THE USA JOURNALS
THE AMERICAN JOURNAL OF ENGINEERING AND TECHNOLOGY (ISSN – 2689-0984)
VOLUME 06 ISSUE12

PUBLISHED DATE: - 31-12-2024
PAGE NO.: - 247-263

RESEARCH ARTICLE **Open Access**

Machine Learning For Multi-Biomarker Screening In Cancer Diagnosis: Integrating Genomic, Proteomic And Metabolic Signatures

Shakhnoza Khamidova
Independent Researcher,USA

Abstract

Cancer screening increasingly requires the integration of molecular signals that capture different aspects of tumor biology. Genomic, proteomic, and metabolic biomarkers provide complementary information, but their high dimensionality, heterogeneous distributions, and biological interdependence make their combined interpretation challenging using conventional statistical approaches. This review examines the application of machine learning (ML) to multi-biomarker cancer screening, with particular emphasis on the integration of genomic, proteomic, and metabolic signatures. Genomic alterations provide relatively stable information concerning tumor identity and molecular lineage, whereas proteomic and metabolic profiles reflect downstream changes in cellular signaling, protein activity, and metabolism. The integration of these molecular layers therefore has the potential to improve discrimination between malignant and non-malignant states while providing information relevant to tumor classification and localization. The review discusses major ML strategies used to harmonize multi-omics data, reduce dimensionality, select informative biomarker features, and integrate heterogeneous molecular measurements through early, intermediate, and late fusion approaches. Evidence from multi-cancer and cancer-specific studies indicates that integrated biomarker models can achieve strong diagnostic discrimination and may outperform individual molecular layers in selected settings. Applications in colorectal, ovarian, breast, hepatocellular, pancreatic, and other cancers demonstrate the potential of combining circulating biomarkers and tissue-derived molecular profiles for early detection and disease characterization. However, promising model performance does not necessarily translate into clinical screening utility. Low tumor-derived signal in early disease, limited cohort sizes, heterogeneous sample processing, class imbalance, external validation gaps, and difficulties in model interpretation remain important limitations. Future progress will require prospective multicenter validation, longitudinal sampling, standardized multi-omics workflows, clinically interpretable biomarker panels, and evaluation in populations containing relevant benign conditions and early-stage disease. Overall, ML-enabled multi-biomarker integration represents a promising framework for cancer screening, but its clinical value will ultimately depend on demonstrating reliable, generalizable, and clinically actionable performance in real-world screening populations.

Keywords Cancer Diagnosis, Proteomic, Metabolic, Genomic, Multi Biomarker

INTRODUCTION

Cancer is a molecularly diverse disease in which genetic changes, abnormal protein signaling and metabolic reprogramming are all intertwined to create a variety of phenotypes in cancer cells. This heterogeneity restricts the use of single molecules as tumor markers as the same molecular change can happen only in a part of the tumors, or it can happen in normal physiological and inflammatory

processes as well. Therefore, emphasis has shifted to multi-biomarker signatures that consider several molecular pattern characteristics together to capture complementary characteristics of malignant transformation. The different measurements (genomic, proteomic and metabolomic) give different information in a biological context. Genomic profiles may detect somatic mutations and copy-number changes, as well as other molecular abnormalities that are linked to the growth of the tumor, while proteomic measurements can detect changes in the abundance and activity of proteins in the tumor. Metabolomic profiles are downstream biochemical effects of the changes in cell regulation and can include changes in energy metabolism and biosynthetic pathways. These molecular layers are linked but not the same expression of a disease. Proportional changes in protein abundance may not result from genetic alterations and metabolic phenotypes may be affected by both the alterations within the tumor and the microenvironment. This integration thus offers the chance to describe the malignancy within various levels of biological organization. The ability of this approach to be diagnostic is demonstrated by large scale molecular profile analysis of circulating molecules. To explore cancer discrimination through machine-learning models, Smelik et al. combined the data from nearly 500,000 individuals from the UK Biobank with 90 million genetic variants, 2,923 circulating proteins, 325 metabolites, and clinical variables. They found significant differences in the amount of predictive information provided by molecular modalities and cancer types, with tissue proteomic information showing better discrimination than circulating molecular information; and blood-based multi-omic models did not show a consistent advantage over clinical variables [1]. This study shows that an increase in the dimensionality of the data does not necessarily lead to an improvement in the diagnostic yield, and that several factors, including the biological compartment, disease phenotype and information content of each modality must be taken into account when building the models. Moreover, because multi-omic data is high dimensional, associated with correlated variables, data is not normally distributed and is often imbalanced (between number of molecular features vs numbers of clinical samples), machine learning is particularly relevant to this study. Regularized regression, tree algorithms, support vector machines, dimensionality-reduction techniques and ensemble learning can be used to learn informative combinations while keeping the model's complexity small. The application of LASSO, random forests, support vector machines, gradient-boosting methods and neural-network approaches to high-throughput tumor-marker selection by Fu et al. highlights the growing trends of computational feature selection and classification in molecular biomarker datasets [2].

But the goal isn't just to maximize classification accuracy; a useful diagnostic signature should also be biologically relevant and be repeatable across other patients in the cohort that it was based upon. Integration can be of special interest, if individual biomarker gives incomplete representation of disease. For instance, metabolomics might be able to detect biochemical changes that are not directly observable by genomics, and proteomics might be able to detect changes in function between the genomic changes and the metabolic phenotype. Multi-omics analyses have demonstrated that integrated transcriptomic, proteomic and metabolomic analyses can recover molecular factors associated with distinct breast cancer subgroups and survival-related phenotypes, highlighting the potential of multi-omics approach to uncover the biological structure spread across different molecular layers. These observations help lay the groundwork for investigating multimodal signatures as opposed to individual analytes, and to highlight the distinction between molecular

association and clinically relevant diagnostic prediction. While this potential is great, there are also significant methodological challenges with multi-biomarker screening. It is possible for apparently strong models to be sub-optimal in external validation due to a variety of batch effects, missing modalities, platform-specific measurement variation, feature redundancy, population heterogeneity, or overfitting. In addition, before the added complexity of a molecular signature can be justified, it is necessary that it be able to provide additional value beyond the already known clinical and/or laboratory markers. To develop clinically useful models, therefore, the biological complementarity, feature selection, multimodal integration, predictive performance, model calibration, and independent validation, must all be considered together.

