., 2014. Machine learning, medical diagnosis, and biomedical engineering research - commentary. Biomed. Eng. OnLine 13, 94. https://doi.org/10.1186/1475-925X-13-94

Francescatto, M., Chierici, M., Rezvan Dezfooli, S., Zandonà, A., Jurman, G., Furlanello, C., 2018. Multi-omics integration for neuroblastoma clinical endpoint prediction. Biol. Direct 13, 5. https://doi.org/10.1186/s13062-018-0207-8

Fridley, B.L., Lund, S., Jenkins, G.D., Wang, L., 2012. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet. Epidemiol. 36, 352–359. https://doi.org/10.1002/gepi.21628

Gammerman, A., 2010. Modern Machine Learning Techniques and Their Applications to Medical Diagnostics, in: Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology. Presented at the IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, Berlin, Heidelberg, pp. 2–2. https://doi.org/10.1007/978-3-642-16239-8_2

Garali, I., Adanyeguh, I.M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., Moszer, I., Guillemot, V., Durr, A., Mochel, F., Tenenhaus, A., 2018. A strategy for multimodal

data integration: application to biomarkers identification in spinocerebellar ataxia. Brief. Bioinform. 19, 1356–1369. https://doi.org/10.1093/bib/bbx060

Giang, T.-T., Nguyen, T.-P., Tran, D.-H., 2020. Stratifying patients using fast multiple kernel learning framework: case studies of Alzheimer's disease and cancers. BMC Med. Inform. Decis. Mak. 20, 108. https://doi.org/10.1186/s12911-020-01140-y

Gibson, G., 2015. A Primer of Human Genetics, 1st ed. 2015 edition. ed. Sinauer, Sunderland, Massachusetts, U.S.A.

Gibson, G., Marigorta, U.M., Ojagbeghru, E.R., Park, S., 2015. PART of the WHOLE: A Case Study in Wellness-Oriented Personalized Medicine. Yale J. Biol. Med. 88, 397–406.

Giovanni Martinelli, Nicholas Foreman, Sean Sanders, 2015. Advancing precision medicine through multi-omics: An integrated approach to tumor profiling.

Glaves, J.P., Li, M.X., Mercier, P., Fahlman, R.P., Sykes, B.D., 2014. High-throughput, multi-platform metabolomics on very small volumes: 1H NMR metabolite identification in an unadulterated tube-in-tube system. Metabolomics 10, 1145–1151. https://doi.org/10.1007/s11306-014-0678-2

Gligorijević, V., Malod-Dognin, N., Pržulj, N., 2016. Integrative methods for analyzing big data in precision medicine. PROTEOMICS 16, 741–758. https://doi.org/10.1002/pmic.201500396

Gligorijević, V., Pržulj, N., 2015. Methods for biological data integration: perspectives and challenges. J. R. Soc. Interface 12, 20150572. https://doi.org/10.1098/rsif.2015.0571

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI—Explainable artificial intelligence. Sci. Robot. 4. https://doi.org/10.1126/scirobotics.aay7120

Guo, L.-Y., Wu, A.-H., Wang, Y., Zhang, L., Chai, H., Liang, X.-F., 2020. Deep learning-based ovarian cancer subtypes identification using multi-omics data. BioData Min. 13, 10. https://doi.org/10.1186/s13040-020-00222-x

Gupta, Y., Lama, R.K., Kwon, G.-R., Initiative, A.D.N., Weiner, M.W., Aisen, P., Weiner, M., Aisen, P., Petersen, R., Jack, C.R.J., Jagust, W., Trojanowki, J.Q., Toga, A.W., Beckett, L., Green, R.C., Saykin, A.J., Morris, J., Shaw, L.M., Khachaturian, Z., Sorensen, G., Carrillo, M., Kuller, L., Raichle, M., Paul, S., Davies, P., Fillit, H., Hefti, F., Holtzman, D., Mesulam, M.M., Potter, W., Snyder, P., Schwartz, A., Green, R.C., Montine, T., Petersen, R., Aisen, P., Thomas, R.G., Donohue, Michael, Walter, S., Gessert, D., Sather, T., Jiminez, G., Balasubramanian, A.B., Mason, J., Sim, I., Beckett, L., Harvey, D., Donohue, Michael, Jack, C.R.J., Bernstein, M., Fox, N., Thompson, P., Schuff, N., DeCArli, C., Borowski, B., Gunter, J., Senjem, M., Vemuri, P., Jones, D., Kantarci, K., Ward, C., Jagust, W., Koeppe, R.A., Foster, N., Reiman, E.M., Chen, K., Mathis, C., Landau, S., Morris, J.C., Cairns, N.J., Franklin, E., Taylor-Reinwald, L., Shaw, L.M., Trojanowki, J.Q., Lee, V., Korecka, M., Figurski, M., Toga, A.W., Crawford, K., Neu, S., Saykin, A.J., Foroud, T.M., Potkin, S., Shen, L., Faber, K., Kim, S., Nho, K., Weiner, M.W., Thal, Lean, Khachaturian, Z., Thal, Leon, Buckholtz, N., Weiner, M.W., Snyder, P.J., Potter, W., Paul, S., Albert, M., Frank, R., Khachaturian, Z., Hsiao, J., Kaye, J., Quinn, J., Silbert, L., Lind, B., Carter, R., Dolen, S., Schneider, L.S., Pawluczyk, S., Becerra, M., Teodoro, L., Spann, B.M., Brewer, J., Vanderswag, H., Fleisher, A., Heidebrink, J.L., Lord, J.L., Petersen, R., Mason, S.S., Albers, C.S., Knopman, D., Johnson, Kris, Doody, R.S., Villanueva-Meyer, J., Pavlik, V., Shibley, V., Chowdhury, M., Rountree, S., Dang, M., Stern, Y., Honig, L.S., Bell, K.L., Ances, B., Morris, J.C., Carroll, M., Creech, M.L., Franklin, E., Mintun, M.A., Schneider, S., Oliver, A., Marson, D., Geldmacher, D., Natelson Love, M., Griffith, R., Clark, D., Brockington, J., Roberson, E., Grossman, H., Mitsis, E., Shah, R.C., deToledo-Morrell, L., Duara, R., Greig-Custo, M.T., Barker, W., Albert,

