Navigating the Multiverse: A Hitchhiker's Guide to Selecting Harmonisation Methods for Multimodal Biomedical Data

Murali Aadhitya Magateshvaren Saras[1,2,3], Mithun K. Mitra[2], Sonika Tyagi[3,4*]

[1] IITB-Monash Research Academy, Mumbai, Maharashtra, India, 400076 .
[2] Indian Institute of Technology Bombay, Mumbai, Maharashtra, India, 400076 .
[3] Central Clinical School, Monash University, Melbourne, Victoria, Australia, 3181 .
[4] School of Computational Technologies, RMIT University, Melbourne, Victoria, Australia, 3000 .

*Corresponding author(s). E-mail(s): sonika.tyagi@monash.edu;

Abstract

Introduction: The application of machine learning (ML) techniques in classification and prediction tasks has greatly advanced our comprehension of biological systems. There is a notable shift in the trend towards integration methods that specifically target the simultaneous analysis of multiple modes or types of data, showcasing superior results compared to individual analyses. Despite the availability of diverse ML architectures for researchers interested in embracing a multimodal approach, the current literature lacks a comprehensive taxonomy that includes the pros and cons of these methods to guide the entire process. Closing this gap is imperative, necessitating the creation of a robust framework. This framework should not only categorise the diverse ML architectures suitable for multimodal analysis but also offer insights into their respective advantages and limitations. Additionally, such a framework can act as a guide for selecting an appropriate workflow for multimodal analysis. This comprehensive taxonomy would furnish a clear guidance and aid in informed decision-making within the progressively intricate realm of biomedical and clinical data analysis, and is imperative for advancing personalised medicine.
Objective: The aims of the work are to comprehensively study and describe the

harmonisation processes that are performed and reported in the literature and present a working guide that would enable planning and selecting an appropriate integrative model.
Methods: A systematic review of publications that report the multimodal harmonisation of biomedical and clinical data has been performed.
Results: We present harmonisation as a dual process of representation and integration, each with multiple methods and categories. The taxonomy of the various representation and integration methods are classified into six broad categories and detailed with the advantages, disadvantages and examples. A guide flowchart that describes the step-by-step processes that are needed to adopt a multimodal approach is also presented along with examples and references.
Conclusions: This review provides a thorough taxonomy of methods for harmonising multimodal data and introduces a foundational 10-step guide for newcomers to implement a multimodal workflow.

Keywords: multimodal integration, feature representation, data integration, deep learning, digital health

1 Introduction

The growth of biological and healthcare data, in terms of volume, velocity and variety, has been exponential and driven by technological advances in electronics, communication and infrastructure (Laney (2001); Dash et al. (2019)). Concurrently, there has been an increase in data analysis tools to understand and analyse the data. Progress in computational techniques, artificial intelligence (AI) and machine learning (ML) methods have been identified to contribute towards the analysis and interpretation better than traditional analytical methods (Acosta et al. (2022); Li and Ngom (2015)).

Data generated in the context of biological systems can manifest in various forms such as quantitative, qualitative or narrative; each of these has its subtypes, which are collectively referred to as a 'modality'. These diverse modalities can capture several aspects of a biological system, such as nucleic acid and protein sequences (Neidle (2008)), gene expression (Raghavachari and Garcia-Reyero (2018)) and the biomolecular structure and its activity (Vergoten and Theophanides (2012)). Other modalities include the epigenetic state and methylation information (Paro et al. (2021)) of the genome, metabolites, and anatomic and phenotypic data.

Each data type has driven research towards elucidating the corresponding functional aspects to understand the system. Numerous studies using a single data modality have presented valuable additions to the literature in disease mapping, pathway and network elucidation (Aburajab et al. (2023); Mansuri et al. (2023); Pang et al. (2023)). However, a vast portion of the biological complexity still requires an explanation, which is an ongoing challenge for the research community.

Different modalities capture different aspects of the system. Thus, integrating them provides a comprehensive multi-view understanding of both biological and clinical conditions (Li and Ngom (2015); Nie et al. (2007)). Combining multiple types of omics data or a 'multiomics' approach to study biological systems has gained momentum lately due to their demonstrated superiority over single-omics approaches (Chen et al. (2021); Chen and Tyagi (2020); Baltrušaitis et al. (2018); Acosta et al. (2022); Summaira et al. (2021)). Furthermore, healthcare data is integrated with omics datasets to reveal their interconnections, providing a comprehensive 360-degree view of an individual's condition. (Schiano et al. (2020); Dargazanli et al. (2020)). Such studies have reported results with significant validation and reliability over independent analysis. Thus, the integration of multiple modalities can reveal synergistic effects, where the combined information enhances the overall performance of the model beyond what individual modalities can achieve.

Existing literature primarily focuses on the model architecture and merits of the general ML methods used for analysis (Li and Ngom (2015); Summaira et al. (2021); Sapoval et al. (2022)). However, a gap still persists in delineating between integrated learning and co-learning (harmonisation). Harmonisation aims to elucidate the low-level relationship between features of different modalities (Baltrušaitis et al. (2018)). Often, articles incorporating 'integration' as part of their pipeline do not necessarily perform a harmonisation process, instead they focus on the correlation between individual data type analysis (Chen et al. (2021)). The effect of an analysis using multiple modalities is not adequately captured by methods that do not harmonise the features. A co-learning set-up is distinct from an individual analysis since it necessitates fusion

of features. Published articles relay information to mitigate data challenges but lack information on the multimodal process (Mirza et al. (2019); Kline et al. (2022); Acosta et al. (2022); Zitnik et al. (2019)). A definitive explanation of the methods involved in a multimodal harmonisation approach is missing. The absence of adequate information impedes interested researchers from fully grasping the process and implementing a workflow. In summary, this literature review addresses the aforementioned gap by offering insights into data modalities, challenges encountered in data, the processes involved in a multimodal setup, and a beginner's guide to multimodal analysis.

2 Methods

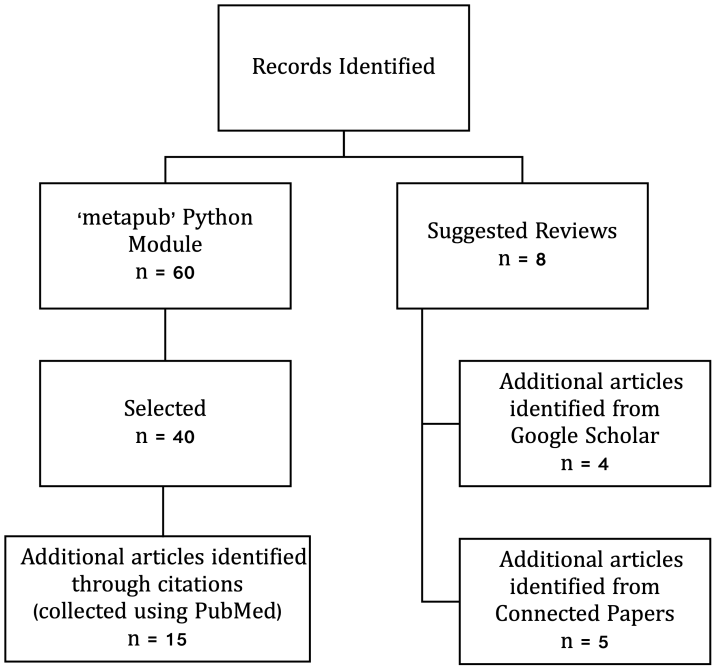

Fig. 1: Flowchart of literature screening. The blocks on the left indicate articles searched for representation and integration methods. The blocks on the right describe the targeted search for review articles on biomedical multimodal harmonisation.

This systematic review was performed based on the standards of the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA 2020) statement (Page et al. (2021)).

Based on existing literature reviews published over the past decade, a general outline was followed to select articles that mentioned multimodal learning techniques. An extensive search of various ML methods focused on biomedical data was initially gathered using the metapub (https://github.com/metapub/metapub) python module

based on the following keywords:
'Multimodal', 'Machine learning', 'Integrated learning', 'Multiomics', 'Genomics', 'Proteomics', 'Biomedical', 'Healthcare data', 'Biological network', 'Deep learning', 'Multitask learning', 'Data fusion', 'Representation', 'Interpretable model', 'Neural network'.

The title and the abstract of the search results reported by the keywords using the python script was used to select papers for complete reading. A few suggested articles from co-authors on were used to initiate the search for reviews on biomedical multimodal integration. Citations from suggested papers were manual searched using Google Scholar and PubMed. Connected Papers (https://www.connectedpapers.com) was used to identify related articles. We identified less than 10 articles that described on the topic. Few papers were selected for reading on general multimodal integration methods. Figure 1 describes the count of articles.

The inclusion criteria for this review primarily focused on studies that incorporated multiple different types of biological and medical data towards a singular analysis using machine learning algorithms. Reports that did not use biomedical datasets but employed a multimodal approach for data analysis were also included for review. The exclusion criteria was marked by the absence of a multimodal approach only. However, studies that focused on representation methods of different data types were included for full-text reading. The last date of article search and selection was 20 October 2023.

The methods for selected articles were reviewed in detail and information on the data type, machine learning framework, model advantages, research gaps were collected. They have been classified into groups and presented in tabular format (Tables 1, 4) and the results are discussed in following sections.

3 Results

In the following sections, we integrate our findings regarding data types, technical hurdles, and harmonisable methodologies gleaned from the literature we have examined.

3.1 Typical Study Designs in biological and clinical studies

All data generated are usually based on a study or an objective that has a focused rationale. The richness of information in collected or generated data depends on its type, determining the range of possible analyses (Ranganathan and Aggarwal (2019)).