This review summarizes the machine-learning methods used to combine the genomic, proteomic and metabolic characteristics in cancer detection and diagnosis. Special focus is given to: (i) the transformation of heterogeneous molecular measurements to composite biomarker signatures, (ii) selection of informative features, and (iii) the evaluation of integrated models in various cancer settings. The principal limitations to clinical translation are also discussed, including lack of reproducibility, lack of interpretability, analytical standardization, and validation in independent patient cohorts

Biological Basis of Genomic, Proteomic and Metabolic Signatures

The concept of combining the genomic, proteomic and metabolic data is based on the hierarchical nature of the relationship between molecular alteration and cellular phenotype in cancer diagnosis. Genomic abnormalities, including mutations, copy-number alteration and other genome instability, may trigger or maintain oncogenic programs, however, the phenotypic effects of the genomic abnormalities are achieved through alterations in transcription and translation. Proteomic measurements thus add a functional dimension between genomic regulation and cellular phenotype and metabolomic measurements represent biochemical outcomes of changes in enzyme activity, nutrient utilization, redox state and biosynthetic demand. Together, these molecular layers can reveal correlations that are not evident when analyzed separately, allowing for the discovery of relationships that are not found by analyzing the data from each modality individually. The results of these layers in the context of integrated tumor profiling show that these are not just redundant measurements of the same process. Using whole-exome sequencing, whole-transcriptome sequencing, proteomics and untargeted metabolomics, and drawing on spatial and single-cell analyses, Hu et al. analyzed 100 treatment-naive clear cell renal cell carcinoma (ccRCC) tumors. These integrated analysis identified four molecular subtypes and highlighted a distinct de-clear cell differentiated (DCCD) subtype associated with clearance of lipid droplets, reduction of metabolic activity, enhancement of nutrient uptake and enhanced proliferative activity. Importantly, these phenotypic characteristics were validated at molecular level by showing pathways changes associated with angiogenesis, glycolysis and immune processes at transcriptomic and proteomic level, and abnormalities of amino-acid and lipid related metabolic pathways at metabolomic level [3]. The ccRCC findings highlight an important concept in screening for multiple biomarkers - that a molecular phenotype could be identified from multiple data modalities, not necessarily a single highly discriminatory variable. Within the DCCD subtype, molecular changes in the metabolism were correlated with transcriptomic and proteomic changes and spatial metabolomics gave information on the spatial distribution of the metabolism changes in tumor tissue. The

investigators also noted that the DCCD phenotype was also correlated with survival even among patients with stage I and II disease, highlighting the fact that molecular heterogeneity may have clinically relevant information that was not fully captured by conventional anatomic staging.

The same integrative relationship was found in breast cancer with transcriomic, proteomic and metabolomic measures of tumors. We applied multi-omics factor analysis (MOFA+) to build latent factors that capture coordinated variation in the molecular modalities of a cohort of 335 untreated patients from the Oslo2 study. The first two latent factors explained 25.7% and 16.2% of the molecular variance, respectively: the latent factor with protein expression levels explained a significant amount of variance, while transcriptomic and metabolomic measurements explained a smaller amount. The molecular groups generated by the clustering based on the survival associated latent factors were found to have significantly different long-term outcomes and the resulting molecular stratification was assessed in the TCGA PanCancer Atlas and METABRIC cohorts [4]. This method illustrates the usefulness of latent-variable modeling, when molecular information exists in feature spaces which are correlated but not identical.

Alternatively, the interactions between specific molecular layers can be seen at the level of specific oncogenic pathways. The 773 Chinese breast cancer patients were comprehensively profiled using genomic, transcriptomic, proteomic, metabolomic, radiomic and digital pathology measurements. The integrated analysis revealed a greater frequency of AKT1 mutations and HER2-enriched features compared to other mostly White populations that have been described to date. In the HR+ / HER2+ subgroup, the amplification of ERBB2 was associated with an increase in abundance of HER2 protein, highlighting how both genetic and proteomic changes can contribute to defining a clinically relevant tumor phenotype. Basal-like tumors were also also linked to ferroptosis processes by integrated proteomic and metabolomic studies [5]. These observations illustrate the need to view a genomic alteration in the context of the functional phenotype that is achieved downstream of the alteration to better understand the molecular or biological significance of the alteration. In fact, at the population level, the specific contribution of each molecular layer may vary greatly. Using a machine-learning pipeline that involved data cleaning, imputation, feature selection, cross-validation and independent holdout testing, Smelik et al compared the genetic variants, proteins, and metabolites of around 90 million individuals in the UK Biobank. In the complex diseases studied, protein measurements yielded more predictive information than did metabolites, and genetic variants yielded the least predictive information among the three classes of molecules. Remarkably, models using only 5 proteins per disease performed very well with AUCs of 0.79 for incident disease, and 0.84 for prevalent disease, highlighting that high-dimensional biomarker systems do not require very high number of features at the end. Smelik et al. used machine-learning models that included proteins, metabolites, genetic variants, and clinical variables on peripheral-blood multi-omics data from about 500,000 individuals in the UK Biobank, including 32,260 with a diagnosis of cancer. The AUCs for collective cancer discrimination against healthy controls were highest for the proteomic features (0.75), clinical variables (0.74), genomic features (0.72) and metabolomic features (0.64). Importantly, discrimination decreased when the reference population was changed from healthy individuals to patients with non-cancer diseases: the proteomic AUC decreased from 0.75 to 0.58, genomic from 0.72 to 0.56, clinical from 0.74 to 0.51 and metabolomic from 0.64 to 0.54. The significant decrease in discrimination suggests that

discrimination signals that were selected against healthy controls may reflect the systemic disease-associated variations instead of cancer-specific molecular signals.

These results collectively favor a biological model where the genomic, proteomic and metabolic signatures ought to be viewed not as identical pools of biomarkers, but as complementary molecular signatures. These changes, at the genomic level, can give clues to the molecular basis of the malignant phenotype; pathway activation and functional consequences can be measured at the proteomic level; and downstream biochemical adaptation at the metabolomic level. Their integration, therefore, cannot be used diagnostically unless the features used are biologically meaningful and clinically relevant. It serves as the basis for machine-learning methods that aim more than just to join molecular datasets to identify cross-modal patterns while reducing redundancy, technical variation and overfitting.

