


ORIGINAL ARTICLE OPEN ACCESS

# MOGAN for LUAD Subtype Classification by Integrating Three Omics Data Types

Haibin He<sup>1</sup> | Longxing Wang<sup>1</sup> | Mingyue Ma<sup>1,2</sup> <sup>1</sup>Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China | <sup>2</sup>Institute of Life Sciences, Chongqing Medical University, Chongqing, China**Correspondence:** Mingyue Ma ([mamingyue@cqupt.edu.cn](mailto:mamingyue@cqupt.edu.cn))**Received:** 31 May 2024 | **Revised:** 21 October 2024 | **Accepted:** 26 November 2024**Funding:** Chongqing Municipal Human Resources and Social Security Bureau (Grant No: 2023C0BSHTB1001); Science and Technology Research Program of Chongqing Municipal Education (Grant No: KJ202200678822935)**Keywords:** cancer subtype classification | deep learning | molecular subtype | multi-omics generative adversarial network | transcriptome-proteome-methylome association analysis

## ABSTRACT

**Background:** Lung adenocarcinoma (LUAD) is a highly heterogeneous cancer type with a poor prognosis. Accurate subtype identification can help guide its treatment. The traditional subtype identification methods using a single-omics approach make it difficult to comprehensively characterize the molecular features of LUAD. Identification of subtypes through multi-omics association strategies can effectively supplement the shortcomings of single-omics information.

**Methods:** In this study, we used the Generative Adversarial Network (GAN) to mine transcriptomic, proteomic, and epigenomic information and generate an integrated data set. The newly integrated data were then used to identify LUAD immune subtypes. In the improved GAN (MOGAN) method, we not only integrated multiple omics datasets but also included the interactions between proteins and genes and between methylation and genes. Thus, we achieved effective complementarity of multi-omics information.

**Results:** Two subtypes, MOGANTPM\_S1 and MOGANTPM\_S2, were identified using immune cell infiltration analysis and the integrated multi-omics data. MOGANTPM\_S1 patients displayed higher immune cell infiltration, better prognosis, and sensitivity to immune checkpoint inhibitors (ICIs), while MOGANTPM\_S2 had lower immune cell infiltration, poorer prognosis, and were insensitive to ICIs. Therefore, immunotherapy was more suitable for MOGANTPM\_S1 patients in clinical practice. In addition, this study developed a LUAD subtype diagnostic model using the transcriptomic and proteomic features of five genes, which can be used to guide clinical subtype diagnosis.

**Conclusions:** In summary, the MOGAN method was applied to integrate three omics data types and successfully identify two LUAD immune subtypes with significant survival differences. This classification method may be useful for LUAD treatment decisions.

**Abbreviations:** GAN, generative adversarial network; ICI, immune checkpoint inhibitor; IDI, integrated discrimination improvement; LUAD, lung adenocarcinoma; MOGAN, multi-omics-based GAN; NRI, net reclassification improvement; PCA, principal component analysis; TIIC, tumor-infiltrating immune cell; TMB, tumor mutation burden; TME, tumor microenvironment.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Cancer Innovation* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

## 1 | Introduction

According to 2022 global cancer statistics, lung cancer was the most frequently diagnosed cancer [1]. In China, the incidence of lung cancer ranks first among cancer types [2]. Lung adenocarcinoma (LUAD) is the main lung cancer subtype (about 40% of cases) [3], with high heterogeneity and poor prognosis [4]. Therefore, many researchers have been committed to identifying ways to classify LUAD better to guide precise diagnosis, treatment, and drug development accurately. Because of the high heterogeneity and complexity of LUAD, traditional histological classification relies on the subjective judgment of pathologists, limiting its utility in diagnosis [5]. Studies have shown that differences in molecular pathology have an important impact on the clinical characteristics of the disease [6]. Therefore, an increasing number of molecular subtypes are being investigated to guide LUAD treatment. For example, Shi et al. used gene methylation levels to classify 335 LUAD patients into seven subtypes, observing significant survival differences among the groups [7]. Wang et al. divided LUAD into high-risk and low-risk subtypes using immune-related genes that were significantly associated with prognosis. The two subtypes exhibited significant differences in survival rates, with the high-risk subtype having a higher tumor mutation burden (TMB) and lower TIDE scores than the low-risk subtype [8]. Zhang et al. clustered LUAD samples into IC1 and IC2 subtypes using CD8 + T cell-related genes. The IC1 subtype displayed greater sensitivity to immune checkpoint inhibitor (ICI) therapy, while the IC2 subtype exhibited a higher TMB and lower immune infiltration scores [9]. However, most LUAD classification studies use only single-omics data, leading to difficulties in comprehensively and systematically explaining the biological characteristics of LUAD.