M., Onyike, C., D'Agostino, D.I., Kielb, S., Sadowski, M., Sheikh, M.O., Anaztasia, U., Mrunalini, G., Doraiswamy, P.M., Petrella, J.R., Borges-Neto, S., Wong, T.Z., Coleman, E., Arnold, S.E., Karlawish, J.H., Wolk, D.A., Clark, C.M., Smith, C.D., Jicha, G., Hardy, P., Sinha, P., Oates, E., Conrad, G., Lopez, O.L., Oakley, M., Simpson, D.M., Porsteinsson, A.P., Goldstein, B.S., Martin, K., Makino, K.M., Ismail, M.S., Brand, C., Potkin, S.G., Preda, A., Nguyen, D., Womack, K., Mathews, D., Quiceno, M., Levey, A.I., Lah, J.J., Cellar, J.S., Burns, J.M., Swerdlow, R.H., Brooks, W.M., Apostolova, L., Tingus, K., Woo, E., Silverman, D.H.S., Lu, P.H., Bartzokis, G., Graff-Radford, N.R., Parfitt, F., Poki-Walker, K., Farlow, M.R., Hake, A.M., Matthews, B.R., Brosch, J.R., Herring, S., van Dyck, C.H., Carson, R.E., MacAvoy, M.G., Varma, P., Chertkow, H., Bergman, H., Hosein, C., Black, S., Stefanovic, B., Caldwell, C., Hsiung, G.-Y.R., Mudge, B., Sossi, V., Feldman, H., Assaly, M., Finger, E., Pasternack, S., Rachisky, I., Trost, D., Kertesz, A., Bernick, C., Munic, D., Mesulam, M.-M., Rogalski, E., Lipowski, K., Weintraub, S., Bonakdarpour, B., Kerwin, D., Wu, C.-K., Johnson, N., Sadowsky, C., Villena, T., Turner, R.S., Johnson, Kathleen, Reynolds, B., Sperling, R.A., Johnson, K.A., Marshall, G., Yesavage, J., Taylor, J.L., Lane, B., Rosen, A., Tinklenberg, J., Sabbagh, M.N., Belden, C.M., Jacobson, S.A., Sirrel, S.A., Kowall, N., Killiany, R., Budson, A.E., Norbash, A., Johnson, P.L., Obisesan, T.O., Wolday, S., Allard, J., Lerner, A., Ogrocki, P., Tatsuoka, C., Fatica, P., Fletcher, E., Maillard, P., Olichney, J., DeCarli, C., Carmichael, O., Kittur, S., Borrie, M., Lee, T.-Y., Bartha, R., Johnson, S., Asthana, S., Carlsson, C.M., Potkin, S.G., Preda, A., Nguyen, D., Tariot, P., Burke, A., Milliken, A.M., Trncic, N., Fleisher, A., Reeder, S., Bates, V., Capote, H., Rainka, M., Scharre, D.W., Kataki, M., Kelley, B., Zimmerman, E.A., Celmins, D., Brown, A.D., Pearlson, G.D., Blank, K., Anderson, K., Flashman, L.A., Seltzer, M., Hynes, M.L., Santulli, R.B., Sink, K.M., Leslie, G., Williamson, J.D., Garg, P., Watkins, F., Ott, B.R., Tremont, G., Daiello, L.A., Salloway, S., Malloy, P., Correia, S., Rosen, H.J., Miller, B.L., Perry, D., Mintzer, J., Spicer, K., Bachman, D., Finger, E., Pasternak, S., Rachinsky, I., Rogers, J., Kertesz, A., Drost, D., Pomara, N., Hernando, R., Sarrael, A., Schultz, S.K., Smith, K.E., Koleva, H., Nam, K.W., Shim, H., Relkin, N., Chiang, G., Lin, M., Ravdin, L., Smith, A., Raj, B.A., Fargher, K., Weiner, M.W., Aisen, P., Weiner, M., Aisen, P., Petersen, R., Green, R.C., Harvey, D., Jack, C.R.J., Jagust, W., Morris, J.C., Saykin, A.J., Shaw, L.M., Toga, A.W., Trojanowki, J.Q., Neylan, T., Grafman, J., Green, R.C., Montine, T., Weiner, M., Petersen, R., Aisen, P., Thomas, R.G., Donohue, Michael, Devon, G., Sather, T., Melissa, D., Morrison, R., Jiminez, G., Neylan, T., Jacqueline, H., Shannon, F., Harvey, D., Donohue, Michael, Jack, C.R.J., Bernstein, M., Borowski, B., Gunter, J., Senjem, M., Kejal, K., Chad, W., Jagust, W., Koeppe, R.A., Foster, N., Reiman, E.M., Chen, K., Landau, S., Morris, J.C., Cairns, N.J., Householder, E., Shaw, L.M., Trojanowki, J.Q., Lee, V., Korecka, M., Figurski, M., Toga, A.W., Karen, C., Scott, N., Saykin, A.J., Foroud, T.M., Potkin, S., Shen, L., Faber, K., Kim, S., Nho, K., Weiner, M.W., Karl, F., Schneider, L.S., Pawluczyk, S., Mauricio, B., Brewer, J., Vanderswag, H., Stern, Y., Honig, L.S., Bell, K.L., Fleischman, D., Arfanakis, K., Shah, R.C., Duara, R., Varon, D., Greig, M.T., Doraiswamy, P.M., Petrella, J.R., James, O., Porsteinsson, A.P., Goldstein, B., Martin, K.S., Potkin, S.G., Preda, A., Nguyen, D., Mintzer, J., Massoglia, D., Brawman-Mintzer, O., Sadowsky, C., Martinez, W., Villena, T., Jagust, W., Landau, S., Rosen, H., Perry, D., Turner, R.S., Behan, K., Reynolds, B., Sperling, R.A., Johnson, K.A., Marshall, G., Sabbagh, M.N., Jacobson, S.A., Sirrel, S.A., Obisesan, T.O., Wolday, S., Allard, J., Johnson, S.C., Fruehling, J.J., Harding, S., Peskind, E.R., Petrie, E.C., Li, G., Yesavage, J.A., Taylor, J.L.,

Furst, A.J., Chao, S., Relkin, N., Chiang, G., Ravdin, L., Ravdin, L., Mackin, S., Aisen, P., Raman, R., Mackin, S., Weiner, M., Aisen, P., Raman, R., Jack, C.R.J., Landau, S., Saykin, A.J., Toga, A.W., DeCarli, C., Koeppe, R.A., Green, R.C., Drake, E., Weiner, M., Aisen, P., Raman, R., Donohue, Mike, Jimenez, G., Gessert, D., Harless, K., Salazar, J., Cabrera, Y., Walter, S., Hergesheimer, L., Shaffer, E., Mackin, S., Nelson, C., Bickford, D., Butters, M., Zmuda, M., Jack, C.R.J., Bernstein, M., Borowski, B., Gunter, J., Senjem, M., Kantarci, K., Ward, C., Reyes, D., Koeppe, R.A., Landau, S., Toga, A.W., Crawford, K., Neu, S., Saykin, A.J., Foroud, T.M., Faber, K.M., Nho, K., Nudelman, K.N., Mackin, S., Rosen, H., Nelson, C., Bickford, D., Au, Y.H., Scherer, K., Catalinotto, D., Stark, S., Ong, E., Fernandez, D., Butters, M., Zmuda, M., Lopez, O.L., Oakley, M., Simpson, D.M., 2019. Prediction and Classification of Alzheimer's Disease Based on Combined Features From Apolipoprotein-E Genotype, Cerebrospinal Fluid, MR, and FDG-PET Imaging Biomarkers. Front. Comput. Neurosci. 13. https://doi.org/10.3389/fncom.2019.00072

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J Mach Learn Res 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. 46, 389–422. https://doi.org/10.1023/A:1012487302797

Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., Ralser, M., 2017. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology. Curr. Opin. Syst. Biol. 6, 37–45. https://doi.org/10.1016/j.coisb.2017.08.009

Haibo He, Yang Bai, Garcia, E.A., Shutao Li, 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Presented at the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hamamoto, R., Suvarna, K., Yamada, M., Kobayashi, K., Shinkai, N., Miyake, M., Takahashi, M., Jinnai, S., Shimoyama, R., Sakai, A., Takasawa, K., Bolatkan, A., Shozu, K., Dozen, A., Machino, H., Takahashi, S., Asada, K., Komatsu, M., Sese, J., Kaneko, S., 2020. Application of Artificial Intelligence Technology in Oncology: Towards the Establishment of Precision Medicine. Cancers 12, 3532. https://doi.org/10.3390/cancers12123532

Hampel, H., O'Bryant, S.E., Castrillo, J.I., Ritchie, C., Rojkova, K., Broich, K., Benda, N., Nisticò, R., Frank, R.A., Dubois, B., Escott-Price, V., Lista, S., 2016. PRECISION MEDICINE - The Golden Gate for Detection, Treatment and Prevention of Alzheimer's Disease. J. Prev. Alzheimers Dis. 3, 243–259. https://doi.org/10.14283/jpad.2016.112

Hampel, H., O'Bryant, S.E., Durrleman, S., Younesi, E., Rojkova, K., Escott-Price, V., Corvol, J.-C., Broich, K., Dubois, B., Lista, S., Alzheimer Precision Medicine Initiative, 2017. A Precision Medicine Initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. Climacteric J. Int. Menopause Soc. 20, 107–118. https://doi.org/10.1080/13697137.2017.1287866

Hanania, N.A., Diamant, Z., 2017. The road to precision medicine in asthma: challenges and opportunities. Curr. Opin. Pulm. Med. https://doi.org/10.1097/MCP.0000000000000444

Handa, A., Sharma, A., Shukla, S.K., 2019. Machine learning in cybersecurity: A review. WIREs Data Min. Knowl. Discov. 9, e1306. https://doi.org/10.1002/widm.1306

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

Hasin, Y., Seldin, M., Lusis, A., 2017. Multi-omics approaches to disease. Genome Biol. 18, 83. https://doi.org/10.1186/s13059-017-1215-1

He, H., Lin, D., Zhang, J., Wang, Y., Deng, H.-W., 2016. Biostatistics, Data Mining and Computational Modeling, in: Application of Clinical Bioinformatics, Translational Bioinformatics. Springer, Dordrecht, pp. 23–57. https://doi.org/10.1007/978-94-017-7543-4_2

Health, C. for D. and R., 2021. Artificial Intelligence and Machine Learning in Software as a Medical Device [WWW Document]. FDA. URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed 2.1.21).