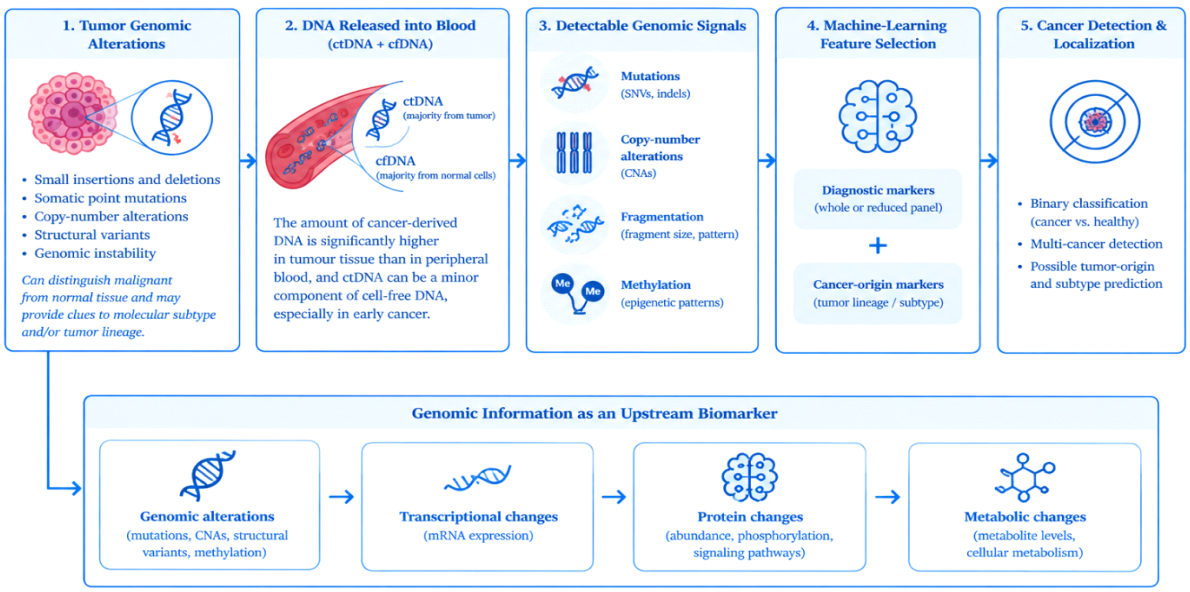
Molecular Biomarker Sources and Their Diagnostic Contribution

Genomic Signatures in Cancer Detection

Genomic alterations serve as a fairly durable molecular sign of malignancy and are thus a highly valuable marker for multi-biomarker screening. Small insertions and deletions, somatic point mutations, copy-number alterations, structural variants, and patterns of genomic instability can help distinguish malignant tissue from normal tissue, and in some cancers, can provide clues to the molecular subtype and/or tumor lineage. The value of genomic information for diagnosis however, is highly dependent upon the biological material that is analyzed. The amount of cancer-derived DNA is significantly higher in the tumour tissue than in the peripheral blood, and circulating tumour DNA (ctDNA) could be a minor component of cell-free DNA, especially in early cancer. This is significant when screening with machine learning, because the more a signal is present in a genome, the more discriminatory it is not. A model may detect a lot of alterations in established tumors, but not be able to detect it in plasma, because these alterations are present only in small numbers. As a result, other characteristics of the genome, such as combinations of mutations, copy-number profiles, fragment characteristics and DNA methylation, have been included in increasingly widespread genomic screening. In particular, methylation is very suitable for liquid biopsy because epigenetic changes are present across broader regions of the genome than individual somatic mutations. In the 2024 TOTEM study, the principle is applied to targeted plasma cfDNA methylation sequencing of 733 patients with seven cancer types and 500 healthy individuals. The investigators chose 57 markers for multiple cancers and 873 cancer-signal-origin markers [6]. They performed well on their binary classification, with 0.907 AUC for training, 0.908 AUC for test data, and 0.868 AUC for an independent validation set; importantly, they were still able to make meaningful predictions if they reduced the model to just 21 diagnostic and 214 origin markers. The results show that a meaningful representation of the diagnostic information in genomic information, or an extremely high dimensional molecular profile, can be reduced to a comparatively lower dimensional representation. However, the moderate sensitivity rates seen for early disease stages point to another key limitation of blood-based genomic screening: the molecular signal can be informative, but not strong. Genomic changes also offer a mechanism to explain the changes in the biomarkers downstream. A mutation in a signalling pathway may influence the abundance of protein, its phosphorylation, metabolism of the cell and transcription. So, genomic variables are not necessarily markers of disease, but may be more like upstream markers that provide clues to the mechanisms behind the metabolic and proteomic perturbations.

Proteomic Signatures and Functional Tumor Phenotypes

Proteomic measurements are a more accurate reflection of the functional state of a tumour cell than genomic measurement alone, since proteins are the main molecular effectors of cellular processes. Changes in protein abundance, secretion, post-translational modification, and changes in pathway activity can then be used to identify phenotypic changes not detectable at the DNA level. Circulating proteins are especially interesting for diagnostic purposes because they can be measured in plasma/serum using widely-used laboratory platforms, and may also be a reflection of tumour derived products and/or systemic responses to malignancy. Importantly, the circulating proteome is not solely comprised of proteins derived from tumor cells. Blood protein concentrations may be significantly affected by inflammation, coagulation, complement activation, tissue injury and host immune responses. A diagnostic protein signature could thus be a mixture of tumour biology and host response and not a direct molecular profile of the tumour. The difference was observed by Xu et al. in [7] in the plasma proteomics study of non-small-cell lung cancer (NSCLC). A total of 942 proteins were quantified and 180 of these were found to be differentially expressed, including those involved in the complement and coagulation systems and in cholesterol metabolism. They then tested 10 protein candidates in addition to amino acids, bile acids and known tumor markers in 110 patients with non-small cell (NSCLC) lung cancer, 108 patients with benign pulmonary disease and 100 healthy controls. Their logistic-regression model with integration performed well with an AUC of 0.959 for the distinction between NSCLC and benign pulmonary disease. The study is of special value for the field of biomarker development as a clinically relevant non-malignant comparator was included in the study, instead of the sole population of healthy individuals.



Genomic signatures as upstream biomarkers in multi-biomarker cancer screening

Even a more direct representation of the tumor phenotype can be achieved using tissue proteomics. Xing et al. 2023 conducted an integrated HCC study that described the genomic, transcriptomic, proteomic and phosphoproteomic features in 160 patients and classified them into three proteomic subtypes characterized by different clinical outcomes [8]. An alternative, simplified nine-protein panel related to the molecular classification by metabolic reprogramming was then suggested and external data sets were examined for reproducibility of the molecular classification. These results are indicative of the ability of proteomic signatures to link molecular changes with the observed tumour characteristics, but it is more difficult to translate tissue-based signatures to minimally invasive screening than circulating ones.