A static study design acquires data as a 'snapshot', that is, collected at a point in time. Case series refer to static data collected from positive-group criteria within a population subgroup, while case-control studies include a negative dataset (controls) for comparison. On the other hand, cohort studies and randomised controlled trials sample data over a period of time, capturing the dynamic nature of a biological system, which allows for a realistic investigation. However, they are resource-intensive methods that must be maintained regularly, and constant follow-up with the subjects considered in the study is crucial.

Many efforts are being taken to enhance the data collection methods and accessibility across various domains, such as in cancer (The Cancer Genome Atlas,

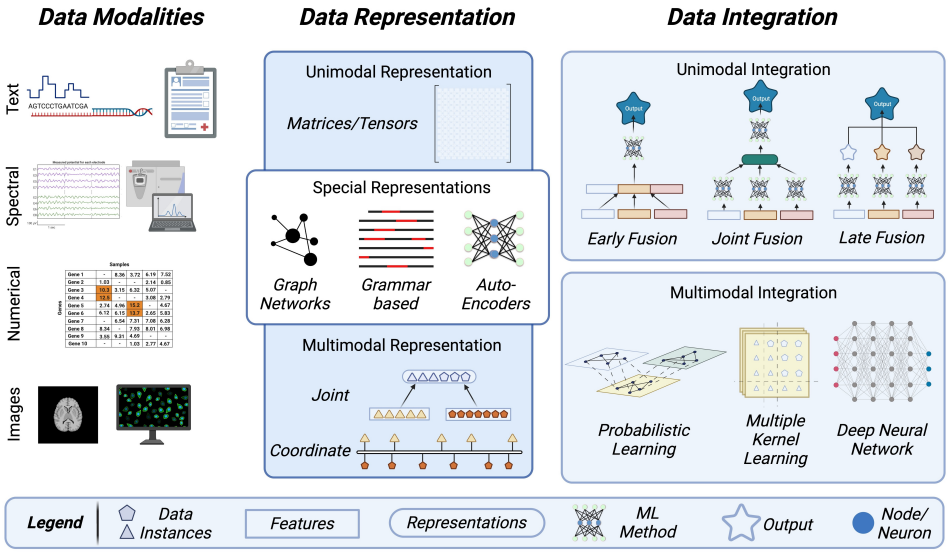

Fig. 2: An illustration briefly depicting the broad categories of data modalities and the representation and integration methods used in a multimodal harmonisation analysis. The representation methods are split into three groups based on the number and type of datasets. The integration methods are split based on the type of fusion performed. Made with BioRender.

https://www.cancer.gov/ccg/) and preterm birth (Garbh-Ini). Current literature predominantly reports on results based on single, static datasets. Correlation studies use multiple datasets to support conclusions through overlapping results (Clarke et al. (2017)). Only a few methods take a complementary approach between modalities (Welch et al. (2017); Chen et al. (2021)).

3.2 Common data modalities studied in the literature

Data can be collected in forms such as text, numbers, and multimedia. Based on the sources, they can be classified as 'biological' data and 'health' data (Dash et al. (2019)). We refer to 'biological data' as information from high-throughput experiments such as sequencing, expression profiling, microscopic imaging and the vast literature corpus for functional annotation. This also includes metadata related to samples, experimental design, assay protocols and technologies. 'Health' sources refer to data primarily collected through healthcare providers in digital forms. This data contains an individual's valuable health and medical history and is stored as time-stamped electronic medical records (EMR). EMR data displays significant structural diversity since it can be structured data, including vital signs and pathology measurements organised in tabular formats, or presented as unstructured data, consisting of clinical notes, images, and documents. Table 1 describes the different modalities that stem from clinical and biomedical sources.

3.2.1 Text Modality

Text as a modality comprises various types, encompassing narrative and sequence forms of data. Sequence data stemming from biological macromolecules such as DNA, RNA, and proteins describe and define the relationship between the genotype and the phenotype of an organism. Differences in sequences among groups differing in demography or phenotype are represented as single nucleotide polymorphism (SNP) or small insertions and deletions of DNA bases and are generated by genome-wide association studies (GWAS) (Uffelmann et al. (2021)). Information on motifs, interaction networks and annotations about biomolecules, drugs and diseases belong to this class. Healthcare data, such as EMR, contain unstructured clinical notes and prescriptions manually entered by medical practitioners, which are included in the text category of datasets (Lima et al. (2019)).

3.2.2 Spectral and Signal Modality

Spectral data, typically acquired through mass spectrometry experiments to study protein molecules and metabolites, provides detailed insights into the structural composition, constitution and organisation of the molecules under investigation (Mansuri et al. (2023); Mou et al. (2022)). ML analysis of spectral data involves features representing three-dimensional conformations and spatial relationships of molecules, enabling classification based on functional groups and elements (Mou et al. (2022); Sachdev and Gupta (2019)). Proteomics, examining proteins through expression, functional relationships, and structural information, includes investigations into protein folding and structural orientations using methods such as NMR and X-ray crystallography (Malet-Martino and Holzgrabe (2011)). Structural metabolomics stores the structural data collected from metabolites.

Healthcare data in spectral form includes time-dependent Electroencephalogram (EEG) and electrocardiogram (ECG) analysed with signal processing methods (Subha et al. (2010); Abarbanel et al. (2009)). Audio data, such as voice notes, undergoes analysis using appropriate methods after feature extraction (Camastra and Vinciarelli (2015)).

3.2.3 Numerical Modality

A numerical form of biological data can be from any quantifiable assay, broadly called 'omics' data. Transcriptomics represents digital counts of identified expressed transcript molecules, available as a two-dimensional (2D) matrix of genes/transcripts and samples. Similarly, proteomic, lipidomic and metabolomic counts data portray the expression levels of proteins, lipids or metabolite molecules as a 2D data matrix. This modal information allows understanding of individual differences regarding genetic expression and linking related biological pathways. EMR readings document vital and pathological parameters such as heart rate, weight, height, age, and blood pressure as numerical time-series data. Frequent time-stamped EMRs enable longitudinal analysis, capturing changes in the recorded values over time (Haghverdi et al. (2016)).

Type	Data Source	Features of Interest	Reference
Text	DNA, RNA and Protein Sequence	Sequence order and motifs	(Tyagi et al. (2008); Andrikos et al. (2022)
	GWAS Data	Genetic variants	(Chang et al. (2018)
	Clinical Notes	Correlated medical terms and phrases	(Banerjee et al. (2019)
Spectral and signal	X-Ray crystallography, NMR, Mass Spectroscopy	Structural composition and identified functional groups, topology	(Mou et al. (2022)
	Audio signals	Speech to text patterns	(Summaira et al. (2021)
Numerical	Biomolecular profiles (Lipids, Metabolites, Nucleic acids, Proteins)	Expression levels of biomolecules	(Zitnik et al. (2019)
	EMR (Vitals, Lab measurements)	Health factors, trends and trajectories	(Banerjee et al. (2019)
	Interaction Networks (Diseases, drugs, genes, proteins)	Regulatory and Functional relationships	(Lee et al. (2020)
Images	EMR (CT, X-ray, Ultrasound)	Patterns and localisations	(Zhao et al. (2021)
	Cell Imaging	Patterns and localisations	(Schiano et al. (2020)

Table 1: The four distinct modalities biological and clinical data investigated in this study are listed and categorised based on their sources and the features targeted for modelling.

3.2.4 Image Modality

This modality encompasses visual information, including images and videos. Videos are also considered under this modality because each frame can be considered an image for processing but in a time-dependent manner (Camastra and Vinciarelli (2015)). Microscopy and cell imaging data are often analysed for morphology studies, protein localisation and DNA tagging (Fu and Rui (2017)). Manual image analysis methods include segmentation tasks to identify regions of interest and cell morphology assessment (Kan (2017)). Cell movement and tracking studies creating animated clips from multiple fluorescence-tagged cell images are a modality of this category. Additionally, X-rays, CT and MRI images from EMRs supporting non-invasive diagnosis fall in this category.

3.3 Common challenges associated with biomedical data

Datasets require extensive pre-processing due to incompleteness and imperfections before analysis (López de Maturana et al. (2019)). Key challenges include high dimensionality, heterogeneity, missing data, class imbalances, bias and accessibility.

Complex high-dimensional data, characterised by large features and file sizes, require extensive computational resources for understanding the variables (Stephens et al. (2015)). Addressing the 'p>>n' problem, considering the ratio of available samples (n) to features (p), is crucial to prevent specific features from being overlooked in small sample groups (Baltrušaitis et al. (2018); Stephens et al. (2015)).

Bias refers to a variety of imbalances found within a dataset and can lead to an unfair interpretation of results. Bias can manifest in various forms, such as representation bias or class imbalance, measurement bias (due to incorrect or unrelated values), aggregation bias (when models are applied to new datasets with a mutually exclusive relationship to training samples) and evaluation bias (when generic models are used as benchmarks for targeted datasets) (Suresh and Guttag (2021)). Comparative analytical methods utilise representative datasets, subsets of the population with samples from distinct groups like case and control. It is crucial for groups to have samples in a comparable and an equivalent number to understand the true differences.

Irregular clinical data collection processes lead to inconsistent data entries and contribute to missing data. In datasets, all samples may not provide data for all possible features, and the resulting matrix could be sparse in a few cases. Importantly, a missing measurement may carry meaning and should be considered subjectively.

Heterogeneity refers to the variety that exists within and across modalities. Within a modality, the data collected across variables can vary in terms of scale, distribution and recorded value, such as discrete, continuous, categories and intervals, due to non-standardised procedures. For example, clinical and genomic data can not be directly compared and analysed, requiring methods to address heterogeneity.

