

Multiview Learning for Omics Data Integration: From Multi-Modal Data Fusion to Systems-Level Biological Insights



Constance B. Bailey^{1*}

Abstract

The rapid expansion of omics technologies has, somewhat paradoxically, both clarified and complicated our understanding of biological systems. While genomics, transcriptomics, proteomics, and related modalities provide unprecedented detail, each—on its own—seems to capture only a fragment of a much larger, deeply interconnected biological narrative. It is within this tension that multiview learning has begun to emerge, not as a definitive solution, but rather as a flexible and evolving framework for integration. This review explores how multiview learning approaches attempt to reconcile heterogeneous, high-dimensional omics datasets into coherent representations of biological systems. We examine the conceptual foundations underlying integration—particularly the balance between consensus and complementarity—and trace the progression from classical statistical models, such as canonical correlation analysis, to more recent deep learning architectures. Along the way, we consider three dominant fusion strategies—early, intermediate, and late integration—each offering distinct advantages and limitations. Particular attention is given to how these methods address persistent challenges, including dimensionality imbalance, modality heterogeneity, and data incompleteness. Through synthesis of methodological and application-oriented studies, this review highlights the growing role of

multiview learning in areas such as cancer subtyping, biomarker discovery, and drug response prediction. Ultimately, the field appears to be shifting—quietly but decisively—toward a central insight: that meaningful biological understanding increasingly depends not on individual data layers, but on how effectively they are integrated. **Keywords:** Multi-omics integration; Multiview learning; Data fusion; Systems biology; Machine learning

Keywords

Multi-omics integration; Multiview learning; Data fusion; Systems biology; Machine learning

1. Introduction

In recent years, the trajectory of biological research—perhaps somewhat quietly at first, and then unmistakably—has shifted toward a data-intensive paradigm. What once relied heavily on reductionist, single-layer observations has now expanded into a multidimensional landscape shaped by high-throughput sequencing and advanced analytical platforms. The emergence of “omics” technologies—spanning genomics, transcriptomics, proteomics, metabolomics, and epigenomics—has not merely added volume to biological data; it has fundamentally altered how we conceptualize biological systems themselves. Rather than discrete,

Author Affiliation:

¹ School of Chemistry, University of Sydney, Camperdown, NSW 2001, Australia.

Please cite this article as:

C. B. B. (2021). "Multiview Learning for Omics Data Integration: From Multi-Modal Data Fusion to Systems-Level Biological Insights", *Bioinfo Chem*, 3(1), 1-12, 10736

© 2021 Bioinfo Chem, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

Significance | Reveals how multiview learning transforms fragmented omics data into integrated biological insight, enabling more precise and interpretable systems-level understanding.

*Correspondence.

School of Chemistry, University of Sydney, Camperdown, NSW 2001, Australia. E-mail: constance.bailey@sydney.edu.au,

Editor Chenxiao Xu, Ph.D. And accepted by the Editorial Board Sep 24, 2021 (received for review Jul 28, 2021)

independently functioning components, genes, proteins, and metabolites are increasingly understood as elements of deeply interconnected networks that co-evolve and co-regulate cellular behavior .

Yet, despite the promise embedded in this abundance of data, a certain tension persists. Single-omics studies, while undeniably informative, often provide only fragmented glimpses into biological processes. They capture snapshots—sometimes precise, sometimes noisy—but rarely the full dynamic interplay that governs phenotype expression or disease progression. This limitation has prompted a growing recognition: to approach something resembling a systems-level understanding, one must move beyond isolated data layers and instead integrate them in a coherent, biologically meaningful way (Hasin et al., 2017; Gomez-Cabrero et al., 2014).

It is within this context that multiview learning has emerged—not as a singular solution, but rather as a conceptual and computational framework capable of navigating the complexity of multi-omics data. At its core, multiview learning treats each omics dataset as a distinct “view” of the same underlying biological entity. These views, while individually informative, collectively encode a richer, more nuanced representation of biological reality. The challenge, however, lies in how to reconcile their differences—differences in scale, noise structure, dimensionality, and even biological interpretation (Li et al., 2016; Sun, 2013).

The difficulty is not trivial. Biological datasets are often characterized by what is commonly referred to as the “large p, small n” problem—thousands, sometimes tens of thousands, of features measured across relatively few samples. This imbalance introduces statistical instability and increases the risk of overfitting. Compounding this issue is the heterogeneity inherent in multi-omics data: transcriptomic measurements may not align neatly with proteomic outputs, and epigenetic modifications may obscure or modulate gene expression in ways that are not immediately apparent. Such discrepancies are not merely technical artifacts; they reflect genuine biological complexity that integration models must account for rather than ignore (Ahmad & Fröhlich, 2016; Rappoport & Shamir, 2018).

To address these challenges, several integration strategies have been proposed, each with its own conceptual strengths and limitations. Early

integration—often described as feature-level fusion—combines data from multiple omics layers into a single matrix before analysis. While appealing in its simplicity, this approach tends to obscure modality-specific structures and may amplify noise when datasets differ substantially in scale or distribution (Pavlidis et al., 2001). Late integration, in contrast, analyzes each omics layer independently and subsequently combines the results, such as clustering outputs or predictive scores. This method preserves modality-specific insights but may miss subtle cross-modal interactions that only become apparent when data are jointly modeled (Bickel & Scheffer, 2004).