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., Boulesteix, A.-L., 2020. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief. Bioinform. https://doi.org/10.1093/bib/bbaa167

Holzinger, E.R., Dudek, S.M., Frase, A.T., Pendergrass, S.A., Ritchie, M.D., 2014. ATHENA: the analysis tool for heritable and environmental network associations. Bioinforma. Oxf. Engl. 30, 698–705. https://doi.org/10.1093/bioinformatics/btt572

Hristoskova, A., Boeva, V., Tsiporkova, E., 2014. A formal concept analysis approach to consensus clustering of multi-experiment expression data. BMC Bioinformatics 15, 151. https://doi.org/10.1186/1471-2105-15-151

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., Qu, S., Mei, X., Chen, H., Yu, T., Sun, N., Rao, J., Wang, J., Zhang, W., Chen, Y., Liao, S., Jiang, H., Liu, X., Yang, Z., Mu, F., Gao, S., 2017. A reference human genome dataset of the BGISEQ-500 sequencer. GigaScience 6, 1–9. https://doi.org/10.1093/gigascience/gix024

Huang, S., Chaudhary, K., Garmire, L.X., 2017. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front. Genet. 8. https://doi.org/10.3389/fgene.2017.00084

Hugenholtz, P., Tyson, G.W., 2008. Metagenomics. Nature 455, 481–483. https://doi.org/10.1038/455481a

Hung, A.J., 2019. Can machine-learning algorithms replace conventional statistics? BJU Int. 123, 1–1. https://doi.org/10.1111/bju.14542

Hwang, B., Lee, J.H., Bang, D., 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. 50, 1–14. https://doi.org/10.1038/s12276-018-0071-8

James, G., Witten, D., Hastie, T., Tibshirani, R., 2017. An Introduction to Statistical Learning: with Applications in R, 1st ed. 2013, Corr. 7th printing 2017 edition. ed. Springer, New York.

Jamil, I.N., Remali, J., Azizan, K.A., Nor Muhammad, N.A., Arita, M., Goh, H.-H., Aizat, W.M., 2020. Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.00944

Jeni, L.A., Cohn, J.F., De La Torre, F., 2013. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Presented at the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245–251. https://doi.org/10.1109/ACII.2013.47

Jiang, T., Gradus, J.L., Rosellini, A.J., 2020. Supervised Machine Learning: A Brief Primer. Behav. Ther. 51, 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed. Springer.
Kalaitzopoulos, D., 2016. The potential of Precision Medicine. New Horiz. Transl. Med. 3, 63–65. https://doi.org/10.1016/j.nhtm.2016.05.001
Kalvari, I., Nawrocki, E.P., Argasinska, J., Quinones-Olvera, N., Finn, R.D., Bateman, A., Petrov, A.I., 2018. Non-Coding RNA Analysis Using the Rfam Database. Curr. Protoc. Bioinforma. 62, e51. https://doi.org/10.1002/cpbi.51
Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27
Karpievitch, Y.V., Polpitiya, A.D., Anderson, G.A., Smith, R.D., Dabney, A.R., 2010. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. Ann. Appl. Stat. 4, 1797–1823. https://doi.org/10.1214/10-AOAS341
Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Comput. Struct. Biotechnol. J. 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
Kim, A.A., Rachid Zaim, S., Subbian, V., 2020. Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data. Int. J. Med. Inf. 141, 104148. https://doi.org/10.1016/j.ijmedinf.2020.104148
Kim, D., Joung, J.-G., Sohn, K.-A., Shin, H., Park, Y.R., Ritchie, M.D., Kim, J.H., 2015. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J. Am. Med. Inform. Assoc. 22, 109–120. https://doi.org/10.1136/amiajnl-2013-002481
Kim, M., Tagkopoulos, I., 2018. Data integration and predictive modeling methods for multi-omics datasets. Mol. Omics 14, 8–25. https://doi.org/10.1039/C7MO00051K
Kim, S., Jhong, J.-H., Lee, J., Koo, J.-Y., 2017. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 10, 2. https://doi.org/10.1186/s13040-017-0126-8
Kim, S., Park, T., Kon, M., 2014. Cancer survival classification using integrated data sets and intermediate information. Artif. Intell. Med. 62, 23–31. https://doi.org/10.1016/j.artmed.2014.06.003
Kirchebner, J., Günther, M.P., Sonnweber, M., King, A., Lau, S., 2020. Factors and predictors of length of stay in offenders diagnosed with schizophrenia - a machine-learning-based approach. BMC Psychiatry 20. https://doi.org/10.1186/s12888-020-02612-1
Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L., 2012. Bayesian correlated clustering to integrate multiple datasets. Bioinforma. Oxf. Engl. 28, 3290–3297. https://doi.org/10.1093/bioinformatics/bts595
Knittelfelder, O.L., Weberhofer, B.P., Eichmann, T.O., Kohlwein, S.D., Rechberger, G.N., 2014. A versatile ultra-high performance LC-MS method for lipid profiling. J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. 951–952, 119–128. https://doi.org/10.1016/j.jchromb.2014.01.011
Kodama, Y., Shumway, M., Leinonen, R., 2012. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 40, D54–D56. https://doi.org/10.1093/nar/gkr854
Köfeler, H.C., Fauland, A., Rechberger, G.N., Trötzmüller, M., 2012. Mass Spectrometry Based Lipidomics: An Overview of Technological Platforms. Metabolites 2, 19–38. https://doi.org/10.3390/metabo2010019
Kohl, M., Megger, D.A., Trippler, M., Meckel, H., Ahrens, M., Bracht, T., Weber, F., Hoffmann, A.-C., Baba, H.A., Sitek, B., Schlaak, J.F., Meyer, H.E., Stephan, C., Eisenacher, M., 2014. A practical data processing workflow for multi-OMICS

projects. Biochim. Biophys. Acta BBA - Proteins Proteomics, Computational Proteomics in the Post-Identification Era 1844, 52–62. https://doi.org/10.1016/j.bbapap.2013.02.029

Kovacs, G.G., 2016. Molecular Pathological Classification of Neurodegenerative Diseases: Turning towards Precision Medicine. Int. J. Mol. Sci. 17. https://doi.org/10.3390/ijms17020189

Kozomara, A., Birgaoanu, M., Griffiths-Jones, S., 2019. miRBase: from microRNA sequences to function. Nucleic Acids Res. 47, D155–D162. https://doi.org/10.1093/nar/gky1141

Kuo, T.-C., Tseng, Y.J., 2018. LipidPedia: a comprehensive lipid knowledgebase. Bioinformatics 34, 2982–2987. https://doi.org/10.1093/bioinformatics/bty213

Kuska, B., 1998. Beer, Bethesda, and biology: how “genomics” came into being. J. Natl. Cancer Inst. 90, 93.

Kwon, M.-S., Kim, Y., Lee, S., Namkung, J., Yun, T., Yi, S.G., Han, S., Kang, M., Kim, S.W., Jang, J.-Y., Park, T., 2015. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. BMC Genomics 16 Suppl 9, S4. https://doi.org/10.1186/1471-2164-16-S9-S4

L, P., H, D., Md, A., Z, B., M, W., L, Y., B, B., R, R., S, S., M, A.-K., P, F., M, J., M, K., I, P., 2018. Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Health. https://doi.org/10.1101/353390

Lambin, P., Leijenaar, R.T.H., Deist, T.M., Peerlings, J., de Jong, E.E.C., van Timmeren, J., Sanduleanu, S., Larue, R.T.H.M., Even, A.J.G., Jochems, A., van Wijk, Y., Woodruff, H., van Soest, J., Lustberg, T., Roelofs, E., van Elmpt, W., Dekker, A., Mottaghy, F.M., Wildberger, J.E., Walsh, S., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. https://doi.org/10.1038/nrclinonc.2017.141

Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S., 2004. A statistical framework for genomic data fusion. Bioinforma. Oxf. Engl. 20, 2626–2635. https://doi.org/10.1093/bioinformatics/bth294

Le, N., Sund, M., Vinci, A., 2016. Prognostic and predictive markers in pancreatic adenocarcinoma. Dig. Liver Dis. 48, 223–230. https://doi.org/10.1016/j.dld.2015.11.001

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lee, C.H., Yoon, H.-J., 2017. Medical big data: promise and challenges. Kidney Res. Clin. Pract. 36, 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3

Lee, G., Bang, L., Kim, S.Y., Kim, D., Sohn, K.-A., 2017. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. BMC Med. Genomics 10, 28. https://doi.org/10.1186/s12920-017-0268-z

Lee, I.-H., Lushington, G.H., Visvanathan, M., 2011. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. J. Clin. Bioinforma. 1, 11. https://doi.org/10.1186/2043-9113-1-11

Lee, J.K. (Ed.), 2010. Statistical bioinformatics: a guide for life and biomedical science researchers. Wiley-Blackwell, Hoboken, N.J.