Metabolic Signatures of Malignant Transformation

The metabolic traits of malignant transformation.Metabolic properties of malignant transformation. Metabolomic biomarkers are located at the bottom of the molecular hierarchy and thus reflect the biochemical consequences of genomic changes, signalling activity, nutrient availability, and tumour to host tissue interactions. Cell metabolism can change in many ways in malignant cells, including glucose utilization, amino-acid metabolism, lipid synthesis and oxidation, nucleotide metabolism, and mitochondrial activity. These variations cause the relative amounts and levels of circulating metabolites to change, which can be detected without the need to directly sample the tumor tissue, potentially producing biochemical signatures. One of the reasons for the diagnostic value of metabolites is the fact that they are close to cell phenotype. A genomic alteration could suggest the possibility of a biological change, while a metabolite concentration could be an indication of the biochemical status present at the time of sampling. This is also a disadvantageous location because metabolomic measures are affected by other factors that are not directly related to cancer, such as diet, medication, renal and hepatic function, inflammation, fasting status, and sample handling. Therefore, the distinction between malignancy and normalcy is not a very specific one when it comes to individual metabolites. The above mentioned complementarity is exemplified in the NSCLC study. In addition to proteins, the researchers measured amino acids and bile acids, and identified altered pathways related to amino acids and lipids that were associated with disease. The metabolites also changed by disease stage; this indicates that metabolic measurements might reflect biological changes that occur by disease stage that are not reflected by a single protein measurement [7]. The results do not suggest that metabolomics should be considered as a standalone set of diagnostic variables, but rather a layer of molecular information that can be added to the information derived from the proteome.

Pathway-level effects of malignant transformation can also be revealed using metabolic signatures. For Wang et al., pancreatic cancer involved a combination of metabolomic and transcriptomic information on oxidative stress and phospholipid metabolism. Together with larger transcriptomic cohorts, larger datasets of metabolic data from pancreatic tumor and non-tumor tissues were analyzed to obtain three molecular subtypes that were associated with differences in prognosis, immune characteristics, and tumor-microenvironment features [9]. Because of this, several components of the signature were further validated at the protein level, supporting the biological relevance of this signature. Such a signature is more of a prognosis than a screening test, but it indicates a way metabolic changes can be analyzed in the context of upstream molecular information instead of as a list of metabolites.

Complementarity Between Molecular Layers

The main reason for combining genomic, proteomic and metabolic (metabonomic) biomarkers is not just to have more variables available. Different biological states are represented by the different layers so there is only partial overlap in information provided. Each of these genomic features represents relatively stable changes in the altered genome of the cancer; proteomic measurements measure functional cellular and extracellular processes; and metabolic measurements measures the biochemical consequences of these processes. It's not a one-to-one correspondence. The relationship between molecular layers can be interrupted and/or modified by epigenetic regulation, post-transcriptional control, protein degradation, post-translational modification, cellular composition and environmental factors. This is why the information gained from genomic, proteomic and metabolic measurements might not be redundant. This is consistent with the evidence provided by integrated clinical models. For NSCLC, Xu et al. [7] were not just adding random molecules found in the plasma proteome, but rather proteins found in the plasma proteome were used to select metabolite candidates, and features identified by these metabolites were then combined with traditional tumor markers. The final model included variables from the different levels of biology and performed better at discriminating than models based on individual molecular categories.



This is also true of the pancreatic cancer Molecular Twin platform by Osipov et al. [10]. The study was conducted on stage I/II pancreatic ductal adenocarcinoma (PDAC) samples by applying multi-omics approaches, including genome and transcriptome sequencing, tissue and plasma proteomics, plasma lipidomics and computational pathology. The resulting 6,363 features were provided to several machine-learning models. Integration did not necessarily require keeping all of the molecular feature space, as the complete multi-omic representation gave the best prediction of disease status, with a significantly reduced model that kept only 589 features keeping similar predictive performance. This is the key to the development of a biomarker-panel. The more predictive a feature is in one molecular layer, the less extra information it may provide when correlated features of another layer are present. On the other hand, a not too highly discriminative metabolite or protein could prove useful if it reflects biological variation not measured in genome-based studies. The goal of multi-biomarker screening is not, therefore, to measure as many molecules as possible, but to determine a short list of complementary features that together define the biology of malignancy with just the right sensitivity and specificity to be clinically useful.

Biological Characteristics and Diagnostic Contributions of Genomic, Proteomic, and Metabolic Biomarkers in Cancer Screening

Molecular Layer	Representative Biomarkers	Biological Information Captured	Potential Diagnostic Contribution	Major Limitations	Role in Multi-Omics Integration
Genomic	Somatic mutations, copy-number alterations, DNA methylation, circulating tumor DNA (ctDNA)	Genetic and epigenetic alterations associated with malignant transformation	Detection of tumor-associated molecular alterations and potential cancer-origin classification	Low tumor-derived DNA abundance in early-stage disease, sequencing errors, and inter-tumor heterogeneity	Provides molecular evidence of malignancy and may help explain downstream protein and metabolic alterations
Proteomic	Circulating proteins, signaling proteins, acute-phase proteins, extracellular-vesicle proteins	Functional cellular activity, signaling pathways, secretion, and host-response mechanisms	Identification of tumor-associated protein patterns and systemic responses to malignancy	Inflammatory confounding, protein abundance variability, platform-dependent measurements, and limited tumor specificity	Captures functional changes that may not be directly reflected by genomic measurements
Metabolic	Amino acids, lipids, bile acids, organic acids, glucose-related metabolites	Biochemical activity, nutrient utilization, energy metabolism, and metabolic reprogramming	Identification of altered metabolic states associated with malignant transformation	Sensitivity to diet, medication, fasting status, organ function, and sample handling	Provides downstream biochemical information that may complement genomic and proteomic signals
Integrated multi-omics	Combinations of genomic, proteomic, and metabolic features	Partially complementary molecular alterations across multiple biological levels	Potential improvement in discrimination, molecular characterization, and diagnostic confidence	High dimensionality, missing modalities, batch effects, cost, and overfitting	Enables joint modeling of complementary signals while supporting feature reduction and biomarker-panel construction