The rapid development of biotechnology has enabled researchers to obtain multilayered information from genomes, epigenomes, transcriptomes, proteomes, metabolomes, and other molecular characteristics [10]. The integration of multi-omics data could more comprehensively reveal the molecular heterogeneity of tumors and promote precision treatment approaches. However, multi-omics studies usually involve complex relationships among multiple variables. Because of the high-dimensional nature of these data, such studies usually require a certain amount of computational resources and complex algorithms for analysis and interpretation. The most basic method of integrating multi-omics data is to directly merge the standardized data from two or more different omics types, such as the association analysis of microRNA (miRNA) expression levels and DNA methylation levels. However, the results when using this approach have not been ideal [11]. Another common strategy is to analyze each data type independently, and then merge them [12]. However, this approach often leads to unreliable conclusions and creates more challenges for integrating multi-omics information.

In recent years, deep learning has become an important algorithm for machine learning. This method uses neural networks composed of hidden layers to perform different operations to discover complex data representations [13]. For example, Ahmed et al. employed a Generative Adversarial Network (GAN) to integrate the gene expression matrix and miRNA expression matrix, leveraging the regulatory network information of miRNAs on genes [14]. Peng et al. considered the crucial

role of genes as bridges between miRNAs and diseases to construct an miRNA-disease association network. The authors introduced a gene layer between the miRNA layer and disease layers, forming a three-layer network that incorporates disease-gene associations and miRNA-gene associations to enhance the miRNA-disease association network [15]. Lee et al. proposed a deep learning-based autoencoder method that combined four omics data types, including mRNA, miRNA, DNA methylation, and copy number variation (CNV). The model learned representative features to distinguish two patient subtypes with significantly different survival rates and concordance indices [16].

Transcriptomics-based analysis is crucial for understanding the biological processes within cells and intercellular communication, providing essential information directly related to cellular phenotypes [17]. Furthermore, DNA methylation regulates gene expression by recruiting proteins involved in gene suppression or inhibiting transcription factor binding to DNA [18]. Proteomics complements other omics technologies, such as genomics and transcriptomics [19], and provides a more accurate reflection of the dynamic changes occurring at the cellular, tissue, or organism level [20]. Therefore, in this study, we aimed to use the GAN method to integrate transcriptomic, methylation, and proteomic data with interaction information among them. The newly generated fused data set was used for LUAD subtype clustering and to explore the heterogeneity among the different subtypes.

## 2 | Materials and Methods

### 2.1 | Omics Data Acquisition and Preprocessing

In this study, RNA sequencing (RNA-seq) gene expression data, DNA methylation data, and corresponding clinical information related to LUAD were obtained from The Cancer Genome Atlas (TCGA) database [21] (<https://portal.gdc.cancer.gov/>) using the UCSC Genome Browser [22]. The RNA-seq gene expression data consisted of 510 samples, with a data type of  $\log_2$  (FPKM+1). To reduce noise and errors and improve computational efficiency, this study removed any genes with zero expression across all samples. Considering the subsequent integration of transcriptomic and proteomic information, only protein-coding genes were included in this study. Finally, a total of 18,171 genes were retained for multi-omics data integration and subsequent analysis.

The DNA methylation data included 455 samples, with a data type of  $\beta$  values. Methylation sites with missing values were filtered out. To ensure the effective integration of methylation-gene association information, this study only selected the methylation sites that could be annotated to genes using the 450k methylation chip annotation file. These annotated methylation sites may be associated with the expression of annotated genes. Finally, this study obtained a  $\beta$  value matrix containing 455 LUAD samples and 365,860 methylation sites. For this methylation matrix,  $\log_2$  transformation was applied to emphasize the relative changes in methylation levels.

The protein expression matrix was obtained from The Cancer Proteome Atlas (TCPA) database [23] (<https://www.tcpaportal.org/>) at Level 3, which encompassed 362 samples.