Lee, T.-Y., Huang, K.-Y., Chuang, C.-H., Lee, C.-Y., Chang, T.-H., 2020. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. Comput. Biol. Chem. 87, 107277. https://doi.org/10.1016/j.compbiolchem.2020.107277

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N.,

Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., Cochrane, G., 2011. The European Nucleotide Archive. Nucleic Acids Res. 39, D28–D31. https://doi.org/10.1093/nar/gkq967

Lévesque, E., Kirby, E., Bolt, I., Knoppers, B.M., de Beaufort, I., Pashayan, N., Widschwendter, M., 2018. Ethical, Legal, and Regulatory Issues for the Implementation of Omics-Based Risk Prediction of Women's Cancer: Points to Consider. Public Health Genomics 21, 37–44. https://doi.org/10.1159/000492663

Li, M., Wang, Y., Zheng, R., Shi, X., li, y, Wu, F., Wang, J., 2019. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Trans. Comput. Biol. Bioinform. 1–1. https://doi.org/10.1109/TCBB.2019.2919581

Li, Q., Freeman, L.M., Rush, J.E., Huggins, G.S., Kennedy, A.D., Labuda, J.A., Laflamme, D.P., Hannah, S.S., 2015. Veterinary Medicine and Multi-Omics Research for Future Nutrition Targets: Metabolomics and Transcriptomics of the Common Degenerative Mitral Valve Disease in Dogs. Omics J. Integr. Biol. 19, 461–470. https://doi.org/10.1089/omi.2015.0057

Li, S., Chen, X., Liu, X., Yu, Y., Pan, H., Haak, R., Schmidt, J., Ziebolz, D., Schmalz, G., 2017. Complex integrated analysis of lncRNAs-miRNAs-mRNAs in oral squamous cell carcinoma. Oral Oncol. 73, 1–9. https://doi.org/10.1016/j.oraloncology.2017.07.026

Li, W., Zhang, A., Zhou, X., Nan, Y., Liu, Q., Sun, H., Fang, H., Wang, X., 2020. High-throughput liquid chromatography mass-spectrometry-driven lipidomics discover metabolic biomarkers and pathways as promising targets to reveal the therapeutic effects of the Shenqi pill. RSC Adv. 10, 2347–2358. https://doi.org/10.1039/C9RA07621D

Liew, A.W.-C., Law, N.-F., Yan, H., 2011. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief. Bioinform. 12, 498–513. https://doi.org/10.1093/bib/bbq080

Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E., Blayney, J.K., 2019. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief. Bioinform. 20, 1795–1811. https://doi.org/10.1093/bib/bby051

Lin, E., Lane, H.-Y., 2017. Machine learning and systems genomics approaches for multi-omics data. Biomark. Res. 5, 2. https://doi.org/10.1186/s40364-017-0082-y

Lindon, J.C., Nicholson, J.K., Holmes, E., 2011. The Handbook of Metabonomics and Metabolomics. Elsevier.

List, M., Hauschild, A.-C., Tan, Q., Kruse, T.A., Mollenhauer, J., Baumbach, J., Batra, R., 2014. Classification of breast cancer subtypes by combining gene expression and DNA methylation data. J. Integr. Bioinforma. 11, 236. https://doi.org/10.2390/biecoll-jib-2014-236

Liu, Y., Ding, J., Reynolds, L.M., Lohman, K., Register, T.C., De La Fuente, A., Howard, T.D., Hawkins, G.A., Cui, W., Morris, J., Smith, S.G., Barr, R.G., Kaufman, J.D., Burke, G.L., Post, W., Shea, S., Mccall, C.E., Siscovick, D., Jacobs, D.R., Tracy, R.P., Herrington, D.M., Hoeschele, I., 2013. Methylomics of gene expression in human monocytes. Hum. Mol. Genet. 22, 5065–5074. https://doi.org/10.1093/hmg/ddt356

Lock, E.F., Dunson, D.B., 2013. Bayesian consensus clustering. Bioinformatics 29, 2610–2616. https://doi.org/10.1093/bioinformatics/btt425

Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B., 2013. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann. Appl. Stat. 7, 523–542. https://doi.org/10.1214/12-AOAS597

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000. Molecular Cell Biology, 4th ed. W. H. Freeman.

López de Maturana, E., Alonso, L., Alarcón, P., Martín-Antoniano, I.A., Pineda, S., Piorno, L., Calle, M.L., Malats, N., 2019. Challenges in the Integration of Omics and Non-Omics Data. Genes 10, 238. https://doi.org/10.3390/genes10030238

López Pineda, A., Ye, Y., Visweswaran, S., Cooper, G.F., Wagner, M.M., Tsui, F. (Rich), 2015. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J. Biomed. Inform. 58, 60–69. https://doi.org/10.1016/j.jbi.2015.08.019

Lorena, A.C., Jacintho, L.F.O., Siqueira, M.F., Giovanni, R.D., Lohmann, L.G., de Carvalho, A.C.P.L.F., Yamamoto, M., 2011. Comparing machine learning classifiers in potential distribution modelling. Expert Syst. Appl. 38, 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., 2017. Transcriptomics technologies. PLOS Comput. Biol. 13, e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Lu, J., Cowperthwaite, M.C., Burnett, M.G., Shpak, M., 2016. Molecular Predictors of Long-Term Survival in Glioblastoma Multiforme Patients. PloS One 11, e0154313. https://doi.org/10.1371/journal.pone.0154313

Luck, K., Sheynkman, G.M., Zhang, I., Vidal, M., 2017. Proteome-Scale Human Interactomics. Trends Biochem. Sci. 42, 342–354. https://doi.org/10.1016/j.tibs.2017.02.006

Lussier, Y.A., Li, H., 2012. Breakthroughs in genomics data integration for predicting clinical outcome. J. Biomed. Inform. 45, 1190–1201. https://doi.org/10.1016/j.jbi.2012.10.003

Ma, A., McDermaid, A., Xu, J., Chang, Y., Ma, Q., 2020. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. Trends Biotechnol. 38, 1007–1022. https://doi.org/10.1016/j.tibtech.2020.02.013

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., Song, F., 2020. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput. Biol. Med. 121, 103761. https://doi.org/10.1016/j.compbiomed.2020.103761

Ma, S., Ren, J., Fenyö, D., 2016. Breast Cancer Prognostics Using Multi-Omics Data. AMIA Summits Transl. Sci. Proc. 2016, 52–59.