Somewhere between these two extremes lies intermediate integration, which arguably represents the most conceptually compelling approach. Here, data are integrated during the learning process itself, allowing models to capture both shared and modality-specific patterns simultaneously. Techniques such as Joint and Individual Variation Explained (JIVE) and multi-omics factor analysis exemplify this strategy, decomposing datasets into components that reflect common biological signals as well as unique variations (Lock et al., 2013; Argelaguet et al., 2018). This dual representation is particularly valuable in biological contexts, where both consensus and diversity across data types carry meaningful information.

Indeed, the theoretical underpinnings of multiview learning often revolve around two complementary principles: the consensus principle and the complementary principle. The former emphasizes agreement across views, seeking latent representations that are consistent across multiple data modalities. The latter, however, acknowledges that each omics layer captures distinct aspects of biological function—DNA methylation patterns, for instance, may reveal regulatory mechanisms that are invisible at the transcriptomic level. Effective integration, therefore, requires a careful balancing act: extracting shared structure without discarding modality-specific insights (Blum & Mitchell, 1998; Li et al., 2016).

Historically, a range of statistical methods laid the groundwork for multiview integration. Canonical Correlation Analysis (CCA), introduced by Hotelling (1936), provided one of the earliest frameworks for identifying relationships between paired datasets. Its modern extensions, including sparse CCA, have been widely in genomic studies to uncover correlated patterns across omics layers (Witten & Tibshirani, 2009). Similarly,

matrix factorization techniques—such as Non-negative Matrix Factorization (NMF)—have enabled the decomposition of complex datasets into interpretable components, facilitating the identification of underlying biological processes (Seung & Lee, 1999; Zitnik & Zupan, 2015).

Network-based approaches have also gained prominence, particularly in the form of Similarity Network Fusion (SNF), which constructs sample similarity networks for each data type and integrates them into a unified representation. This method has demonstrated notable success in cancer subtyping, where integrating genomic, transcriptomic, and epigenomic data can reveal clinically relevant patient clusters that are not apparent from single-omics analyses alone (Wang et al., 2014; Hoadley et al., 2014).

More recently, the advent of deep learning has expanded the methodological landscape even further. Models such as deep canonical correlation analysis and variational autoencoders have introduced the capacity to model highly non-linear relationships across data modalities. These approaches, while often criticized for their limited interpretability, have shown considerable promise in tasks such as disease classification, biomarker discovery, and drug response prediction (Andrew et al., 2013). Their ability to learn latent representations that capture complex cross-modal interactions suggests a powerful, albeit still evolving, direction for multi-omics integration.

And yet, despite these advances, certain questions remain unresolved. How can we ensure that integrated representations are not only statistically robust but also biologically interpretable? To what extent do current models capture causal relationships rather than mere correlations? And perhaps most importantly, how can these computational frameworks be translated into clinically actionable insights?

This review, therefore, seeks to navigate these questions by examining the evolution of multiview learning approaches for omics data integration. By tracing the progression from classical statistical models to contemporary deep learning architectures, we aim to highlight both the conceptual continuity and the methodological innovation that define this field. In doing so, we hope to underscore a central, if somewhat tentative, conclusion: that meaningful biological understanding increasingly depends not on the depth of individual data layers, but on the coherence with which

they are integrated.

2. Methodology

This narrative review was designed to provide a structured yet interpretive synthesis of the evolving landscape of multiview learning approaches in multi-omics data integration. Unlike systematic reviews, which emphasize exhaustive retrieval and quantitative aggregation, the present methodology adopts a concept-driven and theory-informed approach, allowing for a more flexible exploration of methodological developments, conceptual frameworks, and translational implications within the field (Li et al., 2016).

2.1 Literature Identification and Scope Delimitation

The literature base for this review was primarily derived from the reference corpus provided within the source manuscript, supplemented by foundational and highly cited works in multi-view learning, machine learning, and systems biology. Emphasis was placed on studies published prior to and around 2019 to maintain conceptual consistency with the review's temporal scope. Key thematic areas included: (i) multi-omics data integration strategies, (ii) multiview learning principles, (iii) statistical and machine learning models for heterogeneous data, and (iv) biomedical applications such as cancer subtyping and drug response prediction.

Rather than conducting a database-driven systematic search, sources were selected based on their conceptual relevance, citation prominence, and methodological contribution. Seminal works—such as those introducing canonical correlation analysis (Hotelling, 1936), Joint and Individual Variation Explained (Lock et al., 2013), and multi-omics factor analysis (Argelaguet et al., 2018)—were prioritized to anchor the review in established theoretical foundations.