By intersecting the three omics datasets, a total of 305 LUAD samples were obtained for the multi-omics information integration.

LUAD patient single-nucleotide polymorphism (SNP) and CNV data were also obtained from TCGA for mutation analysis. Three LUAD gene expression datasets, GSE30219 ( $n = 85$ ), GSE42127 ( $n = 133$ ), and GSE31210 ( $n = 226$ ), were obtained from the Gene Expression Omnibus (GEO) database [24] (<https://www.ncbi.nlm.nih.gov/geo/>) to validate the immune checkpoint blockade (ICB) response among the subtypes.

## 2.2 | Omics Interaction Network Construction

For the gene–protein interaction relationships, protein–protein interaction (PPI) networks were obtained from the STRING database [25] (<https://string-db.org/>) and the genome annotation files were obtained from the GENECODE database (<https://www.gencodegenes.org/>). This study derived the gene–protein interaction relationships by using the gene–protein correspondence and PPI relationships. An adjacency matrix was then constructed, with the gene–protein relationships with an interaction recorded as 1 and those without an interaction recorded as 0. This adjacency matrix included 9396 genes and 215 proteins.

For the methylation–gene interaction relationships, methylation–gene regulation scores were calculated using 450k methylation microarray annotation files obtained from the GEO database. Because methylation at different gene locations has varied effects, the gene length should be considered when calculating the methylation site–gene regulation score. Many regulatory elements are present near and upstream of the transcription start site (TSS). Therefore, methylation occurring in this region has a greater impact on the gene, leading to a higher methylation score. Methylation sites upstream of the gene are generally located in the promoter region (TSS200–TSS1500) [26].

For methylation sites occurring upstream of a gene, the methylation–gene regulation score was calculated using Formula (1):

$$\text{score} = 1 + \frac{|\text{methsite} - \text{start}|}{\text{width}} \quad (1)$$

For methylation sites occurring in the main gene sequence, the methylation–gene regulation score was calculated using Formula (2):

$$\text{score} = 1 - \frac{|\text{methsite} - \text{start}|}{\text{width}} \quad (2)$$

In this context, “width” represents the total length of the gene, “methsite” represents the methylation occurrence site, and “start” represents the TSS.

## 2.3 | Improving the GAN Model

In a previous study, Ahmed et al. constructed a GAN that could integrate two omics datasets and the interaction network

between them [14]. In this study, we improved this GAN by adding Generator 2 and Discriminator 2 modules, which enabled the integration of three omics data types. During GAN training, omics data 1 was input into Discriminator 1, and omics data 2 and the interaction relations between omics data 1 and 2 were input into Generator 1. Generator 1 generated synthetic data 1 with fusion information and passed it to Discriminator 2. Additionally, omics data 3 and the interaction relations between omics data 1 and 3 were input into Generator 2. Finally, through various iterations, Generator 2 provided a new omics data 1 matrix output with multiple omics information. Dropout layers were added to both the generator and discriminator to avoid overfitting, with a dropout rate of 0.3. The model training epoch was set to 10,000.

To address the issues of mode collapse and training difficulty in the GAN, this study adopted the WGAN architecture as the fundamental structure for both the generator and discriminator. Additionally, Wasserstein distance was employed to quantify the difference between the generated data and real data. The Wasserstein distance was computed using Formula (3):

$$W(P_r, P_g) = \inf_{r \sim \pi(P_r, P_g)} E_{(x,y) \sim \gamma} [\|x - y\|] \quad (3)$$

where  $P_r$  and  $P_g$  represent the probability distributions of the real data and generated data, respectively;  $x$  represents a real sample;  $y$  represents a generated sample;  $\pi(P_r, P_g)$  is the joint probability distribution of all  $(x, y)$  pairs.

The discriminator consisted of three fully connected layers, with a learning rate of  $5 \times 10^{-5}$ . For the discriminator, the goal was to maximize the score of real values and minimize the score of generated values. The loss function was as Formula (4):

$$L_D = D(G(x)) - D(x) \quad (4)$$

The generator consisted of four fully connected layers, with a learning rate of  $5 \times 10^{-6}$ . For the generator, the goal was to maximize the score of generated values. Therefore, the loss function was as Formula (5):

$$L_G = -D(G(x)) \quad (5)$$

In these equations,  $D$  represents the discriminator,  $G$  represents the generator,  $x$  represents the real value,  $G(x)$  represents the generated value, and  $D(x)$  represents the discriminator’s score on the original expression matrix.