Malod-Dognin, N., Petschnigg, J., Pržulj, N., 2017. Precision medicine — a promising, yet challenging road lies ahead. Curr. Opin. Syst. Biol. https://doi.org/10.1016/j.coisb.2017.10.003

Mamoshina, P., Volosnikova, M., Ozerov, I.V., Putin, E., Skibina, E., Cortese, F., Zhavoronkov, A., 2018. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Front. Genet. 9. https://doi.org/10.3389/fgene.2018.00242

Mandel, S.A., Morelli, M., Halperin, I., Korczyn, A.D., 2010. Biomarkers for prediction and targeted prevention of Alzheimer's and Parkinson's diseases: evaluation of drug clinical efficacy. EPMA J. 1, 273–292. https://doi.org/10.1007/s13167-010-0036-z

Mankoo, P.K., Shen, R., Schultz, N., Levine, D.A., Sander, C., 2011. Time to Recurrence and Survival in Serous Ovarian Tumors Predicted from Integrated Genomic Profiles. PLOS ONE 6, e24709. https://doi.org/10.1371/journal.pone.0024709

Margolies, L.R., Pandey, G., Horowitz, E.R., Mendelson, D.S., 2016. Breast Imaging in the Era of Big Data: Structured Reporting and Data Mining. AJR Am. J. Roentgenol. 206, 259–264. https://doi.org/10.2214/AJR.15.15396

Martinez, A.M., Kak, A.C., 2001. PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. 23, 228–233. https://doi.org/10.1109/34.908974

McCabe, S.D., Lin, D.-Y., Love, M.I., 2020. Consistency and overfitting of multi-omics methods on experimental data. Brief. Bioinform. 21, 1277–1284. https://doi.org/10.1093/bib/bbz070

McCarthy, M.I., 2017. Painting a new picture of personalised medicine for diabetes. Diabetologia 60, 793–799. https://doi.org/10.1007/s00125-017-4210-x

McShane, L.M., Cavenagh, M.M., Lively, T.G., Eberhard, D.A., Bigbee, W.L., Williams, P.M., Mesirov, J.P., Polley, M.-Y.C., Kim, K.Y., Tricoli, J.V., Taylor, J.M., Shuman, D.J., Simon, R.M., Doroshow, J.H., Conley, B.A., 2013a. Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. BMC Med. 11, 220. https://doi.org/10.1186/1741-7015-11-220

McShane, L.M., Cavenagh, M.M., Lively, T.G., Eberhard, D.A., Bigbee, W.L., Williams, P.M., Mesirov, J.P., Polley, M.-Y.C., Kim, K.Y., Tricoli, J.V., Taylor, J.M.G., Shuman, D.J., Simon, R.M., Doroshow, J.H., Conley, B.A., 2013b. Criteria for the use of omics-based predictors in clinical trials. Nature 502, 317–320. https://doi.org/10.1038/nature12564

Memon, J., Sami, M., Khan, R.A., Uddin, M., 2020. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). IEEE Access 8, 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

Meng, C., Helm, D., Frejno, M., Kuster, B., 2016a. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J. Proteome Res. 15, 755–765. https://doi.org/10.1021/acs.jproteome.5b00824

Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M., Culhane, A.C., 2016b. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. bbv108. https://doi.org/10.1093/bib/bbv108

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., Rinn, J.L., 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. 30, 99–104. https://doi.org/10.1038/nbt.2024

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386. https://doi.org/10.1186/1471-2105-9-386

Milward, E.A., Shahandeh, A. Heidari, M., Johnstone, D.M., Daneshi, N., Hondermarck, H., 2016. Transcriptomics, in: Encyclopedia of Cell Biology. Academic Press, Waltham, pp. 160–165. https://doi.org/10.1016/B978-0-12-394447-4.40029-5

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., Ping, P., 2019. Machine Learning and Integrative Analysis of Biomedical Big Data. Genes 10, 87. https://doi.org/10.3390/genes10020087

Misra, B.B., Langefeld, C., Olivier, M., Cox, L.A., 2019. Integrated omics: tools, advances and future approaches. J. Mol. Endocrinol. R21–R45. https://doi.org/10.1530/JME-18-0055

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48, D570–D578. https://doi.org/10.1093/nar/gkz1035

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., Hilsenbeck, S.G., 2018. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 19, 71–86. https://doi.org/10.1093/biostatistics/kxx017

Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., Shen, R., 2013. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. 110, 4245–4250. https://doi.org/10.1073/pnas.1208949110

Mostafavi, S., Morris, Q., 2010. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics 26, 1759–1765. https://doi.org/10.1093/bioinformatics/btq262

Mougin, F., Auber, D., Bourqui, R., Diallo, G., Dutour, I., Jouhet, V., Thiessard, F., Thiébaut, R., Thébault, P., 2018. Visualizing omics and clinical data: Which challenges for dealing with their variety? Methods, Comparison and Visualization Methods for High-Dimensional Biological Data 132, 3–18. https://doi.org/10.1016/j.ymeth.2017.08.012

Muehlematter, U.J., Daniore, P., Vokinger, K.N., 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit. Health 0. https://doi.org/10.1016/S2589-7500(20)30292-2

Mutie, P.M., Giordano, G.N., Franks, P.W., 2017. Lifestyle precision medicine: the next generation in type 2 diabetes prevention? BMC Med. 15, 171. https://doi.org/10.1186/s12916-017-0938-x

Nalejska, E., Mączyńska, E., Lewandowska, M.A., 2014. Prognostic and Predictive Biomarkers: Tools in Personalized Oncology. Mol. Diagn. Ther. 18, 273–284. https://doi.org/10.1007/s40291-013-0077-9

Nam, H., Chung, B.C., Kim, Y., Lee, K., Lee, D., 2009. Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. Bioinforma. Oxf. Engl. 25, 3151–3157. https://doi.org/10.1093/bioinformatics/btp558

Nguyen, H., Shrestha, S., Draghici, S., Nguyen, T., 2019. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics 35, 2843–2846. https://doi.org/10.1093/bioinformatics/bty1049

Nguyen, N.D., Wang, D., 2020. Multiview learning for understanding functional multiomics. PLOS Comput. Biol. 16, e1007677. https://doi.org/10.1371/journal.pcbi.1007677

Nguyen, T., Tagett, R., Diaz, D., Draghici, S., 2017. A novel approach for data integration and disease subtyping. Genome Res. 27, 2025–2039. https://doi.org/10.1101/gr.215129.116

Nicolai, M., Bühlmann Peter, 2010. Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72, 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., Bellazzi, R., 2020. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front. Oncol. 10. https://doi.org/10.3389/fonc.2020.01030

Nielsen, J., 2017. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. Cell Metab. 25, 572–579. https://doi.org/10.1016/j.cmet.2017.02.002

Obermeyer, Z., Emanuel, E.J., 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N. Engl. J. Med. 375, 1216–1219. https://doi.org/10.1056/NEJMp1606181

Olson, R.S., Sipper, M., Cava, W.L., Tartarone, S., Vitale, S., Fu, W., Orzechowski, P., Urbanowicz, R.J., Holmes, J.H., Moore, J.H., 2018. A System for Accessible Artificial Intelligence, in: Banzhaf, W., Olson, R.S., Tozier, W., Riolo, R. (Eds.), Genetic Programming Theory and Practice XV, Genetic and Evolutionary Computation. Springer International Publishing, Cham, pp. 121–134. https://doi.org/10.1007/978-3-319-90512-9_8

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J., 2020. Deep Learning vs. Traditional Computer Vision, in: Arai, K., Kapoor, S. (Eds.), Advances in Computer Vision. Springer International Publishing, Cham, pp. 128–144.

Overmyer, K.A., Shishkova, E., Miller, I.J., Balnis, J., Bernstein, M.N., Peters-Clarke, T.M., Meyer, J.G., Quan, Q., Muehlbauer, L.K., Trujillo, E.A., He, Y., Chopra, A., Chieng, H.C., Tiwari, A., Judson, M.A., Paulson, B., Brademan, D.R., Zhu, Y., Serrano, L.R., Linke, V., Drake, L.A., Adam, A.P., Schwartz, B.S., Singer, H.A., Swanson, S., Mosher, D.F., Stewart, R., Coon, J.J., Jaitovich, A., 2020. Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst. https://doi.org/10.1016/j.cels.2020.10.003

Paik, E.S., Choi, H.J., Kim, T.-J., Lee, J.-W., Kim, B.-G., Bae, D.-S., Choi, C.H., 2017. Molecular Signature for Lymphatic Invasion Associated with Survival of Epithelial Ovarian Cancer. Cancer Res. Treat. Off. J. Korean Cancer Assoc. https://doi.org/10.4143/crt.2017.104

Pérez-Cobas, A.E., Gomez-Valero, L., Buchrieser, C., 2020. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microb. Genomics 6, e000409. https://doi.org/10.1099/mgen.0.000409

Peterson, T.A., Doughty, E., Kann, M.G., 2013. Towards precision medicine: advances in computational approaches for the analysis of human variants. J. Mol. Biol. 425, 4047–4063. https://doi.org/10.1016/j.jmb.2013.08.008

Pfützner, A., Forst, T., 2006. High-sensitivity C-reactive protein as cardiovascular risk marker in patients with diabetes mellitus. Diabetes Technol. Ther. 8, 28–36. https://doi.org/10.1089/dia.2006.8.28

Pietzner, M., Engelmann, B., Kacprowski, T., Golchert, J., Dirk, A.-L., Hammer, E., Iwen, K.A., Nauck, M., Wallaschofski, H., Führer, D., Münte, T.F., Friedrich, N., Völker, U., Homuth, G., Brabant, G., 2017. Plasma proteome and metabolome characterization of an experimental human thyrotoxicosis model. BMC Med. 15, 6. https://doi.org/10.1186/s12916-016-0770-8

Pinu, F.R., Beale, D.J., Paten, A.M., Kouremenos, K., Swarup, S., Schirra, H.J., Wishart, D., 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites 9. https://doi.org/10.3390/metabo9040076

Poirion, O.B., Chaudhary, K., Garmire, L.X., 2018. Deep Learning data integration for better risk stratification models of bladder cancer. AMIA Summits Transl. Sci. Proc. 2018, 197–206.