These challenges obstruct the potential in any analysis, but the problem exacerbates when multiple datasets are involved in a multimodal set-up (Zhang et al. (2019)). Multimodal methods are affected by coherence between dataset sets (due to heterogeneity and missing data), accessibility and computational resources (due to high-dimensional datasets and bias).

Data preprocessing steps prepare a path to check, sort, and select data points so that informed decisions can be taken to handle samples with anomalies and poor quality. The lack of data standardisation between multiple collection sites poses a challenge for seamless data harmonisation (Ramakrishnaiah et al. (2023)) and requires specific preprocessing for different sources. While imputation methods partially handle missing data, they are not universal solutions, as approximations may not accurately reflect the system (Schafer (1997)). Hence, more data-driven approaches may be adopted in different scenarios (Ramakrishnaiah et al. (2023)). Large datasets can be converted to latent values to reduce computational load. Representation methods (Section 3.5) effectively resolve these issues.

Biomedical and health data containing personal and sensitive information are restricted for global access, which limits the extent and scope of analysis. Implementing ethical and legal data practices, both nationally and internationally, is crucial for

easing the data sharing process (Tyagi (2023)). These practices establish a structured and transparent approach to handling data, creating an environment conducive to sharing valuable information.

3.4 Multimodal Data Integration: a systematic investigation of data formats and methodologies

Multimodal analytical methods aim to combine information from multiple modalities towards one or many of the following goals: 1) Explain a biological phenomenon or phenotype through overlapping results. 2) Account for and impute missing data in one modality through another linked dataset/modality. 3) Condense the high-dimensional, sparse and noisy data to a low-dimensional latent representation.

The fundamental difference between an unimodal and a multimodal analysis is the number of different modalities used. The complete dataset can be directly fed into an ML architecture for an unimodal analysis, but a multimodal approach requires the fusion of features from multiple datasets. The process of merging and modelling of features can be classified under 'representation' and 'integration'. The choice of method varies depending on the task to be achieved and the dataset combinations. Section 3.5 describes the different ways of feature representation, and section 3.6 briefs about the fusion methods currently used with examples.

3.5 Data Representation

As discussed earlier, biomedical and health data is generated in many forms (Table 1). Data representation methods are crucial as they transform diverse data types into machine-processable formats such as vectors, matrices, or tensors. Vectors are one-dimensional representations of numerical values, while matrices and tensors hold data in 2-dimensional and multidimensional scales. These methods keep track of relationships between elements of each modality via predefined rules, facilitate feature extraction by using relevant information, and help in mapping data from one modality to another. Importantly, the representation methods can be modified to suit the study conducted (Sapoval et al. (2022)). In this context, we have classified three groups of data representation approaches.

3.5.1 Unimodal Data Representation

Unimodal data representation involves using a single mode or source of information to represent features. Each modality qualifies as an unimodal representation when independently transformed into a numerical format through an encoding or embedding approach.

Encoding involves the conversion of original data into a numerical format, whereas embedding refers to portraying the original data in a vector space that incorporates semantic information. The information from biological sequence and text can be encoded by converting them to a numerical representation based on composition (tallying frequency of words/monomers), K-mers (segmenting a biological sequence as a window of 'k' letters) and distribution (percentage of occurrence of each monomer within user-defined ranges) of the sequence (Yang et al. (2020)). K-mers serve as the

counterparts to n-grams or tokens in NLP methods, and ongoing efforts are focused on developing more advanced, data-driven approaches to derive them from sequential data (Chen et al. (2023)). Many tools have been created to embed text data into a numerical representation, such as word2vec (Church (2017)) or doc2vec (Lau and Baldwin (2016)), which preserves the order of information and local neighbouring relationships.

Numerical data, obtained as-is or in other formats is commonly represented as matrices for analysis purposes. Time-series information is represented as tensors, where the data is nested within matrices, extending in dimensions to include the temporal relationship (Zhu et al. (2021)). Similarly, image data is converted to a matrix representation by splitting a digital pixel into a numerical value between 0-255 for constituent colours. Further, spectral information from biological mass spectrometry studies generates coordinate data and is represented as a matrix.

Unimodal representation is the fundamental way to proceed with any ML analysis. The complete set of features obtained through representation methods can vary in size and dimension depending on the dataset. To alleviate the computational load and resources during modelling, feature selection and feature reduction methods reduce the representation into a smaller latent space portraying the complete dataset, which is used for analysis. There are multiple feature reduction methods, such as the Principal Component Analysis (PCA), Joint Non-negative Matrix Factorisation (Joint NMF) and Autoencoders. Wrapper methods (forward, backwards, and stepwise selection), Filter methods (ANOVA, Pearson correlation, variance thresholding), and embedded methods (Lasso, Ridge, Decision Tree) are all part of feature selection techniques (Mirza et al. (2019)).

3.5.2 Multimodal Representation

Multimodal data representation involves using multiple modes or sources of information to represent data. A multimodal representation fuses multiple unimodal representations together onto a shared feature space (joint) or co-represents the features from the different datasets (coordinate).

Each modality is condensed in a joint multimodal representation, and the defining features selected are concatenated to form a single collective representation. The ratio of features from each modality contributing to the concatenated representation is maintained uniformly. This prevents modalities with fewer features from being overwhelmed by modalities with large dimensions. Zhao et al. describe the application of joint representation in two publications using image data and clinical information (Zhao et al. (2020, 2021)). They merge representations of image data (CT scans) and clinical information in different ratios and predict lymph node (LN) metastasis (Zhao et al. (2020)). In a subsequent publication focusing on the same diagnosis, they introduce the 3M-CN architecture that utilises a 'refine layer' to predict LN metastasis (Zhao et al. (2021)). The refine layer is a concatenation of key features identified from clinical information and processed 3D images.

Coordinated representations reduce and present the features within each modality individually but link them towards the same meaning over a common coordinate

space. Trajectory inference or pseudotemporal ordering is a method to classify the different stages of the same cell type along an axis representing evolution (Saelens et al. (2019)). Pseudotime ordering is an excellent example of coordinate representations, where data from single-cell experiments are projected onto an evolutionary axis (Saelens et al. (2019)). The relationships established with identified patterns and domain knowledge help associate the features. MATCHER is a tool that has depicted imputation and correlation between modalities using a coordinate representation (Welch et al. (2017)). The manifold alignment method used in this tool achieved this task by representing data in low dimensions called a manifold and aligning them in a common space (alignment) (Wang and Mahadevan (2009)).

Multimodal representation methods empower ML architectures to investigate the interplay between features across diverse modalities. The entities in a biological system interact with each other in varied ways. Hence, the ratio of representations in the shared space as a parameter also affects the results of a multimodal analysis (Zhao et al. (2020)). Coordinate representations become more difficult than joint representations when there is no common ground to link the features.

3.5.3 Special representations

Special approaches represent data non-conventionally through a generative or a rule-based approach. These are not mutually exclusive to the previous two categories but process one or more source modalities differently to generate a representation. Generative representations learn the underlying patterns and structure of the data and are capable of generating new instances of data that are similar to the examples they were trained on. On the other hand, rule-based representations leverage formal rules and semantics to describe the features within a dataset.

Auto-Encoders (AE)

AE methods compress the entire dataset into a compact set of dimensions through an 'encoding' process, eliminating any non-representative features. A 'decoding' process then reconstructs the original data using the condensed representation, validating the reduced feature space. The decoding layers are generative of the relationships between all the variables within the data, and hence, this method is classified under a generative representation approach.

Detlefsen et al. extensively explores AE-based representations, emphasising the superior results achieved through the non-linear representation method in various tasks (Detlefsen et al. (2022)). Zhang et al. introduce OmiEmbed as a multitask framework, utilising the low-dimensional latent space generated by AEs for downstream tasks like cancer classification and survival prediction (Zhang et al. (2021)). AE representations also find applications in gene identification and cancer detection using expression data (Danaee et al. (2017)) and predicting carcinoma primary sites through DNA methylation data (Leitheiser et al. (2022)).

The encoding process in AEs can incorporate any type of model, such as a fully-connected neural network (FCNN) or convolutional neural network (CNN), to generate the latent space (Zhang et al. (2021)). Multiple modalities can also be combined at the input to generate a joint latent space (Huang et al. (2020)). This allows to generate different variants of the latent space and increases the choices available to work with.

The validation by the decoding process makes the latent space devoid of errors or data misrepresentation. However, the interpretability of AEs is generally low, and reducing the dimensions of the latent layer further diminishes the model understanding.

Graph-based

Graph representations portray relationships between different biological entities as a network by considering all features as 'nodes', and the relationship is depicted using 'edges'. The edge values denote characteristics like similarity, interaction, and affinity between features based on the data. The representation of features as graphs is not limited to local, adjacent points but links them globally with edges.

Intramodal networks describe relationships between identical molecule types (for example protein-protein), while intermodal networks depict links between distinct types (Lee et al. (2020)). Clinical information about diseases and drugs can also be represented as graphs with links depicting common pathologies and targets. Omics modalities such as genome, lipidome, metabolome, proteome and transcriptome can be fused with environment and EMR modalities and represented as a Heterogeneous Multi-layer Network (HMLN) (Himmelstein and Baranzini (2015)).