## 2.4 | Multi-Omics Data Fusion and Subtype Clustering

The fusion of the three omics data types went through two GANs. First, the preprocessed RNA-seq gene expression matrix, protein expression matrix, and gene–protein interactions were input into the GAN. After five iterations of generation, a gene expression matrix containing proteomic and transcriptomic information was generated. Subsequently, the methylation signal value matrix, gene expression matrix integrating

proteomic-transcriptomic information, and methylation-gene regulation score matrix were input into the GAN again. After five iterations, the new gene expression matrix fusing the proteomic, transcriptomic, and epigenomic information was generated. To evaluate the reliability of the newly generated gene expression data, a K-nearest neighbors (KNN) classifier model was constructed using the EGFR mutation information of the patients [27] as labels, with the classifier performance indicator area under the receiver operating characteristic curve (AUC) values used to evaluate its fusion performance. Unsupervised clustering of LUAD samples was subsequently performed using RNA-seq gene expression data fused with multi-omics [28] to distinguish the different LUAD disease subtypes using 50 cycles to ensure the stability of the clusters.

## 2.5 | Evaluation of Subtype Identification Results

The  $-\log_{10} p$  and number of significant clinical parameters were used to evaluate the performance of different methods. We employed the *log-rank* test to assess if there were significant differences in survival among the various subtypes, then transformed the resulting  $p$  to  $-\log_{10}$ . A higher value of this metric indicated a more significant difference in survival rates among the subtypes.

Additionally, we examined the number of significant clinical parameters associated with each subtype. Seven clinical parameters were tested, including age (over 65 years old), gender, smoking status, overall stage, T stage, N stage, and M stage. The *chi-square* test was used to determine the associations between these parameters and the subtypes. *Fisher's exact* test was applied for parameters that did not meet the *chi-square* test conditions. Finally, we counted the number of clinical parameters that were significantly associated with each subtype.

## 2.6 | Statistical Analysis

All statistical analyses were performed using R software (<https://www.r-project.org/>, version 4.2.2). Differences between non-normally distributed variables were analyzed using the *Mann-Whitney U* test for two groups of continuous variables and with the *chi-square* test for noncontinuous variables. The R package *Survival* [29] was used to perform survival analysis,

Kaplan–Meier survival curves were generated to show differences in survival, and the *log-rank* test [30] was used to assess the significance of survival time differences between the two groups. All statistical  $p$  are bilateral and were considered statistically significant when  $p < 0.05$ , if not otherwise stated.