2.2 Inclusion Criteria and Conceptual Filtering

Studies were included if they met at least one of the following criteria: (i) proposed or significantly advanced a multiview or multi-omics integration method; (ii) addressed computational challenges such as dimensionality, heterogeneity, or missing data; (iii) demonstrated application in biomedical contexts, particularly using publicly available datasets (e.g., TCGA, CCLE); (iv) contributed to the theoretical understanding of consensus and complementarity in multiview learning.

Conversely, studies were excluded if they focused solely on single-omics analysis without integrative components, lacked methodological clarity, or provided purely application-based results without computational or conceptual insight. This filtering process ensured that the review remained focused on integration-centric methodologies rather than domain-specific findings alone.

2.3 Thematic Organization and Synthesis Strategy

The synthesis process followed a thematic structuring approach, organizing the literature into interconnected domains rather than chronological progression. Four major thematic axes guided the synthesis:

- **Conceptual Foundations of Multiview Learning** This included principles such as consensus and complementarity (Blum & Mitchell, 1998; Li et al., 2016), which underpin most integration strategies.

- **Fusion Strategies and Integration Architectures** Studies were categorized into early, intermediate, and late integration frameworks, enabling comparative evaluation of their assumptions and limitations (Pavlidis et al., 2001; Bickel & Scheffer, 2004).

- **Algorithmic and Statistical Methodologies** Methods such as CCA, sparse CCA, matrix factorization, and deep learning models were analyzed in terms of mathematical formulation, interpretability, and performance (Witten & Tibshirani, 2009; Andrew et al., 2013; Goodfellow et al., 2016).

- **Biomedical Applications and Translational Relevance** Application-oriented studies were synthesized to highlight how integration frameworks contribute to real-world problems such as cancer classification and drug response prediction (Hoadley et al., 2014; Barretina et al., 2012).

This thematic structuring allowed for a layered interpretation, where methodological developments could be examined alongside their practical implications.

2.4 Comparative and Interpretive Analysis

A qualitative comparative analysis was employed to evaluate different multiview learning approaches. Rather than relying on quantitative metrics alone, the review emphasizes interpretability, robustness, scalability, and biological relevance. For example, classical statistical methods were contrasted with deep learning architectures to highlight the trade-off between transparency and predictive performance (Mo et al., 2013; Liang et al.,

2014).

Tables included in the manuscript (Tables 1–4) were used as integrative tools to systematically map relationships between paradigms, computational challenges, algorithmic performance, and biomedical applications. These tables function not merely as summaries, but as analytical frameworks supporting cross-comparison and synthesis.

2.5 Handling Bias and Limitations in Narrative Synthesis

Given the non-systematic nature of narrative reviews, potential selection bias and subjectivity were addressed through deliberate reliance on widely cited, peer-reviewed sources and foundational studies. Additionally, cross-referencing between studies was used to ensure consistency in interpretation and to avoid over-reliance on single-source perspectives (Gomez-Cabrero et al., 2014; Hasin et al., 2017).

However, it is acknowledged that this approach may underrepresent emerging or less-cited methodologies. The review, therefore, does not aim for exhaustive coverage but rather for conceptual depth and coherence.

2.6 Methodological Orientation

Overall, this narrative review adopts an integrative and interpretive methodological stance. It bridges statistical learning theory, computational biology, and translational medicine to construct a cohesive understanding of multiview learning in omics integration. By focusing on conceptual clarity and methodological evolution, the review seeks to provide a meaningful synthesis that informs both future research directions and practical applications in systems biology and precision medicine.

3. Biomedical Multi-view Learning.

3.1 From Fragmented Signals to Integrated Biological Understanding

It is perhaps not an overstatement—though it may feel like one at first—to say that biology has undergone a quiet but profound redefinition over the past two decades. What was once a discipline rooted largely in observation and descriptive inference has gradually, and almost inevitably, become something far more data-intensive, even predictive in its ambitions. This shift has been driven, in large part, by the rapid maturation of high-throughput and next-generation sequencing technologies, which now allow researchers to interrogate biological systems across multiple molecular strata simultaneously. These

layers—genomics, transcriptomics, epigenomics, proteomics, metabolomics—collectively referred to as “omics,” do not merely add detail; they complicate, enrich, and sometimes even challenge our understanding of how biological systems behave.

And yet, despite this abundance of information, there remains a lingering sense that something is missing. Single-omics analyses, while undeniably powerful, often yield only partial narratives. They isolate one layer of biology, one dimension of regulation, and attempt to explain complex phenotypes through that lens alone. But biological systems, as we increasingly appreciate, are not modular in such a simplistic sense. A change in DNA methylation may ripple through transcriptional networks, alter protein abundance, and ultimately reshape metabolic states. These interactions are not linear, nor are they always predictable. They are, instead, deeply interwoven—suggesting that a more integrative approach is not just beneficial, but necessary (Hasin et al., 2017; Ahmad & Fröhlich, 2016).