Specific ML architectures are devised to best use a graph network representation. Graph Convolutional Networks (GCN) are used to learn local graph structures and scale up based on the number of interactions to represent complex relationships (Kipf and Welling (2016)). Graph Attention Networks (GAT) incorporate attention mechanisms to overcome the structural overfitting for higher order GCNs (Veličković et al. (2017)). Ghorbani et al. present MGCN architecture, which implements graph representations to consolidate multilevel data (Ghorbani et al. (2019)).

Graphs can be generated using data from experimental protocols (e.g. omics), theory and literature (e.g. disease networks) to represent qualitative and quantitative information. Appropriate architectures (Section 3.6.2) assist to map and predict links between nodes using multilevel graph network data. The methods are highly sensitive to missing and unseen information but excel at discovering links within datasets. Sparse data matrices are easily translatable into graphs, as they efficiently condense large dimensional data to relevant nodes.

Grammar-based

Grammar-based methods, or semantic methods, rely on a predetermined, ordered set of 'vocabulary' to generate a representation, usually for a text-based modality. High-level patterns observed in the modality are identified and the complete dataset is represented based on the discovered patterns through a feature generation procedure. Additionally, data can either be embedded based on the provided dataset or referenced from the complete knowledge bank. Dictionary-based embedding methods create embeddings for the complete corpus, and the available data is represented based on the closest relationship from the complete dictionary (Baltrušaitis et al. (2018)).

Tyagi et al. used grammar-based representations to model the syntactic and semantic rule of RNA folding and used context-free grammars (CFG) to generate sequences and parse their structures (Tyagi et al. (2008)). Andikos et al. created Knotify (Andrikos et al. (2022)), a tool to predict RNA pseudoknots using CFG. Onokpasa et al. asserts that CFG representations improve compression ratios of RNA sequences and structures (Onokpasa et al. (2023)). Grammar-based representations have been

used to embed the motif information from sequences with the domain knowledge to depict the functionally connected regulatory regions (Soylu and Sefer (2023)).

Although grammar-based representations are powerful in capturing structured information, they may face challenges in handling the inherent ambiguity and variability in real-world data. They require large amounts of data and computing power to process and generate the rule-based representation (Baltrušaitis et al. (2018)).

In a few cases, ML methods do not differentiate between the representation steps and model training. For instance, dense neural networks and deep-learning architectures do not explicitly have a joint representation stage. They are directly processed for learning the features of the data and training the model (Section 3.6.2).

Table 2: Representation methods detailed with their advantages and disadvantages.

Representation		Advantages	Disadvantages
Unimodal		Simple; Interpretable; Allows interdependencies	Cannot capture contextual information; Susceptible to noise or biases; High feature sizes
Multimodal	Joint	Combines features to common space; Controls modality size effects; Allows interdependencies; Interpretable; Reduces dimensions;	Requires tailored architecture; Relies on meaningful cross-modal relationships;
	Coordinate	Aligns features to common space; Controls modality size effects; Allows interdependencies; Interpretable; Reduces dimensions; Imputes information; Captures contextual Information	Requires common axis for representation; representation depends on quality and definition of common space
Special	AEs	Creates latent representations; Controls modality size effects; Allows interdependecies; Reduces dimensions; Low susceptibility to noise	Low interpretability; Computationally expensive;
	Graph	Represents qualitative and quantitative; Interpretable; scales with feature size; covers global information	Requires domain-specific knowledge for feature extraction; Complex algorithms needed for irregular structures and dynamic graphs; Susceptible to missing information
	Grammar	Applies for text modality; Captures patterns and semantic information; Reduces dimensions; Interpretable;	Affected by ambiguity and complex language constructs; Computationally expensive; Large datasets needed for processing; Susceptible to missing information

3.6 Fusion of Data Modalities: Integrating Multifaceted Information

Data fusion methods harmonise the different data modalities available towards tasks such as clustering, regression or classification, utilising the representations from methods discussed above (Section 3.5). Different modalities can be harmonised using two broad ML approaches: Unimodal learning and Multimodal learning.

3.6.1 Unimodal Learning

Unimodal learning algorithms elicit information from individual modalities to predict an output. Here, three categories of unimodal integration exist based on how the features from different modalities are fused: (i) Early, (ii) Late, and (iii) Joint integration.

Early Fusion

Early fusion methods describe ML architectures that concatenate feature representations from multiple modalities at the input stage for modelling. The data is minimally processed, primarily to resolve heterogeneity, and samples are removed if imputation is impossible for missing data. This method disregards prior selection bias and allows us to investigate all features across modalities. It is time-consuming and computationally expensive to process the complex combinations of all features from modalities (Dash et al. (2019)).

The benefits of using early integration methods are discussed and reported by Barnum et al. (Barnum et al. (2020)). They assert that using immediate fusion techniques to merge modalities before feeding them into a model works better by integrating the lowest statistical correlations between input features. Banerjee et al. describe the PERFORM algorithm, utilising EMR data represented as temporal vectors, to assess its prediction performance in diagnosing acute pulmonary embolism (PE) (Banerjee et al. (2019)) with ElasticNet architecture (Zou and Hastie (2005)).

In the case of early fusion, class imbalance and differing sample sizes across modalities can affect the contribution of individual datasets, potentially biasing the analysis. Moreover, as the number of harmonised modalities increases in early fusion methods, the interpretability of the model drastically reduces. The limited coherence among data modalities restricts their combined usage, creating challenges in achieving a unified representation. For example, input data as a combination of metabolomics and chromatin accessibility data may hinder a unified representation. Chen et al. describe data agnostic and data specific methods, with choices of modalities that can be used for coherent analysis (Chen and Tyagi (2020)).

Late Fusion

Late fusion methods analyse multiple modalities independently using a model that best fits its representations to a predicted output, and the outputs from each are aggregated towards a singular result or inference. Late integration is generally performed either by taking an aggregated average of the predicted probabilities (outputs) from each modality or passing all the predictions from each modality into a FCNN to process a final output.

Wang et al. presented MOGONET as a tool to classify cancer subtypes using three modalities: mRNA expression, methylation and miRNA expression (Wang et al. (2021)). A late fusion architecture was established using GCN to predict an initial class label and an FCNN to generate a final class label prediction. Luo et al. presented a modified version of MOGONET, called GRAMINet with GATs instead of GCNs (Luo et al. (2023)). Both examples pass the same combination of modalities through different architectures to learn a model for biomedical data classification. However, this is not universally applicable since the late integration methods provide flexibility in model selection for different modalities. Huang et al. investigated a multimodal approach to predict PE and reported the results on seven different architectures, one early, two joint and four late fusion architectures (Huang et al. (2020)). The early fusion methods had the highest sensitivity, while the late ElasticNet architecture outperformed in all other metrics, such as accuracy, AUROC, specificity and positive predictive value.

Ensemble learning may be considered a variant of the late fusion model, where the outputs of multiple ML models are combined towards a final decision (Zhou (2012)). The late fusion focuses on combining features or representations after individual processing, whereas ensemble learning leverages the diversity of multiple models to improve overall predictive performance.

Late fusion methods do not directly allow for the interaction of features from multiple modalities. This enables to train each modality with independent, unique models without any interference from other data types. As a result, concerns about different dataset sizes, heterogeneous measurements, and model compatibility vanish. The ability of late fusion methods to capture all information within each modality in an equivalent manner makes it the widely reported harmonisation method in the literature.

Joint Fusion

Joint fusion methods endeavour to extract a representation of all initial modalities and model them together to a predicted output. Direct concatenation methods, often used in early fusion, are not possible between modalities that have different quantities and may require a heavy preprocessing step. In late integration, the interdependency between features across modalities is ignored. Joint integration methods provide an advantage through the interaction of features from different modalities in the training phase, irrespective of the observed heterogeneity. The heterogeneity is mitigated since the feature representation and selection procedures reduce and unify the information numerically.

Joint integration methods have been explored using modalities such as CT scans, EMR, methylation, and expression data to achieve biomedical tasks of classification towards prognosis and diagnosis (Zhao et al. (2020, 2021); Wang et al. (2021); Huang et al. (2020)). Zhao et al. investigated the effect of different ratios of EMR features during an integrated analysis of CT images using the DensePriNet architecture (Zhao et al. (2021)). Huang et al. propose a joint representation of CT images and clinical features in a 'refine layer' to predict an output that performs better in detecting pulmonary embolism in comparison to other models (Huang et al. (2020)). MATCHER

utilises a joint fusion method to interpolate instances based on the alignment of multiple modalities to the pseudotime scale (Welch et al. (2017)).

Joint methods are more complex than late fusion but easier to interpret than early fusion methods. Using representations rather than the raw features from each modality greatly reduces the computational load compared to early fusion. However, joint harmonisation methods are not data agnostic and may require carefully curated model designs.

3.6.2 Multimodal Learning or Co-learning

Multimodal learning is designed to integrate and model the features from different modalities more comprehensively than unimodal learning. Joint fusions merge representations, but multimodal learning enables co-learning through a direct feed of all the features interacting at a higher, complex level. They can be classified into three distinct co-learning methods: probabilistic, multiple kernel learning and deep neural networks.

Probabilistic Models

Probabilistic methods involve building models that capture the relationships and dependencies between different modalities using joint or conditional probabilities. These models are highly interpretable, which allows for the models to integrate expert knowledge in the fusion approach, granting us the ability to interpret the results better than other methods. The random walk method uses probabilistic values to simulate a particle moving between nodes and layers in a network, establishing the relationships and links between the nodes (Baptista et al. (2022)).

MultiXRank module, published by Baptista et al., is an example of the probabilistic method of integration using a multiplex network of intramodal and intermodal interactions (protein-protein interactions, gene multiplex and disease monoplex networks) (Baptista et al. (2022)). Pio-Lopez et al. describe a use case of the random walk with restart architecture, wherein the method predicts long-distance gene-disease interactions using gene interaction network and disease similarity network data (Pio-Lopez et al. (2021)).