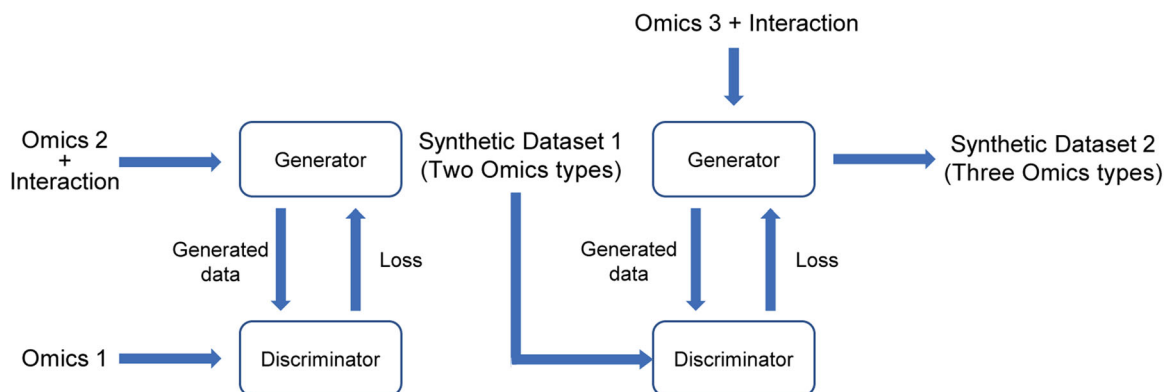
## 3 | Results

### 3.1 | Generation and Evaluation of Fused Data Using Multi-Omics-Based GAN (MOGAN)

The GAN developed by Ahmed and colleagues could only fuse two types of omics data. Here, we improved this GAN to allow it to fuse three types of omics data, aiming to complement the information from multi-omics analyses. The upgraded GAN structure, referred to as MOGAN, is shown in Figure 1. The omics datasets 1 and 2 were subjected to the first round of GAN fusion, resulting in the generation of a data set that follows the distribution of omics data set 1. Therefore, this data set could be considered as a pseudo-real omics data set 1. Subsequently, we performed the second round of GAN fusion between the pseudo-real omics data set 1 and omics data set 3. This process would generate a gene expression matrix that combines all three omics datasets. In the MOGAN, the gene expression matrix, protein expression matrix, and DNA methylation signal matrix were respectively incorporated into omics datasets 1, 2, and 3 (Figure 1). Additionally, the gene–protein interaction relationships and methylation-gene regulatory scores were input into the generator to facilitate the fusion of the three sets of omics information.

To evaluate the newly generated multi-omics fusion data, a KNN classifier was constructed to predict the *EGFR* mutation status of the patients. The performance of the generated data was assessed by comparing its AUC values with those of the original data in the classifier. The fusion data of the three omics (transcriptomic, proteomic, and epigenomic) had the highest AUC value compared with the original gene expression data and gene expression data fused with only the proteome or epigenome. The AUC results are presented in Figure 2, which also demonstrated that the multi-omics fusion gene expression data could more accurately characterize the molecular state of the disease.

While delving into the model prediction capabilities, the AUC value was a pivotal evaluation metric. However, when we

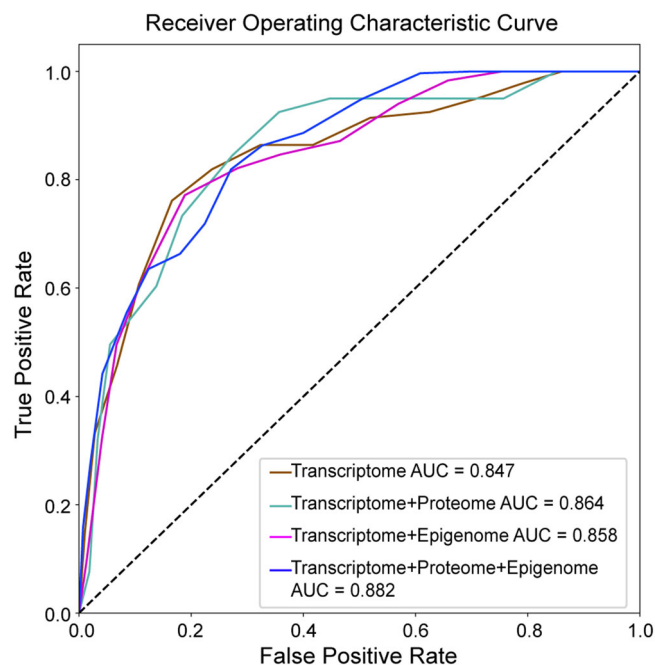


**FIGURE 1** | MOGAN architecture combining three types of omics data. MOGAN, multi-omics-based generative adversarial network.

focused on the nuanced improvements in model performance, especially when enhancing the models by introducing new variables or different datasets, net reclassification improvement (NRI) and integrated discrimination improvement (IDI) emerged as indispensable supplementary tools. These two indices displayed high sensitivity and intuitive interpretability in clinical settings. This provided a unique perspective for evaluating the magnitude of model improvements, thereby facilitating a comprehensive and profound understanding of the models from multiple dimensions. The calculations of the NRI and IDI values are shown in Table 1. When compared with the fusion of transcriptomic data with either proteomic or epigenomic data alone, the three omics fusion data exhibited significantly better performance for both NRI and IDI.

### 3.2 | Identification of LUAD Immune Subtypes Using the Fused Multi-Omics Data

The tumor microenvironment (TME) is closely related to tumor development, recurrence, and metastasis. Tumor-infiltrating immune cells (TIICs) are key immune biomarkers in the TME that play a crucial role in tumor growth and progression, patient prognosis, and immunotherapy response [31, 32]. This study



**FIGURE 2** | AUC analysis for integrating data from different sources. The receiver operating characteristic curves were compared for the original transcriptomic data and the fusion of different omics data using the GAN with EGFR classification. AUC, area under the receiver operating characteristic curve; GAN, generative adversarial network.