Poirion, O.B., Chaudhary, K., Huang, S., Garmire, L.X., 2020. Multi-omics-based pan-cancer prognosis prediction using an ensemble of deep-learning and machine-learning models. medRxiv 19010082. https://doi.org/10.1101/19010082

Proteomics, transcriptomics: what's in a name?, 1999. . Nature 402, 715–715. https://doi.org/10.1038/45354

Pudil, P., Novovičová, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognit. Lett. 15, 1119–1125. https://doi.org/10.1016/0167-8655(94)90127-9

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rappoport, N., Shamir, R., 2019. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics 35, 3348–3356. https://doi.org/10.1093/bioinformatics/btz058

Rappoport, N., Shamir, R., 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 46, 10546–10562. https://doi.org/10.1093/nar/gky889

Rashidi, H.H., Tran, N.K., Betts, E.V., Howell, L.P., Green, R., 2019. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. Acad. Pathol. 6, 2374289519873088. https://doi.org/10.1177/2374289519873088

Ray, P., Zheng, L., Lucas, J., Carin, L., 2014. Bayesian joint analysis of heterogeneous genomics data. Bioinformatics 30, 1370–1376. https://doi.org/10.1093/bioinformatics/btu064

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-Throughput Sequencing Technologies. Mol. Cell 58, 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., Chinnaiyan, A.M., 2005. Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23, 951. https://doi.org/10.1038/nbt1103

Rimoldi, S.F., Scherrer, U., Messerli, F.H., 2014. Secondary arterial hypertension: when, who, and how to screen? Eur. Heart J. 35, 1245–1254. https://doi.org/10.1093/eurheartj/eht534

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D., 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16, 85–97. https://doi.org/10.1038/nrg3868

Rojas-Macias, M.A., Mariethoz, J., Andersson, P., Jin, C., Venkatakrishnan, V., Aoki, N.P., Shinmachi, D., Ashwood, C., Madunic, K., Zhang, T., Miller, R.L., Horlacher, O., Struwe, W.B., Watanabe, Y., Okuda, S., Levander, F., Kolarich, D., Rudd, P.M., Wuhrer, M., Kettner, C., Packer, N.H., Aoki-Kinoshita, K.F., Lisacek, F., Karlsson, N.G., 2019. Towards a standardized bioinformatics infrastructure for N - and O - glycomics. Nat. Commun. 10, 3275. https://doi.org/10.1038/s41467-019-11131-x

Roobaert, D., Karakoulas, G., Chawla, N.V., 2006. Information Gain, Correlation and Support Vector Machines, in: Feature Extraction, Studies in Fuzziness and Soft Computing. Springer, Berlin, Heidelberg, pp. 463–470. https://doi.org/10.1007/978-3-540-35488-8_23

Sakr, S., Elshawi, R., Ahmed, A.M., Qureshi, W.T., Brawner, C.A., Keteyian, S.J., Blaha, M.J., Al-Mallah, M.H., 2017. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. BMC Med. Inform. Decis. Mak. 17. https://doi.org/10.1186/s12911-017-0566-6

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sathyanarayanan, A., Gupta, R., Thompson, E.W., Nyholt, D.R., Bauer, D.C., Nagaraj, S.H., 2020. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief. Bioinform. 21, 1920–1936. https://doi.org/10.1093/bib/bbz121

Saulnier, K.M., Bujold, D., Dyke, S.O.M., Dupras, C., Beck, S., Bourque, G., Joly, Y., 2019. Benefits and barriers in the design of harmonized access agreements for international data sharing. Sci. Data 6, 297. https://doi.org/10.1038/s41597-019-0310-4

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. Off. J. Int. Neural Netw. Soc. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schumacher, A., Rujan, T., Hoefkens, J., 2014. A collaborative approach to develop a multi-omics data analytics platform for translational research. Appl. Transl. Genomics, Global Sharing of Genomic Knowledge in a Free Market 3, 105–108. https://doi.org/10.1016/j.atg.2014.09.010

Schwarz, D.F., König, I.R., Ziegler, A., 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 26, 1752–1758. https://doi.org/10.1093/bioinformatics/btq257

Seal, D.B., Das, V., Goswami, S., De, R.K., 2020. Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. Genomics 112, 2833–2841. https://doi.org/10.1016/j.ygeno.2020.03.021

Senft, D., Leiserson, M.D.M., Ruppin, E., Ronai, Z.A., 2017. Precision Oncology: The Road Ahead. Trends Mol. Med. 23, 874–898. https://doi.org/10.1016/j.molmed.2017.08.003

Seoane, J.A., Day, I.N.M., Gaunt, T.R., Campbell, C., 2014. A pathway-based data integration framework for prediction of disease progression. Bioinformatics 30, 838–845. https://doi.org/10.1093/bioinformatics/btt610

Sharifi-Noghabi, H., Zolotareva, O., Collins, C.C., Ester, M., 2019. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 35, i501–i509. https://doi.org/10.1093/bioinformatics/btz318

Shaw, A., Bradley, M.D., Elyan, S., Kurian, K.M. 2015. Tumour biomarkers: diagnostic, prognostic, and predictive. BMJ 351, h3449. https://doi.org/10.1136/bmj.h3449

Shen, H.-B., Chou, K.-C., 2006. Ensemble classifier for protein fold pattern recognition. Bioinforma. Oxf. Engl. 22, 1717–1722. https://doi.org/10.1093/bioinformatics/btl170

Shen, R., Olshen, A.B., Ladanyi, M., 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25, 2906–2912. https://doi.org/10.1093/bioinformatics/btp543

Shin, H., Hill, N.J., Lisewski, A.M., Park, J.-S., 2010. Graph Sharpening. Expert Syst Appl 37, 7870–7879. https://doi.org/10.1016/j.eswa.2010.04.050

Shin, H., Lisewski, A.M., Lichtarge, O., 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 23, 3217–3224. https://doi.org/10.1093/bioinformatics/btm511

Shrivastava, A.K., Singh, H.V., Raizada, A., Singh, S.K., 2015. C-reactive protein, inflammation and coronary heart disease. Egypt. Heart J. 67, 89–97. https://doi.org/10.1016/j.ehj.2014.11.005

Singhal, A., Simmons, M., Lu, Z., 2016. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. PLoS Comput. Biol. 12, e1005017. https://doi.org/10.1371/journal.pcbi.1005017

Sonsare, P.M., Gunavathi, C., 2019. Investigation of machine learning techniques on proteomics: A comprehensive survey. Prog. Biophys. Mol. Biol. 149, 54–69. https://doi.org/10.1016/j.pbiomolbio.2019.09.004

Sorzano, C.O.S., Vargas, J., Montano, A.P., 2014. A survey of dimensionality reduction techniques. ArXiv14032877 Cs Q-Bio Stat.

Speicher, N.K., Pfeifer, N., 2015. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics 31, i268–i275. https://doi.org/10.1093/bioinformatics/btv244

Stańczyk, U., Jain, L.C. (Eds.), 2015. Feature selection for data and pattern recognition, Studies in computational intelligence. Springer, Berlin.