Probabilistic methods are applicable to any combination of modalities as long as they form a multiplex network. They rely heavily on theoretical knowledge to bridge relationships between elements of multiple domains and hence can be applied to data from any domain with multiplex and bipartite networks (Pio-Lopez et al. (2021)).

Multiple Kernel Learning

Kernels are linear classifiers that divide the data linearly using lenient boundaries, and a combined multitude of them assist in classifying non-linear heterogeneous data (Gönen and Alpaydın (2011)). This method is implemented in support vector machines (SVM), a popular method to analyse complex data.

Liu et al. used SVM to model MRI datasets from multiple sources towards Alzheimer's disease classification (Liu et al. (2013)). Lancktiet et al. predict the functions of yeast proteins using kernel-based learning (Lanckriet et al. (2003)). Multiple matrices describing the protein data were used in the algorithm, and results were reported on the different combinations of kernels used to classify the proteins as per their functions (Lanckriet et al. (2003)).

In MKL, different kernels are applied to each modality, and the combination of these kernels is learned to optimise the model's overall performance. Kernels can identify linear boundaries in datasets, making MKL a highly suitable method for classification tasks (Wilson et al. (2019)). Kernels can be combined in different ways (sum, product) to generate new kernels. A combination of multiple kernels accounts for a better classifier than using a single kernel (Hofmann et al. (2008)). This method is resistant to outliers but is susceptible to missing data (Wilson et al. (2019)).

Deep neural networks (DNNs)

Deep neural integration methods are characterised by a substantial number of neurons and layers constituting neural networks with significant depth and complexity. DNNs utilise representations of different modalities to reduce the features and pass them through high-level, intricate architectures, which enable them to uncover hidden information within the datasets.

DNNs are extensively used to understand data at a microscopic level, especially in the biomedical domain. EMR data can be modelled with omics modalities to shed light on physical and phenotypic changes and their relationships across time. Zhu et al. address an ML model to fuse and learn time-series data, with the use of Stacked Sparse Auto-Encoder (SSAE) and Long Short-Term Memory (LSTM) architecture (Zhu et al. (2021)). Zhang et al. have reported about OmiEmbed, a multitask deep-learning framework based on an autoencoder architecture (Zhang et al. (2021)). AffinityNet, proposed by Ma et al., uses k-nearest neighbours (kNN) attention pooling where the cluster representations of the data is processed as a GAT (Ma and Zhang (2019)). The method has asserted good performance for both labelled and unlabelled datasets.

DNNs are computationally expensive to perform due to their dense and complex architecture. They model data at high degrees of non-linearity, but the process becomes hard to decipher and elucidate. The meaning of data is lost when modelling and remains a black box with very low interpretability.

Table 3: A description of the integration methods and their advantages and disadvantages for a multimodal set-up.

Integration		Advantages	Disadvantages
Unimodal	Early	Performs lowest level of statistical correlation between features; Disregards selection bias	Computationally complex and expensive; Requires tailored models for effectively modelling and interpreting features
	Joint	Mitigates heterogeneity during modelling; Reduced feature size due to representations eases computational load	Requires tailored models for modelling joint representations
	Late	Independent models for different modalities; Ignores representation bias	Does not allow for feature interaction
Multimodal	Probabilistic	Highly interpretable; Establishes links between entities across different modalities	Requires domain-specific multilayer network information; Susceptible to missing data
	Multiple kernel Learning	Models based on overlapping results; Applicable for non-linear relationships; Resistant to outliers; Interpretable	Susceptible to missing data
	DNNs	Uncovers hidden information without explicit rules; Utilises complex architectures to understand the non-linear relationships between features and modalities	Low interpretability; Computationally complex and expensive

4 Guidelines for Model Selection

We propose ten recommendations for initiating a multimodal harmonisation analysis (Figure 3. Before starting out on an analysis, clearly articulate the objectives and aims of the study before initiating the harmonisation analysis. These objectives will guide subsequent data collection, representation, and model selection steps.

1. **Tailor Study Design to Objectives:** Tailor the study design to the defined objectives, taking into consideration the scale of the study and available resources. Ensure effective study design for sample identification and data collection that aligns with the study's goals.
2. **Implement Optimised Experimental Protocols:** The data is either already available or generated through new experiments. Employ or select optimised experimental protocols and assays for data collection, ensuring consistency and reliability. These protocols form the foundation for subsequent analysis steps and contribute to the quality of collected data.
3. **Digitised Data and Global Sharing:** Digitise collected data to facilitate analysis and global data sharing through repositories and databanks. Adopting data and metadata standards enhances data sharing and harmonisation. This step is crucial for collaborative research efforts and ensures data accessibility for future studies.

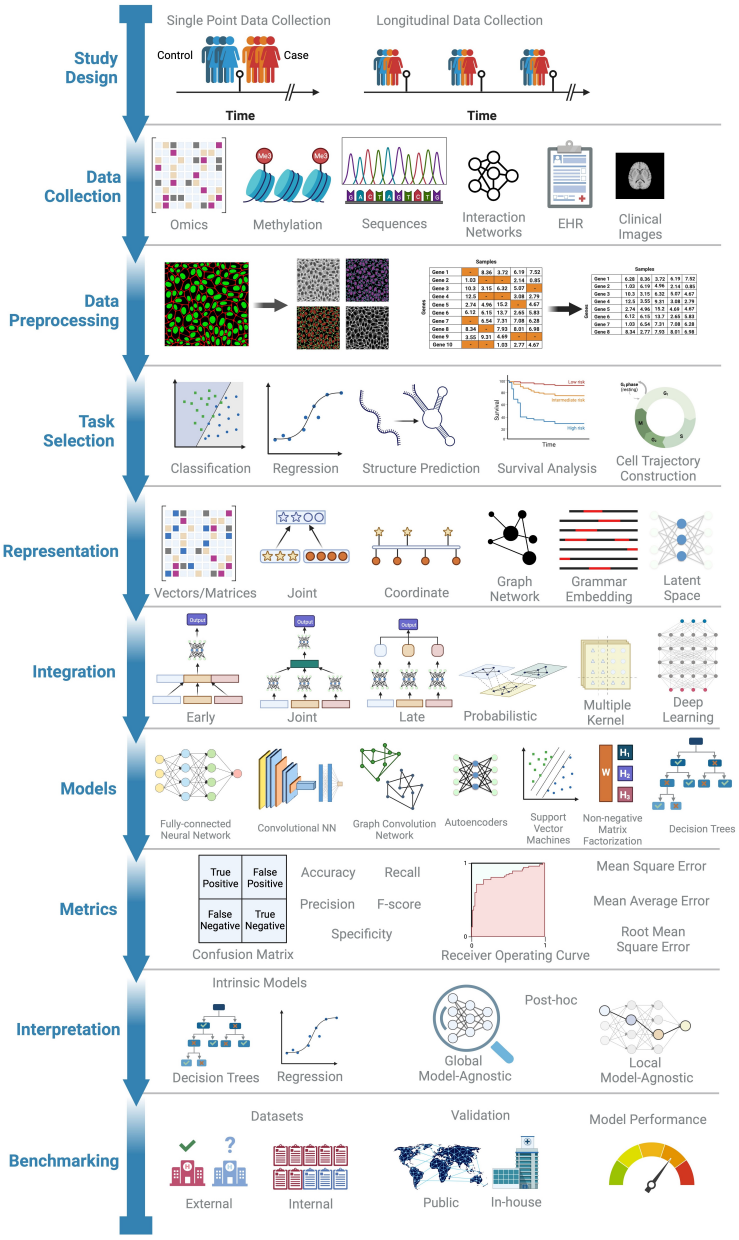

Fig. 3: A ten step guide flowchart that describes the process and order of execution to perform a multimodal integration. The titles on the left of the timeline describe the task order. The illustrations on the right are representative examples of different methods under each category. Made with BioRender.

4. **Modality Identification and Data Preprocessing:** Classify the collected data into the broad modalities discussed here and systematically process the data individually to create a working subset free of artefacts and low-quality elements. The analysis and results can vary depending on the level of preprocessing done.
5. **Task Selection:** Task selection should be guided by the study's goals, setting the stage for subsequent processing and analysis. Choose tasks that reflect the aims and are applicable to the dataset.
6. **Choose Feature Representation and Integration Methods Wisely:** The choice of representation or integration methods influences each other. Hence, select them based on the data type, number of modalities, level of harmonisation, and modality coherence. Recognise that a one-size-fits-all approach is impractical, and tailored methods may be needed for different tasks.
7. **Navigate Model Selection Complexity:** Different ML models can be employed for the same harmonisation set-up. Employ different models of varying complexity to assess the data and evaluate their performance using appropriate metrics.
8. **Model Performance Metrics:** Select model performance metrics corresponding to the task to compare and choose the optimal model. Provide an explanation of the metrics used and their relevance to the task.
9. **Prioritise Interpretable Models:** Prioritise using interpretable models, either intrinsic or through post-hoc interpretation. Especially in clinical settings, understanding how a model arrives at conclusions enhances trust and reliability.
10. **Validate and Benchmark Models:** Validate models on different datasets and sources to ensure robustness and generalizability. Benchmark models against state-of-the-art approaches and external datasets to mitigate aggregation and evaluation biases.

In the end, ensure that these recommendations are adapted to the specific context and goals of your multimodal harmonisation analysis.