3.2 The Emergence of Multi-View Learning in Biomedical Contexts

It is within this conceptual gap that biomedical multi-view learning (MVL) has found its footing. At a glance, MVL might appear to be simply another branch of machine learning, but its significance lies in how it reframes the problem of data integration. Rather than treating heterogeneous datasets as inconvenient or incompatible, MVL embraces them as complementary “views” of the same biological system. Each omics layer, in this framework, becomes a distinct yet related representation of an underlying biological truth. The analytical challenge, then, is not merely to combine these views, but to do so in a way that preserves their individual contributions while uncovering the relationships that bind them together (Li et al., 2016).

This, of course, is easier said than done. The literature leading up to 2019 consistently highlights the technical and statistical barriers that complicate such integration. Biological data are notoriously heterogeneous—not just in content, but in structure. Transcriptomic data, for instance, may be represented as count-based measurements, while proteomic data often reflect continuous abundance values influenced by post-translational modifications. These discrepancies are not trivial; they introduce layers of noise and bias that can obscure meaningful biological signals if

not carefully addressed (Rappoport & Shamir, 2018).

3.3 The Multi-Modal Challenge: Dimensionality, Noise, and Heterogeneity

Compounding this issue is the well-known “curse of dimensionality.” In most omics studies, the number of measured features—genes, transcripts, proteins—vastly exceeds the number of available samples. This imbalance ($n \ll p$) poses significant challenges for statistical modeling, increasing the risk of overfitting and reducing the generalizability of learned patterns. Models trained under such conditions may perform well on training data, yet fail to capture stable, reproducible biological relationships (Vapnik, 1999).

Moreover, the noise inherent in biological data is not merely technical—it is often biological in origin. Variability between samples may reflect genuine biological diversity, but it may also arise from experimental inconsistencies or measurement errors. Distinguishing between these sources of variation is, in many cases, non-trivial. Naïve approaches—such as simple feature concatenation—tend to amplify these issues, blending signal and noise in ways that are difficult to disentangle.

3.4 Conceptual Foundations: Balancing Consensus and Complementarity

To navigate these complexities, multiview learning relies on two foundational principles: consensus and complementarity. At first glance, these concepts might seem almost contradictory. The consensus principle seeks commonality—it assumes that different data views should, in some latent sense, agree with one another. By identifying shared structures across modalities, models can enhance robustness and reduce susceptibility to view-specific noise (Blum & Mitchell, 1998).

Complementarity, on the other hand, emphasizes difference. It acknowledges that each omics layer provides unique information that cannot be inferred from others. For example, while genomic mutations may indicate potential disruptions in gene function, epigenomic modifications reveal regulatory mechanisms that operate independently of sequence variation. Ignoring these distinctions would mean discarding valuable biological insight (Li et al., 2016).

The challenge, then, lies in balancing these two principles—extracting shared biological signals without suppressing modality-specific nuances. It is a delicate

equilibrium, one that many models strive for, though not all achieve with equal success.

3.5 Fusion Strategies: Early, Intermediate, and Late Integration

The question of when to integrate data—rather than simply how—has led to the development of three primary fusion strategies: early, intermediate, and late integration. Each approach reflects a different philosophical stance on the nature of multi-omics data.

Early integration, or feature-level fusion, combines all data modalities into a single matrix prior to analysis. This approach is appealing in its simplicity and compatibility with standard machine learning algorithms. However, it often overlooks the unique statistical properties of each data type, leading to potential distortions and redundancy (Pavlidis et al., 2001).

Late integration takes a more modular approach. Here, each omics dataset is analyzed independently, and the resulting outputs—such as cluster assignments or predictive scores—are combined at a later stage. While this preserves modality-specific insights, it may fail to capture subtle interactions that only emerge through joint analysis (Bickel & Scheffer, 2004).

Intermediate integration, by contrast, attempts to bridge these two extremes. Data are integrated during the learning process itself, allowing models to capture cross-modal relationships in a unified framework. Techniques such as Joint and Individual Variation Explained (JIVE) exemplify this approach, decomposing datasets into shared and modality-specific components (Lock et al., 2013). Similarly, multi-omics factor analysis provides a probabilistic framework for uncovering latent structures that span multiple data types (Argelaguet et al., 2018). These methods, while computationally more complex, offer a more nuanced representation of biological systems.

3.6 Methodological Milestones: From Linear Models to Deep Architectures

Historically, the field of multiview learning has been shaped by a series of methodological milestones. Canonical Correlation Analysis (CCA), introduced by Hotelling (1936), remains one of the earliest and most influential techniques for identifying relationships between paired datasets. Its extensions, such as sparse CCA, have been particularly valuable in high-dimensional settings, enabling the selection of relevant features while

maintaining interpretability (Witten & Tibshirani, 2009).

Matrix factorization techniques have also played a central role. Methods such as JIVE and related frameworks decompose data into interpretable components, separating shared biological signals from modality-specific variation. This decomposition is especially useful in biological contexts, where understanding the source of variation is as important as modeling it (Lock et al., 2013).

Network-based approaches, such as Similarity Network Fusion (SNF), represent another important advancement. By constructing similarity networks for each data modality and integrating them into a consensus network, SNF captures non-linear relationships that may be missed by linear models. Its success in cancer subtyping underscores the potential of integrative approaches in clinical applications (Wang et al., 2014; Shen et al., 2009).