Stetson, L.C., Pearl, T., Chen, Y., Barnholtz-Sloan, J.S., 2014. Computational identification of multi-omic correlates of anticancer therapeutic response. BMC Genomics 15, S2. https://doi.org/10.1186/1471-2164-15-S7-S2

Strimbu, K., Tavel, J.A., 2010. What are Biomarkers? Curr. Opin. HIV AIDS 5, 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Subramaniam, S., 2007. LMSD: LIPID MAPS structure database. Nucleic Acids Res. 35, D527–D532. https://doi.org/10.1093/nar/gkl838

Tan, K., Huang, W., Hu, J., Dong, S., 2020. A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. BMC Med. Inform. Decis. Mak. 20, 129. https://doi.org/10.1186/s12911-020-1114-3

Tan, X., Yu, Y., Duan, K., Zhang, J., Sun*, P.S. and H., 2020. Current Advances and Limitations of Deep Learning in Anticancer Drug Sensitivity Prediction [WWW Document]. Curr. Top. Med. Chem. URL https://www.eurekaselect.com/183610/article (accessed 11.2.20).

Tang, B., Pan, Z., Yin, K., Khateeb, A., 2019. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00214

Taskesen, E., Babaei, S., Reinders, M.M.J., de Ridder, J., 2015. Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. BMC Bioinformatics 16 Suppl 4, S5. https://doi.org/10.1186/1471-2105-16-S4-S5

Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., Gojobori, T., 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. 30, 27–30. https://doi.org/10.1093/nar/30.1.27

Tepeli, Y.I., Ünal, A.B., Akdemir, F.M., Tastan, O., 2019. PAMOGK: A Pathway Graph Kernel based Multi-Omics Clustering Approach for Discovering Cancer Patient Subgroups. bioRxiv 834168. https://doi.org/10.1101/834168

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515. https://doi.org/10.1093/nar/gky1049

Thomas, T., Stefanoni, D., Dzieciatkowska, M., Issaian, A., Nemkov, T., Hill, R.C., Francis, R.O., Hudson, K.E., Buehler, P.W., Zimring, J.C., Hod, E.A., Hansen, K.C., Spitalnik, S.L., D'Alessandro, A., 2020. Evidence for structural protein damage and membrane lipid remodeling in red blood cells from COVID-19 patients. medRxiv 2020.06.29.20142703. https://doi.org/10.1101/2020.06.29.20142703

Thudumu, S., Branch, P., Jin, J., Singh, J. (Jack), 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. J. Big Data 7, 42. https://doi.org/10.1186/s40537-020-00320-x

Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R.D., York, W.S., Karlsson, N.G., Lisacek, F., Packer, N.H., Campbell, M.P., Aoki, N.P., Fujita, A., Matsubara, M., Shinmachi, D., Tsuchiya, S., Yamada, I., Pierce, M., Ranzinger, R., Narimatsu, H., Aoki-Kinoshita, K.F., 2017. GlyTouCan: an accessible glycan structure repository. Glycobiology 27, 915–919. https://doi.org/10.1093/glycob/cwx066

Timp, W., Timp, G., 2020. Beyond mass spectrometry, the next step in proteomics. Sci. Adv. 6, eaax8978. https://doi.org/10.1126/sciadv.aax8978

Tini, G., Marchetti, L., Priami, C., Scott-Boyer, M.-P., 2019. Multi-omics integration—a comparison of unsupervised clustering methodologies. Brief. Bioinform. 20, 1269–1279. https://doi.org/10.1093/bib/bbx167

Tipping, M.E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. J Mach Learn Res 1, 211–244. https://doi.org/10.1162/15324430152748236

Tong, L., Wu, H., Wang, M.D., 2020. Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer. Methods. https://doi.org/10.1016/j.ymeth.2020.07.008

Tsuda, K., Shin, H., Schölkopf, B., 2005. Fast protein classification with multiple networks. Bioinformatics 21, ii59–ii65. https://doi.org/10.1093/bioinformatics/bti1110

Uddin, S., Khan, A., Hossain, M.E., Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Inform. Decis. Mak. 19, 281. https://doi.org/10.1186/s12911-019-1004-8

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., Ponten, F., 2010. Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. 28, 1248–1250. https://doi.org/10.1038/nbt1210-1248

Van Deun, K., Smilde, A.K., van der Werf, M.J., Kiers, H.A., Van Mechelen, I., 2009. A structured overview of simultaneous component based data integration. BMC Bioinformatics 10, 246. https://doi.org/10.1186/1471-2105-10-246

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.

Vasta, G.R., Ahmed, H., 2008. Animal Lectins: A Functional View. CRC Press.

Vickers, A.J., Elkin, E.B., 2006. Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 26, 565–574. https://doi.org/10.1177/0272989X06295361

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. J. Mach. Learn. Res. 11, 3371–3408.

Vineetha, S., Chandra Shekara Bhat, C., Idicula, S.M., 2013. MicroRNA-mRNA interaction network using TSK-type recurrent neural fuzzy network. Gene 515, 385–390. https://doi.org/10.1016/j.gene.2012.12.063

Vivian, J., Eizenga, J.M., Beale, H.C., Vaske, O.M., Paten, B., 2020. Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples. JCO Clin. Cancer Inform. 4. https://doi.org/10.1200/CCI.19.00095

Vogel, F., Motulsky, A.G., 1997. Vogel and Motulsky's Human Genetics: Problems and Approaches. Springer Science & Business Media.

Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A., 2014. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337. https://doi.org/10.1038/nmeth.2810

Wang, D., Gribskov, M., 2005. Examining the architecture of cellular computing through a comparative study with a computer. J. R. Soc. Interface 2, 187–195. https://doi.org/10.1098/rsif.2005.0038

Wang, L., 2010. Pharmacogenomics: a systems approach. Wiley Interdiscip. Rev. Syst. Biol. Med. 2, 3–22. https://doi.org/10.1002/wsbm.42

Wang, M., Wang, C., Han, R.H., Han, X., 2016. Novel Advances in Shotgun Lipidomics for Biology and Medicine. Prog. Lipid Res. 61, 83–108. https://doi.org/10.1016/j.plipres.2015.12.002

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K., 2020. MORONET: Multi-omics Integration via Graph Convolutional Networks for Biomedical Data Classification. bioRxiv 2020.07.02.184705. https://doi.org/10.1101/2020.07.02.184705

Waring, J., Lindvall, C., Umeton, R., 2020. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif. Intell. Med. 104, 101822. https://doi.org/10.1016/j.artmed.2020.101822

Watanabe, K., Yasugi, E., Oshima, M., 2000. How to Search the Glycolipid data in "LIPIDBANK for Web" the Newly Developed Lipid Database in Japan. Trends Glycosci. Glycotechnol. 12, 175–184. https://doi.org/10.4052/tigg.12.175

Watt, J., Borhani, R., Katsaggelos, A.K., 2020. Machine learning refined: foundations, algorithms, and applications, Second Edition. ed. Cambridge University Press, New York.

Weisz Hubshman, M., Broekman, S., van Wijk, E., Cremers, F., Abu-Diab, A., Khateb, S., Tzur, S., Lagovsky, I., Smirin-Yosef, P., Sharon, D., Haer-Wigman, L., Banin, E., Basel-Vanagaite, L., de Vrieze, E., 2018. Whole-exome sequencing reveals POC5 as a novel gene associated with autosomal recessive retinitis pigmentosa. Hum. Mol. Genet. 27, 614–624. https://doi.org/10.1093/hmg/ddx428

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12. https://doi.org/10.1371/journal.pone.0174944

Wilkins, M.R., Appel, R.D., 2007. Ten Years of the Proteome, in: Wilkins, M.R., Appel, R.D., Williams, K.L., Hochstrasser, D.F. (Eds.), Proteome Research: Concepts, Technology and Application. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–13. https://doi.org/10.1007/978-3-540-72910-5_1

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C., Scalbert, A., 2018. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 46, D608–D617. https://doi.org/10.1093/nar/gkx1089

Wong, A.J., Kanwar, A., Mohamed, A.S., Fuller, C.D., 2016. Radiomics in head and neck cancer: from exploration to application. Transl. Cancer Res. 5, 371–382. https://doi.org/10.21037/8305

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., Ma, S., 2019. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput 8, 4. https://doi.org/10.3390/ht8010004

Wu, C.-C., Asgharzadeh, S., Triche, T.J., D'Argenio, D.Z., 2010. Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. Bioinformatics 26, 807–813. https://doi.org/10.1093/bioinformatics/btq044

Wu, D., Wang, D., Zhang, M.Q., Gu, J., 2015. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics 16, 1022. https://doi.org/10.1186/s12864-015-2223-8

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., Xu, H., 2020. Deep learning in clinical natural language processing: a methodical review. J. Am. Med. Inform. Assoc. 27, 457–470. https://doi.org/10.1093/jamia/ocz200