5 Discussion

Lack of Comprehensive Reviews:

The article points out a noticeable gap in the existing literature regarding comprehensive explanations of workflow and procedures for integrating biomedical multimodal data. Multiple reviews for machine learning strategies to process multimodal data are available, but there is a deficit of articles relating them to biological and clinical data. A predominant part of research literature presents results with information from a single modality. Studies that utilise different biological modalities often interpret the results of independent analyses together. The concept of co-analysis, or more aptly, 'co-learning' is missed. There is a lack of clarity on how to effectively integrate data from disparate sources at the lowest item level to extract holistic knowledge.

Diverse Taxonomies in Multimodal Analysis:

Biomedical multimodal data from the same sample set is now routinely available from various research and development activities and healthcare. In the context of multimodal analysis, there is a distinction in the representation and integration steps

compared to unimodal analysis. We highlight the various analysis methods and the data types available under a limited set of taxonomic categorisation.

This classification of data types from biological and clinical sources allows one to identify methods that will suit the analysis of specific combinations and evaluate the advantages of each. We describe data harmonisation as a split of representation and integration methods, each with six distinct categories. Most steps are similar to an unimodal analysis, and the distinction in a multimodal analysis arises in the representation and the integration steps. The representation methods proposed emphasise on the features within the data. The various types of representation methods are key to uniformly present the multimodal data prior to an analysis. The section on integration focuses on the various methods to feed the data into ML architectures.

The article discusses this difference and provides insights into how to handle representation and integration methods for multimodal data effectively.

Framework and Model Suggestions for Biomedical Data Combinations:

There is a need for a structured framework or guideline to facilitate the harmonization process for multimodal data. The article aims to address this gap by presenting the first guideline framework towards a data harmonization process and providing a complete workflow. The recommended procedure consists of 10 steps to plan through towards a multimodal analysis.

To assist those undertaking harmonisation for the first time, we present a guide matrix showcasing examples from published literature, illustrating different combinations of data modalities. The combinations between the representation and integration methods are presented as a non-exhaustive list in table 4. Existing studies show that different choices can yield different results when using the same datasets (Huang et al. (2020)). The diverse taxonomies outlined in this paper can assist in understanding the significance of choosing an appropriate integration model for analysis, considering the concern related to biomedical data and model challenges.

Future Focus for Harmonisable Models:

The article acknowledges the challenges related to data and model selection in the context of multimodal analysis. It suggests that diverse taxonomies outlined in the paper can assist in understanding the significance of choosing an appropriate integration model for analysis, considering these challenges.

Data related challenges and model related challenges both arise when implementing a multimodal analysis. In addition to the challenges described in section 3.3 related to biomedical data, concerns on data acquisition and maintenance also require attention. The quality of biomedical data collected needs to be maintained, with appropriate measures taken for de-identification of the data and global sharing. A vast majority of the published literature on biomedical multimodal analysis focuses on the model metrics and parameters scores. However, due focus should be given to the model interpretability as well. Multimodal analysis with complex architectures may yield high performance scores, but they cannot be used to understand the biological and clinical data if the models are not interpretable. Interpretable models are needed to understand the process, especially with biomedical data to relate to further procedures, such as diagnosis and intervention strategies .

6 Conclusion:

The article highlights a significant gap in existing literature regarding the integration of multimodal data, noting a lack of comprehensive explanations and holistic views in current research. While recognizing the transformative potential of multimodal integration, it emphasizes the need for clarity on effectively integrating disparate data sources to extract comprehensive knowledge. Acknowledging challenges in data and model selection, the article proposes using diverse taxonomies to aid integration model selection. Addressing the distinction between unimodal and multimodal analysis, the article provides insights into representation and integration methods for multimodal data. Furthermore, it underscores the necessity for a structured framework to facilitate the harmonization process, presenting the first guideline framework and workflow. Additionally, it aims to assist researchers new to harmonization by offering a guide matrix featuring examples from published literature, aiding in the selection of appropriate integration models.

Declarations

Funding. MAMS acknowledges DBT and IITB-Monash research academy for the PhD sponsorship. ST acknowledges AISRF EMCR fellowship from Australian Academy of Science. MKM acknowledges support from the IITB Monash research academy and from Core Research Grant by Science and Engineering Research Board, India (Grant number: CRG/2022/008142).

Competing 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.

CRediT Authorship Contribution Statement. textbfMurali Aadhitya M S: Data curation, Visualisation, Writing - original draft. **Mithun Mitra:** Supervision, Validation, Writing - review & editing. **Sonika Tyagi:** Conceptualisation, Data curation, Supervision, Writing - original draft, Writing - review & editing

Acknowledgements. We thank Tyrone Chen for his suggestions regarding concepts of the work done.

Ethics approval and consent to participate: 'Not applicable'
Consent for publication: 'Not applicable'
Data availability: 'Not applicable'
Materials availability: 'Not applicable'
Code availability: 'Not applicable'

Table 4: Guide matrix providing examples from literature based on different combinations of modalities and the corresponding tasks, study design, representation, integration, model used, interpretability and reference. * are all representation methods

Modalities Integrated	Repn	Integration	Study Design	Task	Model Used	Model Interp	Validation	Reference
EMR*	Unimodal	MKL	Static	Prediction	SVM	High	Public, Internal	(Xu et al. (2010)
	Joint	Joint	Time Series	Risk Prediction	EN	Low	In-house, External	(Banerjee et al. (2019)
Gene Sequence*	Grammar	-	Static	RNA Structure Prediction	CFG	High	Public, Internal	(Tyagi et al. (2008)
	Grammar	-	Static	RNA Structure Prediction	CFG	High	In-house, Internal	(Andrikos et al. (2022)
	Grammar	-	Static	RNA Structure Prediction	CFG	High	Public, Internal	(Onokpasa et al. (2023)
Protein Sequence*	AE	-	Static	Clustering	VAE, ResNet, LSTM	Low	Public, Internal	(Detlefsen et al. (2022)
Disease network, miRNA	Graph	Probabilistic	Static	Disease Prediction	Katz	High	Public, Internal	(Zhang et al. (2017)
Image, EMR	Unimodal, Joint	Early, Joint, Late	Static	Classification	FCNN, CNN	Low	In-house, Internal	(Huang et al. (2020)
	Unimodal	Joint	Static	Classification	DDB	Low	In-house, Internal	(Zhao et al. (2020)
	Joint	Joint	Static	Classification, Segmentaion	DB, CNN, FCNN	Low	In-house, External	(Zhao et al. (2021)
Methylation, Image	Unimodal	Early	Static	Multi-variate Regression	GLM	High	In-house, Internal	(Schiano et al. (2020)
Metabolomics, Proteomics	Unimodal	Late	Static	Differential Correlation	-	High	In-house, Internal	(An et al. (2022)
Trancriptomics, DNA-Protein interaction network	Unimodal	Late	Static	Differential Correlation	-	High	In-house, Internal	(Wang et al. (2019)
Transctiptomics, Methylation	Unimodal	Late	Static	Differential Correlation	-	high	Public, Internal	(Xu et al. (2019)
	AE	Joint	Static	Classification	AE, FCNN	Low	Public, Internal	(Zhang et al. (2019)
	Graph, Joint	Late	Static	Classification	SNF	High	Public, Internal	(Wang et al. (2014)
Non-Bio	Graph	Probablistic	Static	Classification	GCN	Low	Public, Internal	(Ghorbani et al. (2019)
	AE	Joint	Static	Classification, Regression	SSAE, LSTM	Low	Public, Internal	(Zhu et al. (2021)
Proteomics, EMR	Joint	MKL	Static	Clasification	SVM	High	In-house, Internal	(Dargazanli et al. (2020)

Note: The table is continued in Table 5

Table 5: Continuation of Table 4. Guide matrix providing examples from literature based on different combinations of modalities and the corresponding tasks, study design, representation, integration, model used, interpretability and reference. * are all representation methods

Modalities Integrated	Repn	Integration	Study Design	Task	Model Used	Model Interp	Validation	Reference
Mutation; Drugs	Unimodal	Joint, Late	Static	Drug Prediction	CNN	Low	Public, Internal	(Chang et al. (2018)
Gene Expression, Chromatin Accessibility	Coordinate	Late	Time Series	Cell Trajectory Construction	MA	High	Public, Internal	(Welch et al. (2017)
Networks of Gene; Drug; Disease	Graph	Probablistic	Static	Link Prediction	RWR		Public, Internal	(Pio-Lopez et al. (2021)
	Graph	Probabilistic	Static	Link Prediction	RWR	High	Public, Internal	(Baptista et al. (2022)
Methylation, mRNA, miRNA	AE	DNN	Static	Classification	AE	Low	Public, Internal	(Franco et al. (2021)
	AE	Joint	Static	Classification, Regression, Survival Analysis	AE, FCNN, CNN	Low	Public, Internal	(Zhang et al. (2021)
	Graph	Joint	Static	Classification	GCN	Low	Public, Internal	(Wang et al. (2021)
	Unimodal, Graph	Late	Static	Classification	GCN	Low	Public, Internal	(Luo et al. (2023)
Lipidomics; Metabolomics; Proteomics; Transcriptomics	Unimodal	Joint	Static	Drug Prediction	PLSDA		Public, Internal	(Chen et al. (2021)
Methylation; Gene Expression; Mutation; Copy Number Alteration; Clinical	AE	Joint, Late	Static	Survival Analysis	FCNN	Low	Public, Internal	(Zhao et al. (2021)
Gene expression; Metabolomics; Proteomics; Cytokine measurements; Cytometric measurements; Microbiome	Unimodal	Late	Time Series	Multi-variate Estimation	EN	Low	In-house, Internal	(Ghaemi et al. (2018)
Interaction networks of drug compound, pharmacologic class, gene of action, pathway, biological process, disease, side effect, symptom, and anatomy	Graph	Probabilistc	Static	Link Prediction	EN	High	Public, Internal	(Himmelstein and Baranzini (2015)