The advent of deep learning has, more recently, expanded the methodological horizon. Deep autoencoders and variational autoencoders offer powerful tools for dimensionality reduction, learning latent representations that capture complex, non-linear interactions across modalities (Goodfellow et al., 2016). Deep canonical correlation analysis further extends these ideas, enabling the modeling of relationships that are not easily captured by traditional methods (Andrew et al., 2013). Still, these models introduce new challenges—particularly in terms of interpretability and reproducibility—that the field continues to grapple with.

3.7 Clinical Relevance and Emerging Directions

Ultimately, the promise of multiview learning lies not in methodological elegance, but in its potential for clinical translation. In the context of precision medicine, integrated multi-omics analyses have already begun to reshape how diseases—particularly cancers—are classified and treated. By identifying molecular subtypes that respond differently to therapy, clinicians can move toward more personalized treatment strategies, improving outcomes while minimizing unnecessary interventions (Hoadley et al., 2014). Beyond classification, MVL also plays a role in drug discovery and response prediction. By linking molecular profiles with pharmacological data, researchers can identify candidate drugs that are more likely to be effective for specific patient populations. This, in turn, opens the door to more targeted and efficient therapeutic development.

And yet, despite these advances, the field remains in flux. Questions of interpretability, scalability, and clinical applicability continue to shape ongoing research. There is a growing recognition that predictive accuracy alone is not sufficient; models must also provide insights that are biologically meaningful and clinically actionable.

4. Synthesising the Landscape of Multi-View Omics Integration

4.1 From Computational Possibility to Biological Necessity

If one looks closely—perhaps more closely than is comfortable—the evolution of multi-view learning (MVL) in omics integration does not appear as a sudden methodological breakthrough, but rather as a gradual, almost inevitable convergence of computational need and biological reality. What began as an abstract challenge—how to combine heterogeneous datasets—has, over time, become central to how we interpret biological systems themselves. Multi-omics integration, in this sense, is no longer optional; it is foundational to systems biology.

The synthesis of prior work, particularly before 2019, reveals a field moving through stages of experimentation toward something more structured, even principled. The four tables presented earlier—spanning MVL paradigms, computational solutions, algorithmic performance, and biomedical applications—collectively provide a scaffold for understanding this progression. As summarised in Table 1, MVL frameworks evolved not simply as technical tools, but as conceptual models that reflect different assumptions about biological data integration. Meanwhile, Table 2 outlines how these frameworks respond to persistent computational challenges, Table 3 compares their performance trade-offs, and Table 4 grounds these methods in real-world biomedical datasets and applications.

Together, these tables do not just catalogue methods—they map a shifting intellectual landscape.

4.2 The Subtle Ascendancy of Intermediate Fusion

One of the more consistent patterns emerging from the literature is the gradual, and perhaps unsurprising, preference for intermediate fusion strategies. Early integration—essentially concatenating features across modalities—once seemed like a reasonable starting point. It is, after all, straightforward. But this simplicity quickly reveals its limitations. High dimensionality, differing

statistical distributions, and modality-specific noise often render such approaches unstable or, at best, biologically shallow (Ahmad & Fröhlich, 2016; Rappoport & Shamir, 2018).

Late fusion, on the other hand, offers flexibility. By modeling each omics layer independently, it preserves modality-specific structure. Yet, it also introduces a kind of fragmentation—an inability to capture the subtle “cross-talk” between layers that defines biological regulation (Rappoport & Shamir, 2018).

It is here that intermediate fusion begins to feel less like an option and more like a necessity. As reflected in Table 1, methods such as iCluster exemplify how joint latent variable models can uncover shared biological states across genomic and epigenomic layers (Shen et al., 2009). These approaches do not merely combine data—they allow the integration process itself to become a site of learning. In doing so, they align more closely with the hierarchical and interconnected nature of biological systems (Li et al., 2016).

One might even argue, cautiously, that intermediate fusion represents a computational analogue of biological integration itself.

4.3 Algorithmic Trade-offs: Interpretability Versus Complexity

The comparative performance of MVL algorithms—summarised in Table 3—reveals a recurring tension. On one side are classical statistical models, valued for their interpretability. On the other are more complex, often non-linear approaches, which offer improved predictive performance but at the cost of transparency.

Correlation-based methods, particularly sparse canonical correlation analysis (CCA), have long served as a foundation. Their appeal lies in their clarity: they identify linear relationships between datasets, often revealing biologically interpretable associations such as SNP–gene interactions (Witten & Tibshirani, 2009). Similarly, matrix factorization techniques like JIVE decompose data into shared and modality-specific components, providing a structured way to separate biological signal from technical noise (Lock et al., 2013). Yet, these methods are not without limitations. Their reliance on linear assumptions can constrain their ability to capture the non-linear dynamics inherent in biological systems. Graph-based

Table 1. Taxonomy and Core Paradigms of Multi-View Learning. This table summarizes the major learning paradigms in multi-view learning (MVL), outlining their objectives, integration logic, and representative methods. It highlights how different paradigms leverage modality-specific and shared information to address biological complexity in multi-omics data integration.