Machine Learning Strategies for Multi-Omics Integration

Data Preprocessing and Harmonization

The integration of genomic, proteomic and metabolic data starts with a preprocessing problem which is significantly more complex than the typical single-omics classification. The nature of features, measurements, missing features, noise characteristics, and biological interpretations vary across each of the molecular layers. Genomic datasets may include binary mutation indicators, copy-number measurements, or continuous methylation values while proteomic and metabolomic datasets are often abundance measurements that have very different distributions. Features of one modality may dominate the learning process if these features are directly used together, just because of the scale of the numbers or dimensionality of the features. Thus, in general, the two main preprocessing steps are QC, transformation, normalization, filtering of features, and scaling before integration. Logarithmic or variance-stabilizing transformations can stabilize abundance distributions that are highly skewed; standardized or robust-scaled representations can normalize

different measurements of molecules to be on similar scales. Samples collected from different laboratories, sequencing centres, experimental runs or analytical platforms at the same time are especially suitable for batch correction. Importantly, any processing operation that involves estimating parameters based on the full data set should be carried out on the training data set while developing the model. If not, data from the validation/training set could be incorporated into the learning process, leading to potentially optimistic estimates of predictive performance. This is more challenging when patients do not have measurements for all molecular layers. Often, multi-omics experiments do not have isolated missing values but missing blocks, for example, the genomic and transcriptomic data may exist for one patient and the proteomic measurements may be missing. Picard et al. highlighted the fact that the two datasets within a multi-omics study may have different dimensions and lack complete observations, and presented existing solutions and methods, including: dataset integration before transformation, transformation of each individual dataset followed by integration, and integration of each individual dataset at prediction level. Therefore it is not only a technical pre-processing step to perform the harmonization [11]. This decides the difference between learning biological relations and technical differences between molecular platforms. It is important to harmonize in clinical screening applications owing to the fact that the samples may be collected at different times, and the conditions under which they are analyzed may vary from those in which the model was constructed.

- *Dimensionality Reduction and Latent Representation*

A major statistical challenge with the multi-omics datasets is the dimensionality, that is, the number of molecular variables can be very large compared with the number of patients available. There may be thousands of genes, methylation sites, proteins, and/or metabolites in a single cancer dataset and only a few hundred individuals that have been clinically characterized. All of the variables can then be concatenated to create a very high-dimensional feature matrix, which can suffer from overfitting, multicollinearity and unstable feature selection.

The goal of dimensionality reduction is to reduce the number of measurements in the original molecular data set without losing information that is useful for the biological or predictive problem of interest. The classical techniques are principal component analysis and related matrix factorization techniques, while the newer methods are autoencoders and neural embedding architectures. The goal is not just to save computation time but to build models that are able to describe correlated molecular measurements in the latent space which can be later fed to classifiers. In order to solve the problem of a high dimensional of variables compared with the samples, Chai et al. proposed a denoising autoencoder deep-learning framework for multi-omics data. They first learned representations from multiple molecular layers, and then applied these representations to the prediction of cancer prognosis. The authors also independently evaluated a simplified model using mRNA data alone, and found that a simplified model could sometimes capture much of the predictive power of a complex integrated model, as demonstrated in three external Gene Expression Omnibus (GEO) datasets, which illustrates an important principle of multi-omics modelling: an integrated model can reveal useful biological structure, but a simplified one can sometimes retain much of this predictive information [12].

In OmiEmbed, Zhang et al. proposed a similar approach which mapped high-dimensional measurements of omics to low-dimensional embeddings before applying downstream tasks

[13]. The framework enabled dimensionality reduction, tumor-type classification, multi-omics integration, clinical-feature reconstruction, and survival prediction. It was multi-task architecture and it enabled the shared representation to be learned from multiple related objectives instead of a single prediction task, and the authors report that performance from multi-task training is superior to training the downstream tasks independently. When some of the measurements made on different molecular layers are partially redundant, latent representations can be very useful. The embedding can represent the global states of the molecules rather than individual genes and proteins. But dimensionality reduction comes with a "lossy" interpretability vs compression cost. A latent component could be very predictive, but not be directly linked to a known biological pathway or clinically measurable biomarker. Per diagnosis, therefore, representation should also preferably be accompanied by the tools that enable important latent dimensions to be traced back to individual genes, proteins, metabolites, or pathways.

Feature Selection and Biomarker Panel Construction

There are two problems in dimensionality reduction and feature selection: they are related but not the same. Dimensionality reduction aims to transform the original variables to their representation, while feature selection aims to select a subset of the original molecular measurements. The latter has special significance for clinical translation as a model involving a small number of measurable biomarkers is more easily assayable, interpretable, standardizable and validated than one that relies on thousands of molecular variables. Feature selection can be done independently on each omics layer or along with training the model. The traditional methods are univariate filtering, correlation-based filtering, recursive feature elimination, regularized regression and the importance measures provided by trees. Penalized regression is particularly useful when there are many more predictors than observations, since regularization will limit the complexity of the model, and can help minimize the impact of the variables that are less stable. Ma et al. [14] presented an important application of feature selection in multi-omics classification, where they were able to differentiate early- and late-stage cancers using TCGA datasets including data from mRNA expression, DNA methylation, and copy-number variation data using an XGBoost model. They tested different combinations of molecular data (increasingly complex) and a number of machine-learning approaches. The measurement of DNA methylation appeared to be more accurate and stable for the task of stage discrimination than the measurement of mRNA and miRNA; integrating multiple molecular layers by using an autoencoder increased the accuracy of the classification. Investigators also ranked genes by the XGBoost model and then analyzed the importance of each gene in terms of its biological function.

In this study it is shown how and why biological interpretation of a model should be included in the process of feature selection and not only as a mathematical optimization step. Selection of a feature consistently across multiple independent data sets and to a biological coherent pathway will offer more evidence for relevance to biomarker than a single selection based on statistical association in one training cohort. One of the key differences to consider in multi-omics research is between predictive importance and biomarker utility. A feature may be correlated with another molecular variable, but contribute little to the classification, and may improve the classification. On the other hand, a moderately predictive feature on another omics layer can add valuable complementary information as it captures biological processes that are not reflected in the dominant modality. Therefore, the process of

feature selection should not only be based on the association of each individual feature with the disease status, but also on its incremental predictive contribution, biological coherence and stability.

Early, Intermediate and Late Data Fusion

The key integration challenge after normalizing and representing the individual molecular datasets is where to integrate them. There are generally three different approaches, early fusion, intermediate fusion, and late fusion, although methodological reviews have suggested finer-grained distinctions, such as mixed and hierarchical integration.