References

Abarbanel, A., Evans, J.R., Budzynski, T.H., Budzynski, H.K.: Introduction to Quantitative EEG and Neurofeedback: Advanced Theory and Applications. Academic Press, ??? (2009)

Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature Medicine **28**(9), 1773–1784 (2022)

Andrikos, C., Makris, E., Kolaitis, A., Rassias, G., Pavlatos, C., Tsanakas, P.: Knotify: An efficient parallel platform for rna pseudoknot prediction using syntactic pattern recognition. Methods and Protocols **5**(1), 14 (2022)

Aburajab, R., Pospiech, M., Alachkar, H.: Profiling the epigenetic landscape of the antigen receptor repertoire: The missing epi-immunogenomics data. Nature Methods **20**(4), 477–481 (2023)

An, R., Yu, H., Wang, Y., Lu, J., Gao, Y., Xie, X., Zhang, J.: Integrative analysis of plasma metabolomics and proteomics reveals the metabolic landscape of breast cancer. Cancer & Metabolism **10**(1), 13 (2022) https://doi.org/10.1186/s40170-022-00289-6

Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence **41**(2), 423–443 (2018)

Baptista, A., Gonzalez, A., Baudot, A.: Universal multilayer network exploration by random walk with restart. Communications Physics **5**(1), 170 (2022)

Banerjee, I., Sofela, M., Yang, J., Chen, J.H., Shah, N.H., Ball, R., Mushlin, A.I., Desai, M., Bledsoe, J., Amrhein, T., *et al.*: Development and performance of the pulmonary embolism result forecast model (perform) for computed tomography clinical decision support. JAMA network open **2**(8), 198719–198719 (2019)

Barnum, G., Talukder, S., Yue, Y.: On the benefits of early fusion in multimodal representation learning. arXiv preprint arXiv:2011.07191 (2020)

Clarke, T.-K., Adams, M.J., Davies, G., Howard, D.M., Hall, L.S., Padmanabhan, S., Murray, A.D., Smith, B.H., Campbell, A., Hayward, C., *et al.*: Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in uk biobank (n= 112 117). Molecular psychiatry **22**(10), 1376–1384 (2017)

Church, K.W.: Word2vec. Natural Language Engineering **23**(1), 155–162 (2017)

Chen, T., Philip, M., Lê Cao, K.-A., Tyagi, S.: A multi-modal data harmonisation approach for discovery of covid-19 drug targets. Briefings in bioinformatics **22**(6),

737 185 (2021)

738 Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T.S., Jung, J., Shin,
739 J.-M.: Cancer drug response profile scan (cdrscan): a deep learning model that
740 predicts drug effectiveness from cancer genomic signature. Scientific reports **8**(1),
741 8857 (2018)

742 Chen, T., Tyagi, S.: Integrative computational epigenomics to build data-driven gene
743 regulation hypotheses. GigaScience **9**(6), 064 (2020)

744 Chen, T., Tyagi, N., Chauhan, S., Peleg, A.Y., Tyagi, S.: genomicbert and data-free
745 deep-learning model evaluation. bioRxiv, 2023–05 (2023)

746 Camastra, F., Vinciarelli, A.: Machine Learning for Audio, Image and Video Analysis:
747 Theory and Applications. Springer, ??? (2015)

748 Danaee, P., Ghaeini, R., Hendrix, D.A.: A deep learning approach for cancer detection
749 and relevant gene identification. In: Pacific Symposium on Biocomputing 2017, pp.
750 219–229 (2017). World Scientific

751 Detlefsen, N.S., Hauberg, S., Boomsma, W.: Learning meaningful representations of
752 protein sequences. Nature communications **13**(1), 1914 (2022)

753 Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S.: Big data in healthcare:
754 management, analysis and future prospects. Journal of big data **6**(1), 1–25 (2019)

755 Dargazanli, C., Zub, E., Deverdun, J., Decourcelle, M., Bock, F., Labreuche, J.,
756 Lefèvre, P.-H., Gascou, G., Derraz, I., Riquelme Bareiro, C., Cagnazzo, F., Bonafé,
757 A., Marin, P., Costalat, V., Marchi, N.: Machine learning analysis of the cerebrovas-
758 cular thrombi proteome in human ischemic stroke: An exploratory study. Frontiers
759 in Neurology **11** (2020) https://doi.org/10.3389/fneur.2020.575376

760 Fu, J., Rui, Y.: Advances in deep learning approaches for image tagging. APSIPA
761 Transactions on Signal and Information Processing **6**, 11 (2017)

762 Franco, E.F., Rana, P., Cruz, A., Calderon, V.V., Azevedo, V., Ramos, R.T., Ghosh,
763 P.: Performance comparison of deep learning autoencoders for cancer subtype
764 detection using multi-omics data. Cancers **13**(9), 2013 (2021)

765 Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. The Journal of Machine
766 Learning Research **12**, 2211–2268 (2011)

767 Ghorbani, M., Baghshah, M.S., Rabiee, H.R.: Mgcn: semi-supervised classification in
768 multi-layer graphs with graph convolutional networks. In: Proceedings of the 2019
769 IEEE/ACM International Conference on Advances in Social Networks Analysis and
770 Mining, pp. 208–211 (2019)

Ghaemi, M.S., DiGiulio, D.B., Contrepois, K., Callahan, B., Ngo, T.T.M., Lee-McMullen, B., Lehallier, B., Robaczewska, A., Mcilwain, D., Rosenberg-Hasson, Y., Wong, R.J., Quaintance, C., Culos, A., Stanley, N., Tanada, A., Tsai, A., Gaudilliere, D., Ganio, E., Han, X., Ando, K., McNeil, L., Tingle, M., Wise, P., Maric, I., Sirota, M., Wyss-Coray, T., Winn, V.D., Druzin, M.L., Gibbs, R., Darmstadt, G.L., Lewis, D.B., Partovi Nia, V., Agard, B., Tibshirani, R., Nolan, G., Snyder, M.P., Relman, D.A., Quake, S.R., Shaw, G.M., Stevenson, D.K., Angst, M.S., Gaudilliere, B., Aghaeepour, N.: Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. Bioinformatics **35**(1), 95–103 (2018) https://doi.org/10.1093/bioinformatics/bty537 https://academic.oup.com/bioinformatics/article-pdf/35/1/95/48962254/bioinformatics_35_1_95_s1.pdf

Garbh-Ini https://www.garbhinicohort.in

Himmelstein, D.S., Baranzini, S.E.: Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS computational biology **11**(7), 1004259 (2015)

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., Theis, F.J.: Diffusion pseudotime robustly reconstructs lineage branching. Nature methods **13**(10), 845–848 (2016)

Huang, S.-C., Kothari, T., Banerjee, I., Chute, C., Ball, R.L., Borus, N., Huang, A., Patel, B.N., Rajpurkar, P., Irvin, J., *et al.*: Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. NPJ digital medicine **3**(1), 61 (2020)

Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I., Lungren, M.P.: Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. Scientific reports **10**(1), 22147 (2020)

Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning (2008)

Kan, A.: Machine learning applications in cell image analysis. Immunology and cell biology **95**(6), 525–530 (2017)

Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., Luo, Y.: Multimodal machine learning in precision health: A scoping review. npj Digital Medicine **5**(1), 171 (2022)

Laney, D.: 3d data management: Controlling data volume, velocity and variety. META Group Research Note (2001)

807 Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into
808 document embedding generation. arXiv preprint arXiv:1607.05368 (2016)

809 Leitheiser, M., Capper, D., Seegerer, P., Lehmann, A., Schüller, U., Müller, K.-R.,
810 Klauschen, F., Jurmeister, P., Bockmayr, M.: Machine learning models predict the
811 primary sites of head and neck squamous cell carcinoma metastases based on dna
812 methylation. The Journal of Pathology **256**(4), 378–387 (2022)

813 Lanckriet, G.R., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S.: Kernel-
814 based data fusion and its application to protein function prediction in yeast. In:
815 Biocomputing 2004, pp. 300–311. World Scientific, ??? (2003)

816 Maturana, E., Alonso, L., Alarcón, P., Martín-Antoniano, I.A., Pineda, S., Piorno, L.,
817 Calle, M.L., Malats, N.: Challenges in the integration of omics and non-omics data.
818 Genes **10**(3), 238 (2019)

819 Luo, H., Liang, H., Jia, M., Li, Y., Yao, X., Cong, S.: Multi-omics integration for dis-
820 ease prediction via multi-level graph attention network and adaptive fusion. bioRxiv,
821 2023–03 (2023)

822 Li, Y., Ngom, A.: Data integration in machine learning. In: 2015 IEEE International
823 Conference on Bioinformatics and Biomedicine (BIBM), pp. 1665–1671 (2015).
824 IEEE

825 Lima, D.M., Rodrigues-Jr, J.F., Traina, A.J., Pires, F.A., Gutierrez, M.A.: Transform-
826 ing two decades of epr data to omop cdm for clinical research. Stud Health Technol
827 Inform **264**(August), 233–7 (2019)

828 Lee, B., Zhang, S., Poleksic, A., Xie, L.: Heterogeneous multi-layered network model
829 for omics data integration and analysis. Frontiers in genetics **10**, 1381 (2020)