Learning Paradigm	Primary Objective	Integration Logic	Core Principle	Representative Method	Common Modalities	Key Advantage	Reference
Supervised	Predict known labels	Task-guided fusion	Task relevance	MFDA	mRNA + Mutation	High predictive power	Chen & Sun (2009)
Unsupervised	Discover latent groups	Hypothesis-free	Pattern discovery	iCluster	DNAm + miRNA	Unbiased discovery	Shen et al. (2009)
Semi-supervised	Use unlabeled data	Agreement-based	Co-training	Co-training SVM	MRI + Clinical	Efficient with sparse labels	Blum & Mitchell (1998)
Representation Learning	Extract embeddings	Dimensionality reduction	Shared latent space	JIVE	Proteomics + mRNA	Reduces data complexity	Lock et al. (2013)
Deep Learning	Model non-linearities	Neural architectures	Abstraction	Deep CCA	SNPs + Imaging	Captures complex cross-talk	Andrew et al. (2013)
Graph-based	Preserve topology	Network fusion	Connectivity	SNF	CNV + mRNA	Robust to local noise	Wang et al. (2014)
Matrix Factorization	Disentangle variation	Decomposition	Joint & individual structure	iNMF	Metabolomics + Gene Expression	Separates shared vs specific signals	Yang & Michailidis (2016)
Kernel Methods	Map to high-dim space	Similarity-based	Complementarity	MKL	PPI + DNAm	Flexible across data types	Gönen & Alpaydin (2011)
Probabilistic	Model uncertainty	Generative latent variables	Consensus	MOFA	scRNA + Methylation	Handles missing data	Argelaguet et al. (2018)
Ensemble	Aggregate decisions	Late fusion	Robustness	Random Forest	Clinical + Multi-omics	Reduces overfitting	Breiman (2001)

Table 2. Methodological Solutions for Computational Challenges in Multi-Omics. This table outlines key computational challenges encountered in multi-omics integration and the methodological strategies developed to address them. It links biological issues with algorithmic solutions and highlights their role in improving interpretability and robustness.

Challenge Category	Biological Issue	Computational Solution	Integration Stage	Implementation	Modality Context	Biological Goal	Reference
Dimensionality	High features, low samples	Sparse regularization	Intermediate	Sparse CCA	mRNA + miRNA	Biomarker identification	Witten & Tibshirani (2009)
Heterogeneity	Scaling differences	Metric fusion	Intermediate	Conflation	Metabolomics + Proteomics	Cross-modal alignment	Mitra et al. (2020)*
Noisy Modalities	Experimental error	Denoising autoencoder	Intermediate	DAE-Net	DNA methylation	Robust feature extraction	Vincent et al. (2010)
Regulatory Linkage	Information flow	Knowledge-guided graph	Intermediate	PARADIGM	CNV + mRNA	Pathway activity inference	Vaske et al. (2010)
Missing Data	Partial observations	Matrix completion	Intermediate	PVC	Single-cell data	Maximize data usage	Li et al. (2014)
Batch Effects	Technical variation	Normalization	Early	XPN	Cross-platform GE	Comparative analysis	Shabalin et al. (2008)
Data Interaction	Complex cross-talk	Non-linear embedding	Intermediate	Deep DBN	Drug + Gene Expression	Drug synergy modeling	Liang et al. (2014)
Redundancy	Overlapping signals	Joint decomposition	Intermediate	JIVE	MRI + PET	Separate shared vs specific	Lock et al. (2013)
Inconsistency	Modality mismatch	Agreement penalty	Intermediate	Co-regularization	GE + Proteomics	Consensus clustering	Sindhwani et al. (2005)
Systemic Logic	Hierarchical interactions	Multi-layer networks	Intermediate	HMLN	Genome → Phenome	Mechanistic modeling	Lee et al. (2014)

*Note: Mitra et al. (2020) builds on earlier conflation principles (Hill, 2011).

methods, such as Similarity Network Fusion (SNF), offer a different perspective. By focusing on relationships between samples rather than features, they construct similarity networks that can be integrated into a global consensus. As demonstrated in cancer datasets, SNF often outperforms feature-based approaches in tasks such as patient stratification and survival prediction (Wang et al., 2014). This robustness, particularly in the presence of noisy data, makes graph-based models particularly appealing.

Then there is deep learning—a category that, prior to 2019, was still emerging within this domain. Methods like Deep CCA and deep belief networks introduced the ability to model highly non-linear interactions across modalities (Andrew et al., 2013; Liang et al., 2014). Their performance, especially in clustering and classification tasks, is often superior. And yet, there remains a hesitation—a recognition that these models, while powerful, are not easily interpretable. In clinical contexts, where understanding why a model makes a prediction can be as important as the prediction itself, this limitation cannot be ignored (Mo et al., 2013).