In early fusion, features from multiple molecular layers are concatenated into a single matrix before model training:

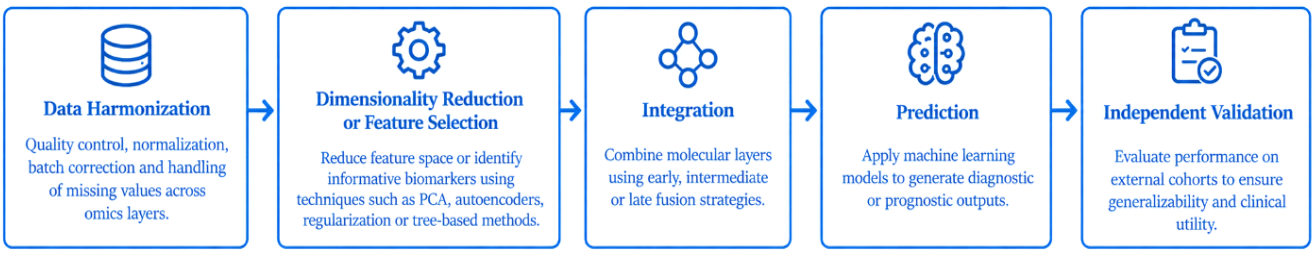
.This combined representation is then used to train a single classifier. The most important benefit is its simplicity: it is able to directly learn interactions between variables coming from different layers of the molecules. Concatenation's downside is that it can produce a very large and diverse feature space. One omics dataset may contain much more of a number of variables than another, and thus might have a disproportionate effect on the model. To resolve this issue, intermediate fusion transforms each molecular layer into a lower dimensional representation, then it fuses these representations. These representation-learning algorithms can then be used to generate representations for each modality, prior to a shared classifier or prediction layer. This can capture modality-specific information, whilst enabling the model to learn relationships between the modality-specific representations. In the case of late fusion, models are kept as separate models for each omics layer. A prediction, probability or latent decision representation is created by each model, and the models combine their outputs. It can be helpful when the molecular layers are very different among the samples or when the molecular layers are absent from a specific sample. But it might not be able to capture detailed molecular interaction since the modalities has been processed independently prior to prediction.

Picard et al. [11] detailed five different integration paradigms, with the early and the mixed ones designed to maintain the modality-specific structure of the data, while the intermediate, late and hierarchical ones focused on leveraging cross-omic relationships. In the context of genomic–proteomic–metabolic integration, this framework is especially applicable, as the dimensionality and measurement properties of the three layers vary greatly. The methodological literature also shows that there is no optimum fusion strategy in general. Cai et al. carried out a review of general-purpose and task-specific approaches to cancer multi-omics integration, and a comparison and evaluation of five methods with the Cancer Cell Line Encyclopedia data for cancer-type classification and drug-response prediction [15]. They found that the efficiency of the algorithm varied based on the problem that was set downstream, and did not have a universal sequence of how well the methods worked. In terms of biomarker development for diagnosis, this translates to a need to choose the fusion strategy based on the clinical question, sample structure, missingness and interpretability desired, and not the apparent complexity of the algorithm.

Ensemble and Deep Learning Approaches

Machine-learning architectures have evolved from classic classifiers to representation-learning architectures, ensemble models and graph-based models as the number of multi-omics datasets has grown increasingly diverse. Traditional algorithms like logistic regression, support vector machines, random forests and gradient boosting will still be helpful for many reasons and can be used following appropriate feature selection, and offer reasonably interpretable estimates of feature importance. Their main drawback is that they require a lot of

pre-processing when modelling interactions among thousands of molecular variables. Tree-based ensemble methods allows for modeling of non-linear relationship and interactions without explicit modeling of the relationship. The goal of identifying informative molecules features along with the classification is a typical example of this approach, as illustrated by the XGBoost framework by Ma et al. [16]. A solution using deep learning methods is to learn nonlinear representations directly from the high-dimensional molecular measurements. Each omics layer can be encoded into latent variables using an Autoencoder, and Denoising Autoencoders can further promote representation of the data that is robust to noise in the input data. OmiEmbed furthered this idea by adding multiple prediction tasks as downstream components and leveraging deep embeddings to share information that is useful for several related tasks [13]. More modern architectures have started to emulate the relational model rather than tabular data with multi-omics datasets. To tackle these challenges, Ouyang et al. [17] proposed an end-to-end framework called MOGLAM, which integrates multi-omics attention, feature selection, dynamic graph convolution, and integrated representation learning. Rather than assuming that a constant similarity structure between patients is the correct one, the model learned the relationships between the patients within each layer of omics samples adaptively. The attention mechanism then acted as a weight gate to selectively distribute the learned representations, and the integrated representation module tried to learn common and complementary information between the molecular layers. The authors tested the framework on three different cancer types: kidney, squamous-cell carcinoma (pan-cancer), and breast invasive carcinoma (subtype) cancer, and showed that it outperforms other multi-omics methods used for cancer classification in the article. However, the more complex the model, the better the biomarker system is not necessarily. When several omics matrices are aggregated, deep neural networks can have far more parameters than are supported by the number of patient samples, especially when dealing with high dimensional matrices. The models that emerge could be specific for the cohort noise, batch properties or clinical covariates that are not captured. Therefore, it is necessary to have a sequence of up-scaling steps as described in Fig. 1 that is required for the development of multi-omics diagnostic models, not just simply growing the architecture.



Progression of multi-omics diagnostic models

Cancer-Specific Applications and Clinical Evaluation of Integrated Biomarker Models

Multi-Biomarker Screening Across Cancer Types

The goal of integrating molecular biomarkers into clinical practice is to diagnose malignancy in the early stages without invoking an unnecessary diagnostic process and with high specificity. The initial blood-based methods have integrated genomic changes from the tumor with measurements of protein levels in the blood and laid the groundwork for future multi-omics screening strategies. Cohen et al. have created a blood-based test that identifies mutations in 16 cancer genes and eight proteins in the blood, called the CancerSEEK. A total of 1,005 patients with clinically diagnosed, non-metastatic cancers of the ovary, liver, stomach, pancreas, esophagus, colorectum, lung and breast and 812 healthy controls were evaluated. A median of 70% of the eight types of cancer were detected, ranging from 69% to 98% sensitivity, for five cancers with no established population screening tests. Specificity was >99% (7 healthy controls were positive). In addition, the investigators have created a machine-learning algorithm to determine the anatomical location of cancers they detected, though the accuracy of the localization was dependent on the type of cancer [18].