830 Liu, F., Zhou, L., Shen, C., Yin, J.: Multiple kernel learning in the primal for mul-
831 timodal alzheimer's disease classification. IEEE journal of biomedical and health
832 informatics **18**(3), 984–990 (2013)

833 Malet-Martino, M., Holzgrabe, U.: Nmr techniques in biomedical and pharmaceutical
834 analysis. Journal of pharmaceutical and biomedical analysis **55**(1), 1–15 (2011)

835 Mou, M., Pan, Z., Lu, M., Sun, H., Wang, Y., Luo, Y., Zhu, F.: Application of machine
836 learning in spatial proteomics. Journal of Chemical Information and Modeling
837 **62**(23), 5875–5895 (2022)

838 Mansuri, M.S., Williams, K., Nairn, A.C.: Uncovering biology by single-cell pro-
839 teomics. Communications Biology **6**(1), 381 (2023)

840 Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., Ping, P.: Machine learning
841 and integrative analysis of biomedical big data. Genes **10**(2), 87 (2019)

Ma, T., Zhang, A.: Affinitynet: semi-supervised few-shot learning for disease type prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1069–1076 (2019)

Neidle, S.: 2 - the building-blocks of dna and rna. In: Neidle, S. (ed.) Principles of Nucleic Acid Structure, pp. 20–37. Academic Press, New York (2008). https://doi.org/10.1016/B978-012369507-9.50003-0 . https://www.sciencedirect.com/science/article/pii/B9780123695079500030

Nie, L., Wu, G., Culley, D.E., Scholten, J.C., Zhang, W.: Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Critical reviews in biotechnology **27**(2), 63–75 (2007)

Onokpasa, E., Wild, S., Wong, P.W.: Rna secondary structures: from ab initio prediction to better compression, and back. In: 2023 Data Compression Conference (DCC), pp. 278–287 (2023). IEEE

Paro, R., Grossniklaus, U., Santoro, R., Wutz, A.: Introduction to Epigenetics. Springer, ??? (2021)

Pio-Lopez, L., Valdeolivas, A., Tichit, L., Remy, É., Baudot, A.: Multiverse: a multiplex and multiplex-heterogeneous network embedding approach. Scientific reports **11**(1), 8794 (2021)

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. Bmj **372** (2021)

Pang, B., Weerd, J.H., Hamoen, F.L., Snyder, M.P.: Identification of non-coding silencer elements and their regulation of gene expression. Nature Reviews Molecular Cell Biology **24**(6), 383–395 (2023)

Ranganathan, P., Aggarwal, R.: Study designs: Part 3-analytical observational studies. Perspectives in clinical research **10**(2), 91 (2019)

Raghavachari, N., Garcia-Reyero, N.: Gene Expression Analysis. Springer, ??? (2018)

Ramakrishnaiah, Y., Macesic, N., Webb, G., Peleg, A.Y., Tyagi, S.: Ehr-qc: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes. Journal of Biomedical Informatics, 104509 (2023)

Sapoval, N., Aghazadeh, A., Nute, M.G., Antunes, D.A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., *et al.*: Current progress and open challenges for applying deep learning across the biosciences. Nature Communications **13**(1), 1728 (2022)

Schiano, C., Benincasa, G., Infante, T., Franzese, M., Castaldo, R., Fiorito, C., Mansueto, G., Grimaldi, V., Della Valle, G., Fatone, G., Soricelli, A., Nicoletti, G.F., Ruocco, A., Mauro, C., Salvatore, M., Napoli, C.: Integrated analysis of dna methylation profile of hla-g gene and imaging in coronary heart disease: Pilot study. PLOS ONE **15**(8), 1–16 (2020) https://doi.org/10.1371/journal.pone.0236951

Schafer, J.L.: Analysis of Incomplete Multivariate Data. CRC press, ??? (1997)

Saelens, W., Cannoodt, R., Todorov, H., Saeys, Y.: A comparison of single-cell trajectory inference methods. Nature biotechnology **37**(5), 547–554 (2019)

Sachdev, K., Gupta, M.K.: A comprehensive review of feature based methods for drug target interaction prediction. Journal of biomedical informatics **93**, 103159 (2019)

Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–9 (2021)

Subha, D.P., Joseph, P.K., Acharya U, R., Lim, C.M.: Eeg signal analysis: a survey. Journal of medical systems **34**, 195–212 (2010)

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E.: Big data: astronomical or genomical? PLoS biology **13**(7), 1002195 (2015)

Summaira, J., Li, X., Shoib, A.M., Li, S., Abdul, J.: Recent advances and trends in multimodal deep learning: a review. arXiv preprint arXiv:2105.11087 (2021)

Soylu, N.N., Sefer, E.: Bert2ome: Prediction of 2'-o-methylation modifications from rna sequence by transformer architecture based on bert. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2023)

Tyagi, S., Vaz, C., Gupta, V., Bhatia, R., Maheshwari, S., Srinivasan, A., Bhattacharya, A.: Cid-mirna: a web server for prediction of novel mirna precursors in human genome. Biochemical and biophysical research communications **372**(4), 831–834 (2008)

Tyagi, S.: Technical issues in implementing ai in healthcare. In: Translational Application of Artificial Intelligence in Healthcare, pp. 60–70. Chapman and Hall/CRC, ??? (2023)

Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D.: Genome-wide association studies. Nature Reviews Methods Primers **1**(1), 59 (2021)

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)

912 Vergoten, G., Theophanides, T.: Biomolecular Structure and Dynamics vol. 342.
913 Springer, ??? (2012)

914 Welch, J.D., Hartemink, A.J., Prins, J.F.: Matcher: manifold alignment reveals cor-
915 respondence between single cell transcriptome and epigenome dynamics. Genome
916 biology **18**(1), 1–19 (2017)

917 Wilson, C.M., Li, K., Yu, X., Kuan, P.-F., Wang, X.: Multiple-kernel learning for
918 genomic data mining and prediction. BMC bioinformatics **20**, 1–7 (2019)

919 Wang, C., Mahadevan, S.: A general framework for manifold alignment. In: 2009 AAAI
920 Fall Symposium Series (2009)

921 Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B.,
922 Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic
923 scale. Nature Methods **11**(3), 333–337 (2014) https://doi.org/10.1038/nmeth.2810

924 Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K.: Mogonet
925 integrates multi-omics data using graph convolutional networks allowing patient
926 classification and biomarker identification. Nature Communications **12**(1), 3445
927 (2021)

928 Wang, B., Wang, P., Parobchak, N., Treff, N., Tao, X., Wang, J., Rosen, T.: Integrated
929 rna-seq and chip-seq analysis reveals a feed-forward loop regulating h3k9ac and key
930 labor drivers in human placenta. Placenta **76**, 40–50 (2019)

931 Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.R.: Simple and efficient multiple kernel
932 learning by group lasso. In: Proceedings of the 27th International Conference on
933 Machine Learning (ICML-10), pp. 1175–1182 (2010). Citeseer

934 Xu, W., Xu, M., Wang, L., Zhou, W., Xiang, R., Shi, Y., Zhang, Y., Piao, Y.:
935 Integrative analysis of dna methylation and gene expression identified cervical
936 cancer-specific diagnostic biomarkers. Signal transduction and targeted therapy
937 **4**(1), 55 (2019)

938 Yang, S., Wang, Y., Lin, Y., Shao, D., He, K., Huang, L.: Lncmirnet: predicting lncrna–
939 mirna interaction based on deep learning of ribonucleic acid sequences. Molecules
940 **25**(19), 4372 (2020)

941 Zhao, L., Dong, Q., Luo, C., Wu, Y., Bu, D., Qi, X., Luo, Y., Zhao, Y.: Deepomix: A
942 scalable and interpretable multi-omics deep learning framework and application in
943 cancer survival analysis. Computational and structural biotechnology journal **19**,
944 2719–2725 (2021)

945 Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal
946 of the Royal Statistical Society Series B: Statistical Methodology **67**(2), 301–320
947 (2005)

Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms. CRC press, ??? (2012)

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., Hoffman, M.M.: Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion **50**, 71–91 (2019)

Zhao, X., Wang, X., Xia, W., Li, Q., Zhou, L., Li, Q., Zhang, R., Cai, J., Jian, J., Fan, L., *et al.*: A cross-modal 3d deep learning for accurate lymph node metastasis prediction in clinical stage t1 lung adenocarcinoma. Lung Cancer **145**, 10–17 (2020)

Zhao, X., Wang, X., Xia, W., Zhang, R., Jian, J., Zhang, J., Zhu, Y., Tang, Y., Li, Z., Liu, S., *et al.*: 3d multi-scale, multi-task, and multi-label deep learning for prediction of lymph node metastasis in t1 lung adenocarcinoma patients' ct images. Computerized Medical Imaging and Graphics **93**, 101987 (2021)

Zhang, X., Xing, Y., Sun, K., Guo, Y.: Omiembed: a unified multi-task deep learning framework for multi-omics data. Cancers **13**(12), 3047 (2021)

Zhang, X., Zou, Q., Rodriguez-Paton, A., Zeng, X.: Meta-path methods for prioritizing candidate disease mirnas. IEEE/ACM transactions on computational biology and bioinformatics **16**(1), 283–291 (2017)

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., Guo, Y.: Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 765–769 (2019). IEEE

Zhang, S.-F., Zhai, J.-H., Xie, B.-J., Zhan, Y., Wang, X.: Multimodal representation learning: advances, trends and challenges. In: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1–6 (2019). IEEE

Zhu, Q., Zhang, S., Zhang, Y., Yu, C., Dang, M., Zhang, L.: Multimodal time series data fusion based on ssae and lstm. In: 2021 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–7 (2021). IEEE