4.4 Addressing Computational Bottlenecks: Lessons from Methodological Innovations

If one turns to Table 2, a different narrative begins to emerge—one focused not on performance, but on problem-solving. Each row in the table represents a challenge: dimensionality, heterogeneity, noise, missing data. And each solution reflects a broader methodological shift. Sparse regularization, for instance, addresses the imbalance between features and samples, enabling models to focus on the most relevant variables (Witten & Tibshirani, 2009). Denoising autoencoders attempt to filter out experimental noise, preserving meaningful biological patterns (Vincent et al., 2010). Knowledge-guided models such as PARADIGM incorporate prior biological information, moving beyond purely data-driven approaches (Vaske et al., 2010). What is perhaps most striking is how these solutions increasingly blur the boundary between statistical modeling and biological reasoning. Integration is no longer just about combining datasets—it is about embedding biological knowledge into computational frameworks.

4.5 Translational Impact Across Biomedical Domains

The true measure of any methodological framework, however, lies in its application. As shown in Table 4, MVL

has been extensively applied across major biomedical datasets, each offering a distinct lens on disease. In precision oncology, datasets such as The Cancer Genome Atlas (TCGA) have demonstrated the power of integrated analysis. By combining multiple omics layers, researchers have identified molecular subtypes that transcend traditional tissue classifications—an insight that challenges conventional approaches to cancer treatment (Hoadley et al., 2014). This has, in turn, influenced the design of “basket trials,” where therapies are matched to molecular profiles rather than anatomical origin. In pharmacogenomics, integration of datasets like CCLE and GDSC has improved drug response prediction. By linking molecular profiles with chemical properties, models can identify potential drug–target interactions and even suggest synergistic combinations (Barretina et al., 2012; Li et al., 2016).

Neuroinformatics offers yet another example. In Alzheimer’s disease research, combining imaging data with genomic information has enabled more accurate prediction of disease progression. Integrated models from the ADNI dataset have shown that multimodal approaches outperform single-modality analyses in predicting cognitive decline (Ahmed et al., 2013). These applications suggest that MVL is not merely a computational convenience—it is a pathway to clinically meaningful insight.

4.6 Navigating Incomplete Data: The Role of Probabilistic Models

A persistent challenge in multi-omics research is missing data. Not all samples are measured across all modalities, leading to incomplete datasets that can undermine traditional integration methods. Probabilistic frameworks, particularly Multi-Omics Factor Analysis (MOFA), have emerged as a compelling solution. By modeling latent variables and leveraging correlations across views, MOFA can effectively “fill in the gaps,” allowing for the integration of partially observed datasets (Argelaguet et al., 2018). Its application in diseases such as chronic lymphocytic leukemia has revealed latent biological processes that were previously undetectable. This ability to work with incomplete data is not just a technical advantage—it reflects a deeper shift toward more flexible, adaptive modeling strategies.

4.7 Concluding Perspective

Taken together, the synthesis of pre-2019 MVL research suggests a field that is both mature in its foundations and

Table 3. Performance Characteristics of Multi-View Integration Algorithms. This table compares widely used MVL algorithms based on their mathematical foundations, fusion strategies, and performance metrics. It provides insight into their suitability for specific biomedical applications.

Algorithm	Mathematical Basis	Fusion Approach	Latent Representation	Dimensionality Handling	Best Application	Performance Metric	Reference
iCluster	Probabilistic PCA	Early/Intermediate	Latent variables	Low-dimensional	Cancer subtyping	Cluster purity	Shen et al. (2009)
SNF	Graph diffusion	Intermediate	Consensus network	Iterative	Patient stratification	Survival significance	Wang et al. (2014)
JIVE	Matrix factorization	Intermediate	Shared + specific	PCA-based	Signal separation	Variance explained	Lock et al. (2013)
Sparse CCA	Correlation maximization	Intermediate	Linear projections	SVD-based	SNP-gene associations	Canonical correlation	Witten & Tibshirani (2009)
MCCA	Generalized CCA	Intermediate	Joint variates	High $\rightarrow$ Low	Multi-modal imaging	Sum correlation	Kettenring (1971)
NMF	Matrix decomposition	Intermediate	Part-based	Non-negative factors	Module discovery	Reconstruction error	Lee & Seung (1999)
PINS	Perturbation-based	Late	Stability-driven	Variable	Disease subtyping	Stability index	Nguyen et al. (2017)
LRACluster	Low-rank approximation	Early	Global latent	Reduced rank	Pan-cancer clustering	Log-likelihood	Wu et al. (2015)
Deep CCA	Neural networks	Intermediate	Non-linear latent	Multi-layer	Vision & omics	Non-linear correlation	Andrew et al. (2013)
iClusterPlus	Bayesian regression	Intermediate	Latent variables	Mixed types	Heterogeneous omics	Deviance	Mo et al. (2013)

Table 4. Key Biomedical Datasets and Multi-View Application Domains. This table presents widely used biomedical datasets that have driven multi-view learning applications. It highlights their disease contexts, omics modalities, and clinical relevance in translational research.