Wu, X., Hasan, M.A., Chen, J.Y., 2014. Pathway and Network Analysis in Proteomics. J. Theor. Biol. 0, 44–52. https://doi.org/10.1016/j.jtbi.2014.05.031

Xu, D., Tian, Y., 2015. A Comprehensive Survey of Clustering Algorithms. Ann. Data Sci. 2, 165–193. https://doi.org/10.1007/s40745-015-0040-1

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, Hussain, Dawood, Hassan, 2019. A hierarchical integration deep flexible neural forest framework for cancer subtype

classification by integrating multi-omics data. BMC Bioinformatics 20, 527. https://doi.org/10.1186/s12859-019-3116-7

Xu, X., Liang, T., Zhu, J., Zheng, D., Sun, T., 2019. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. Neurocomputing, Chinese Conference on Computer Vision 2017 328, 5–15. https://doi.org/10.1016/j.neucom.2018.02.100

Yan, K.K., Zhao, H., Pang, H., 2017. A comparison of graph- and kernel-based –omics data integration algorithms for classifying complex traits. BMC Bioinformatics 18, 539. https://doi.org/10.1186/s12859-017-1982-4

Yan, Z., Xiong, Y., Xu, W., Li, M., Cheng, Y., Chen, F., Ding, S., Xu, H., Zheng, G., 2012. Identification of recurrence-related genes by integrating microRNA and gene expression profiling of gastric cancer. Int. J. Oncol. 41, 2166–2174. https://doi.org/10.3892/ijo.2012.1637

Yang, K., Han, X., 2016. Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. Trends Biochem. Sci. 41, 954–969. https://doi.org/10.1016/j.tibs.2016.08.010

Young, F.W., Hamer, R.M., 1987. Multidimensional scaling: history, theory, and applications. L. Erlbaum Associates, Hillsdale, N.J.

Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage Clin. 2, 735–745. https://doi.org/10.1016/j.nicl.2013.05.004

Yu, X.-T., Zeng, T., 2018. Integrative Analysis of Omics Big Data. Methods Mol. Biol. Clifton NJ 1754, 109–135. https://doi.org/10.1007/978-1-4939-7717-8_7

Yuan, Y., Savage, R.S., Markowetz, F., 2011. Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. PLOS Comput. Biol. 7, e1002227. https://doi.org/10.1371/journal.pcbi.1002227

Yue, Z., Meng, D., He, J., Zhang, G., 2017. Semi-supervised learning through adaptive Laplacian graph trimming. Image Vis. Comput., Regularization Techniques for High-Dimensional Data Analysis 60, 38–47. https://doi.org/10.1016/j.imavis.2016.11.013

Zampieri, M., Sekar, K., Zamboni, N., Sauer, U., 2017. Frontiers of high-throughput metabolomics. Curr. Opin. Chem. Biol., Omics 36, 15–23. https://doi.org/10.1016/j.cbpa.2016.12.006

Zhang, L., Dong, X., Lee, M., Maslov, A.Y., Wang, T., Vijg, J., 2019. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc. Natl. Acad. Sci. 116, 9014–9019. https://doi.org/10.1073/pnas.1902510116

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., Shi, T., 2018. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Front. Genet. 9. https://doi.org/10.3389/fgene.2018.00477

Zhang, Q., Burdette, J.E., Wang, J.-P., 2014. Integrative network analysis of TCGA data for ovarian cancer. BMC Syst. Biol. 8, 1338. https://doi.org/10.1186/s12918-014-0136-9

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P.W., Zhou, X.J., 2012. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 40, 9379–9391. https://doi.org/10.1093/nar/gks725

Zhang, Y., Wang, B., Jin, W., Wen, Y., Nan, L., Yang, M., Liu, R., Zhu, Y., Wang, C., Huang, L., Song, X., Wang, Z., 2019. Sensitive and robust MALDI-TOF-MS glycomics analysis enabled by Girard's reagent T on-target derivatization (GTOD) of

reducing glycans. Anal. Chim. Acta 1048, 105–114. https://doi.org/10.1016/j.aca.2018.10.015
Zhao, J., Xie, X., Xu, X., Sun, S., 2017. Multi-view learning overview: Recent progress and new challenges. Inf. Fusion 38, 43–54. https://doi.org/10.1016/j.inffus.2017.02.007
Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., Ma, S., 2015. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief. Bioinform. 16, 291–303. https://doi.org/10.1093/bib/bbu003
Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., Liu, X., 2014. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE 9, e78644. https://doi.org/10.1371/journal.pone.0078644
Zhou, B., Xiao, J.F., Tuli, L., Ressom, H.W., 2012. LC-MS-based metabolomics. Mol. Biosyst. 8, 470–481. https://doi.org/10.1039/c1mb05350g
Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S.G., Alvarez-Cohen, L., 2015. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. mBio 6. https://doi.org/10.1128/mBio.02288-14
Zhou, J.T., Pan, S.J., Tsang, I.W., 2019. A deep learning framework for Hybrid Heterogeneous Transfer Learning. Artif. Intell. 275, 310–328. https://doi.org/10.1016/j.artint.2019.06.001
Zhou, Y., Hou, Y., Shen, J., Mehra, R., Kallianpur, A., Culver, D.A., Gack, M.U., Farha, S., Zein, J., Comhair, S., Fiocchi, C., Stappenbeck, T., Chan, T., Eng, C., Jung, J.U., Jehi, L., Erzurum, S., Cheng, F., 2020. A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19. PLOS Biol. 18, e3000970. https://doi.org/10.1371/journal.pbio.3000970
Zhu, J., Sova, P., Xu, Q., Dombek, K.M., Xu, E.Y., Vu, H., Tu, Z., Brem, R.B., Bumgarner, R.E., Schadt, E.E., 2012. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. PLOS Biol. 10, e1001301. https://doi.org/10.1371/journal.pbio.1001301
Zhu, W., Xie, L., Han, J., Guo, X., 2020. The Application of Deep Learning in Cancer Prognosis Prediction. Cancers 12. https://doi.org/10.3390/cancers12030603
Zierer, J., Pallister, T., Tsai, P.-C., Krumsiek, J., Bell, J.T., Lauc, G., Spector, T.D., Menni, C., Kastenmüller, G., 2016. Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model. Sci. Rep. 6, 37646. https://doi.org/10.1038/srep37646
Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. J. Am. Stat. Assoc. 101, 1418–1429.
Zou, Q., Chen, L., Huang, T., Zhang, Z., Xu, Y., 2017. Machine learning and graph analytics in computational biomedicine. Artif. Intell. Med. https://doi.org/10.1016/j.artmed.2017.09.003

FIGURE LEGENDS

Figure 1: Number of publications published per year on different search keywords. *For the year 2020, the annual count was extrapolated using the count of publications until October 2020.

Figure 2: Workflow pipelines for different types of integration methods for multi-omics analysis.

Figure 3: Recommendation flowchart for choosing a method for multi-omics integration. For abbreviations please refer to *List of Abbreviations*.

CRediT author statement

Parminder S. Reel: Conceptualization, Methodology, Validation, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

Smarti Reel: Conceptualization, Methodology, Investigation, Writing - Review & Editing, Visualization

Ewan Pearson: Writing - Review & Editing

Emanuele Trucco: Writing - Review & Editing

Emily Jefferson: Conceptualization, Writing - Review & Editing, Supervision

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Title: Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review

Parminder S. Reel[1*], Smarti Reel[1*], Ewan Pearson[1], Emanuele Trucco[2], Emily Jefferson[1]

[1]Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, United Kingdom.
[2]VAMPIRE project, Computing, School of Science and Engineering, University of Dundee, Dundee, United Kingdom.

* These authors contributed equally to this study (shared first authorship)

***Corresponding author*:**

Prof. Emily Jefferson

Division of Population Health and Genomics,
School of Medicine,
University of Dundee, Dundee, UK
Email: e.r.jefferson@dundee.ac.uk

HIGHLIGHTS

- Machine learning methods are novel techniques to integrate omics datasets
- Recently, publications based on 'multi-omics integration' have gained popularity
- Integration of omics data using concatenation, model- or transformation-based methods
- Multi-omics studies offer a more comprehensive view of complex diseases
- Recommendation flowchart included for interdisciplinary professionals