CancerSEEK is especially important because it illustrates the significant promise of using both genomic and proteomic signals, instead of one type of biomarker. However it had patients with established cancers, not the unselected population of people being screened for the presence of cancer. As a result, the reported sensitivity and specificity are not necessarily used to conclude the effectiveness of population level screening. Putcha et al. conducted a prospective study of a multi-omics blood test for diagnosis of CRC in a second study. The platform integrated whole genome sequencing, bisulfite sequencing, and protein measurements to enable the capture of signals from the tumor and the host response. The investigators reported a mean sensitivity of 92% and specificity of 90% at detecting early-stage colorectal adenocarcinoma in 591 participants (43 with colorectal adenocarcinoma and 548 with no colorectal cancer who had been diagnosed by colonoscopy). In all pathological subtypes, the sensitivity was 80% early. The authors also found that this was less effective than a single assay test when tested individually [19]. The following results are for an example complementary molecular signals contributing to a screening oriented classifier. However, the number of subjects with cancer was limited and the report was a conference abstract of interim results. As such, the findings must be viewed as preliminary and not as clinical validation.

This study aims to investigate the feasibility of using molecular markers for the selection of patients and for stage classification of colorectal cancer.

The next example of linking biomarker selection with diagnostic classification using a machine learning approach is colorectal cancer. Battista et al. created the BH-index, an ensemble-learning system, based on the 17 circulating protein markers measured in 345 individuals with 248 patients having colorectal neoplasia. The panel comprised of carcinoembryonic antigen, CA 19-9, C-reactive protein, ceruloplasmin, haptoglobin, transferrin, ferritin, complement proteins, and other serum analytes

Cancer: Biomarker Selection and Stage Classification

The investigators created two separate predictors: one for disease presence and one for TNM stage. The ceruloplasmin and α-2-macroglobulin were informative variables for the early CRC diagnosis for XGBoost and random forest models; the

contribution of CA 50 and α-1-antitrypsin to the early diagnosis was observed to be less. A study on using a majority-voting ensemble to decrease prediction variability was also undertaken [20].This investigation has not incorporated genomic and metabolomic data, but it illustrates an important translational rule: A diagnostic panel should be considered on the basis of the contribution of each of its biomarkers, rather than the number of analytes included. The serum protein panel also provides an example of how easily accessible assays can complement the more expensive assays of molecular assays. In 2022, Su et al. analyzed colorectal cancer classification using the gene-expression profiles from the TCGA. They analysed the data using weighted gene co-expression network analysis, feature selection by least absolute shrinkage and selection operator, and random forest, support vector machine, and decision-tree classifiers. Random forest gave an average cross-validation accuracy of 99.81% for cancer versus control classification. The reported average accuracy was 91.5% for stages I–IV, and was lower for the classification of the stages than for the binary diagnostic task. The authors also noted that they were able to identify candidate prognostic genes, such as GCNT2, GLDN, SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5 and CPT2 [21]. Given that this study involved transcriptomic data, rather than multi-omics data, the results can be considered as complementary evidence for feature selection and classification. It is also useful to note that the high level of internal accuracy further demonstrates that results derived from a cross-validation on a molecular dataset is not enough to prove readiness for clinical screening.

Ovarian Cancer: Probabilistic Metabolomic Diagnosis

Ovarian cancer is especially challenging to diagnose due to the fact that most of the time, there are not any particular signs of the disease and existing population screening methods are limited. Ban et al. studied the machine-learning-based diagnosis based on serum metabolomic profiles of 431 ovarian cancer patients from four geographic locations, together with 133 controls. Four analytical datasets obtained from the four different chromatographic and ionization conditions were used to analyze serum samples by ultra-performance liquid chromatography and high resolution mass spectrometry. A recursive feature elimination with repeated cross-validation (RFECV) algorithm was employed to select informative metabolic features. A consensus prediction system consisting of five classifiers: logistic regression, random forest, support vector machine, k-nearest neighbors and adaptive boosting were formed. The final model was able to detect if the sample was cancer or healthy and it was accurate to 93%. In addition to simply determining a binary classification, the investigators took the distribution of individual model probabilities into account to create a probabilistic diagnostic framework. The majority of the metabolites that were chemically annotated were lipids, and only a small fraction of the most informative metabolites were chemically annotated [22].

This study was a metabolomics-based study and not a genomic–proteomic–metabolic integration study, but is applicable to the clinical evaluation of molecular classifiers. Positive or negative results may be better used in combination with individual probability estimates to better inform risk-based diagnostic decisions. But the cited accuracy should not be taken as the sensitivity or positive predictive value of population screening, as the case-control design is different from the prevalence in routine screening.

Diagnostic Performance, Calibration and Clinical Generalizability

In all of these investigations, the performance of the

models is reported, depending on the type of cancer, the molecular platform, the comparator population, and the design of the validation. The measures of accuracy and area under the receiver operating characteristic curve are good indicators of discrimination but do not determine clinical reliability of a model to provide risk estimates. A classifier may have a good ability to differentiate normal controls from established cancer cases, but have far less ability when compared to patients presenting with benign inflammatory, metabolic and organ specific disease. Care is especially needed to ensure specificity in screening at the population level. If the prevalence of disease is low, a percentage of false-positive results will produce a significant number of cases that will need further imaging, endoscopy, biopsy, or specialist evaluation. Therefore, the positive predictive value needs to be applied to the group (population) being screened, not to be inferred from an artificially balanced case control set. A complementary measure is carried out in calibration, which involves looking for consistency between the frequencies of the diseases observed and the probabilities predicted. A model with a 20% risk of malignancy should detect cancer in about 20% of people within an appropriate group that has been calibrated to the model correctly. Instead of using traditional discrimination measures, it is important to assess the calibration of the model, model sensitivity at a predetermined level of specificity, and assess the model using decision analytic tools. It is also crucial to also get external validation. Random splitting of dataset patients to training and testing sets is not sufficient to ensure generalizability if both sets are recruited and tested on the same platform, have similar demographics, and have the same distributions of disease spectra. Ideally, independent validation should include another institution, geographic population, analysis batch and clinically relevant control population. Lastly, a molecular integration's incremental value needs to be proven in relation to appropriate comparators. Individual microarray, metabolomics, or proteomics classifiers, as well as traditional clinical parameters and diagnostic biomarkers, should be compared in a multi-omics classifier. If a reduced protein or metabolite panel does well in comparison to a large integrated panel, then the reduced panel may provide practical benefits in terms of analytical cost, complexity of the assays and reproducibility. The evidence presented here clearly indicates the possibility of molecular classifiers that are integrated but also shows that diagnostic performance, clinical calibration, and generalizability to new external populations are distinct needs. Multi-biomarker models can only be considered as potential for routine cancer screening if they can be evaluated in combination.