Dataset	Disease Context	Omics Views	Sample Size	Integration Goal	Clinical Utility	Access	Reference
TCGA	Pan-cancer	mRNA, DNAm, CNV	Thousands	Subtype classification	Precision oncology	Public	Hoadley et al. (2014)
GDSC	Cancer cell lines	Mutation, Drug IC50	~1,000	Drug response prediction	Therapy selection	Public	Yang et al. (2012)
CCLE	Cancer sensitivity	GE, pharmacogenomics	~900	Sensitivity modeling	Biomarker discovery	Public	Barretina et al. (2012)
ENCODE	Human genome	ChIP-seq, RNA-seq	Thousands	Functional annotation	Regulatory biology	Open	ENCODE (2012)
ADNI	Alzheimer's	MRI, PET, SNP	Hundreds	Disease progression	Early diagnosis	Restricted	Ahmed et al. (2013)
GTEx	Healthy tissues	GE, genotype	~900	eQTL mapping	Gene regulation	Public	GTEx Consortium (2013)
ROADMAP	Epigenomics	DNAm, histone marks	Hundreds	Epigenomic mapping	Risk factors	Public	Bernstein et al. (2010)
CTRP	Drug response	Bioactivity + omics	Thousands	Target identification	Drug repositioning	Public	Seashore-Ludlow (2015)
LINCS	Cell perturbation	L1000, chemicals	Millions	Signature mapping	Mechanism discovery	Open	Duan et al. (2014)
METABRIC	Breast cancer	GE, CNV, survival	~2,000	Prognostic subtyping	Risk stratification	Public	Curtis et al. (2012)

still evolving in its ambitions. Statistical models such as CCA and JIVE established the groundwork for interpretability. Graph-based methods expanded the ability to model complex relationships. Deep learning introduced a new level of representational power—though not without trade-offs. What remains consistent, however, is the central insight: that integrating multiple omics views provides a more robust, nuanced, and clinically relevant understanding of biological systems than any single modality alone. And perhaps, as these methods continue to evolve, the question will shift—from how we integrate data, to how we interpret the increasingly complex representations that integration produces.

5. Limitations

This review, by design, adopts a narrative rather than a systematic approach, and this choice inevitably introduces certain limitations. The selection of literature, while guided by relevance and citation prominence, remains somewhat subjective, potentially overlooking emerging or less-cited contributions. Additionally, the focus on pre-2019 developments—intended to maintain conceptual consistency—means that more recent advances, particularly in deep learning and large-scale integrative frameworks, may not be fully represented. There is also an inherent imbalance in the literature itself. Many studies emphasize methodological innovation over biological validation, which complicates efforts to assess real-world applicability. Furthermore, while this review discusses interpretability as a key concern, it does so largely at a conceptual level rather than through formal evaluation metrics. As such, the conclusions drawn here should be understood not as definitive, but as informed interpretations shaped by the current—and still evolving—state of the field.

6. Conclusion

Multiview learning, in many ways, reflects a broader shift in biological research—from isolated observation toward integrative understanding. What emerges from this review is not a single dominant method, but rather a spectrum of approaches, each attempting—imperfectly, yet progressively—to reconcile the complexity of multi-omics data. There remains, however, a certain tension between predictive performance and interpretability, between computational sophistication and biological meaning. And perhaps that tension is not a weakness, but a signal—an indication that the field is still negotiating its direction. If anything seems clear, it is this: integration is no longer

optional. It is, increasingly, the foundation upon which meaningful biological insight must be built.

Author Contributions

C.B.B. conceptualized the study, designed the review framework, conducted literature synthesis and analysis, interpreted the findings, and drafted, reviewed, and finalized the manuscript.

References

- Ahmad, A., & Fröhlich, H. (2016). Integrating heterogeneous omics data via statistical inference and learning techniques. *gene expression*, 4, 5.
- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 1247–1255).
- Argelaguet, R., Velten, B., Arnl, D., Dietrich, S., Zenz, T., Marioni, J. C., ... & Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124.
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., ... & Cancer Genome Atlas Network. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), 603–607.
- Bickel, S., & Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the 4th IEEE International Conference on Data Mining* (pp. 19–26).
- Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92–100).
- Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merckenschlager, M., Gisel, A., ... & Tegnér, J. (2014). Data integration in the era of omics: Current and future challenges. *BMC Systems Biology*, 8(Suppl 2), I1.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18, 1–15.
- Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., ... & Cancer Genome Atlas Network. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4), 929–944.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.
- Li, Y., Wu, F. X., & Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2), 325–340.
- Liang, M., Li, Z., Chen, T., & Zeng, J. (2014). Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4), 928–937.
- 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. *Annals of Applied Statistics*, 7(1), 523–542.
- Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., ... & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11), 4245–4250.
- Nguyen, T., Tagett, R., Diaz, D., & Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12), 2025–2039.
- Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2001). Gene functional classification from heterogeneous data. In *Proceedings of the 5th International Conference on Computational Molecular Biology* (pp. 242–248).
- Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20), 10546–10562.
- Seung, H. S., & Lee, D. D. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791.
- 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(22), 2906–2912.
- Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7), 2031–2038.
- Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 988–999.
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3), 333–337.
- Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), Article 28.
- Yang, Z., & Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1), 1–8.
- Zitnik, M., & Zupan, B. (2015). Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1), 41–53.