Conclusion and Research Directions

The integration of genomic, proteomic, and metabolic biomarkers together, in a machine-learning framework, allows for the analysis of the molecular changes associated with cancer on various levels of biological organization. The genomic measurements are associated with changes linked to malignant transformation; the proteomic profiles include changes in function and signaling, and the metabolomic measurements describes the biochemical changes associated with tumor development and reaction to the tumor by the host. Together, their analysis provides a chance to discover diagnostic patterns, which might otherwise not be adequately represented by individual molecular layers. The evidence presented in this review, however, suggests that the more molecular measurements that are added, the less the specificity of the diagnosis or the usefulness of the test for the clinical situation. The analysis power of multi-omics studies has been greatly boosted by machine-learning techniques such as regularized regression, ensemble classifiers, latent representation learning, and attention-based

integration. However, there are significant challenges in this characterization of cancer signals versus the physiological ones, reducing dimensionality, assuring analytical reproducibility, and avoiding overfitting. Moreover, the diagnostic discrimination is not enough to conclude the clinical usefulness. In addition, false-positive and false-negative results and their consequences must be considered, as well as the predictive value and the calibration.

Further studies are needed to look into the molecular changes prior to diagnosis and during the course of the disease from a longitudinal multi-omic profiling perspective. Frequent sampling can aid in the identification of persistent cancer-associated alterations from transient metabolic, inflammatory and/or treatment-induced alterations. Collaboration with clinical parameters, imaging studies and known laboratory markers also should be explored to determine the value of incremental diagnostic information given molecular measurements. Another key area of development is the creation of small, easily understood biomarker panels. Future models should instead be able to define combinations of genomic, protein, and metabolic markers that would retain the diagnostic utility with fewer molecular features and lower assay complexity and costs. Biological pathway analysis, evaluation of biomarker stability and independent replication may be useful to determine if chosen biomarkers are associated with disease or statistical associations among the parameters in a cohort. Before use of integrated biomarker systems for population screening could be suggested, they must be prospective, multicenter, and validated in asymptomatic population before. Benign disease controls should be included in such investigations, as should be prespecified diagnostic thresholds, calibration analysis and downstream diagnostic procedures. To ensure reproducibility and clinical feasibility, standardization of analytical platforms, demographic population and laboratories is also required. Machine-learning screening of multiple biomarkers is a good research avenue but not a proven alternative to traditional cancer screening routes. Whether or not it will be clinically useful will depend on the ability to show that molecular integration leads to reproducible and clinically meaningful, biologically interpretable changes in the early detection. For the shift from computational classification to a robust cancer screening tool, then, there is a need for simultaneous progress in the standardization of molecular assays, the validation of models, prospective clinical study, and implementation science.

REFERENCES

1. Smelik, M., Zhao, Y., Mansour Aly, D., Mahmud, A. F., Sysoev, O., Li, X., & Benson, M. (2024). Multiomics biomarkers were not superior to clinical variables for pan-cancer screening. Communications Medicine, 4(1), 234.
2. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., ... & Krishna, R. (2024, September). Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision (pp. 148-166). Cham: Springer Nature Switzerland.
3. Hu, J., Wang, S. G., Hou, Y., Chen, Z., Liu, L., Li, R., ... & Chen, K. (2024). Multi-omic profiling of clear cell renal cell carcinoma identifies metabolic reprogramming associated with disease progression. Nature genetics, 56(3), 442-457.
4. Sharma, A., Debik, J., Naume, B., Ohnstad, H. O., Bathen, T. F., & Giskeødegård, G. F. (2024). Comprehensive multi-omics analysis of breast

cancer reveals distinct long-term prognostic subtypes. Oncogenesis, 13(1), 22.

5. Jiang, Y. Z., Ma, D., Jin, X., Xiao, Y., Yu, Y., Shi, J., ... & Shao, Z. M. (2024). Integrated multiomic profiling of breast cancer in the Chinese population reveals patient stratification and therapeutic vulnerabilities. Nature cancer, 5(4), 673-690.
6. Xiong, D., Han, T., Li, Y., Hong, Y., Li, S., Li, X., ... & Li, C. (2024). TOTEM: a multi-cancer detection and localization approach using circulating tumor DNA methylation markers. BMC cancer, 24(1), 840.
7. Xu, R., Wang, J., Zhu, Q., Zou, C., Wei, Z., Wang, H., ... & Zhang, S. (2023). Integrated models of blood protein and metabolite enhance the diagnostic accuracy for Non-Small Cell Lung Cancer. Biomarker Research, 11(1), 71.
8. Xing, X., Hu, E., Ouyang, J., Zhong, X., Wang, F., Liu, K., ... & Liu, X. (2023). Integrated omics landscape of hepatocellular carcinoma suggests proteomic subtypes for precision therapy. Cell reports medicine, 4(12).
9. Wang, H., Guo, H., Sun, J., & Wang, Y. (2023). Multi-omics analyses based on genes associated with oxidative stress and phospholipid metabolism revealed the intrinsic molecular characteristics of pancreatic cancer. Scientific Reports, 13(1), 13564.
10. Osipov, A., Nikolic, O., Gertych, A., Parker, S., Hendifar, A., Singh, P., ... & Theodorescu, D. (2024). The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. Nature cancer, 5(2), 299-314.
11. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis.
12. Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., & Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Computers in biology and medicine, 134, 104481.
13. Zhang, X., Xing, Y., Sun, K., & Guo, Y. (2021). OmiEmbed: a unified multi-task deep learning framework for multi-omics data. Cancers, 13(12), 3047.
14. Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine, 121, 103761.
15. Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. Iscience, 25(2).
16. Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine, 121, 103761.
17. Ouyang, D., Liang, Y., Li, L., Ai, N., Lu, S., Yu, M., ... & Xie, S. (2023). Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Computers in biology and medicine, 164, 107303.
18. Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., ... & Papadopoulos, N. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 359(6378), 926-930.
19. Putcha, G., Liu, T. Y., Ariazi, E., Bertin, M., Drake, A., Dzamba, M., ... & Sansanwal, P. (2020). Blood-based detection of early-stage colorectal cancer using multiomics and machine learning. J Clin Oncol, 38(suppl 4), 66.

20. Battista, A., Battista, R. A., Battista, F., Iovane, G., & Landi, R. E. (2021). BH-index: A predictive system based on serum biomarkers and ensemble learning for early colorectal cancer diagnosis in mass screening. Computer Methods and Programs in Biomedicine, 212, 106494.

21. Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., ... & Lv, X. (2022). Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Computers in biology and medicine, 145, 105409.

22. Ban, D., Housley, S. N., Matyunina, L. V., McDonald, L. D., Bae-Jump, V. L., Benigno, B. B., ... & McDonald, J. F. (2024). A personalized probabilistic approach to ovarian cancer diagnostics. Gynecologic oncology, 182, 168-175.