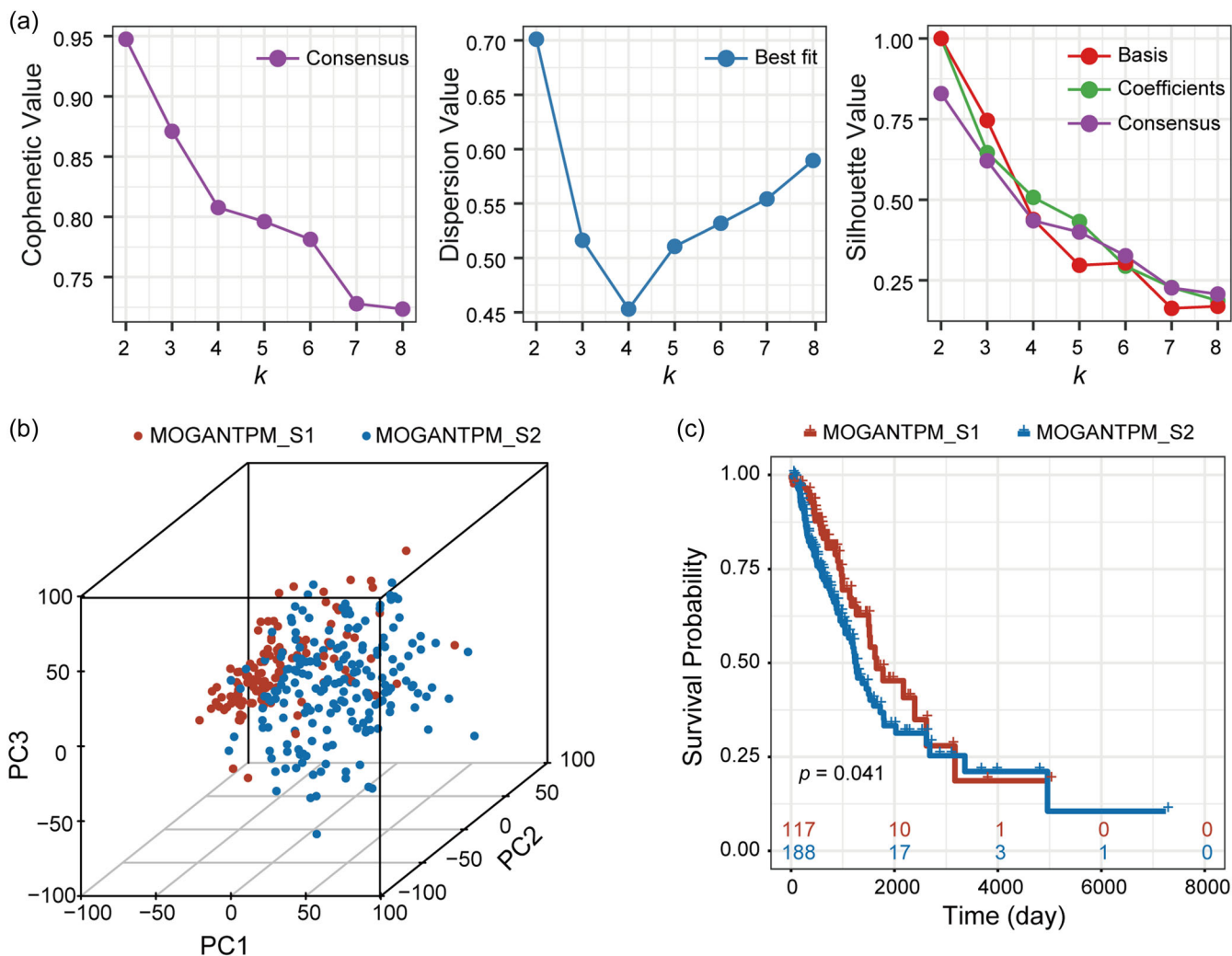
**TABLE 1** | Clinical characteristic differences in LUAD patients.

	Transcriptome+Proteome	Transcriptome+Epigenome	Transcriptome+Proteome+Epigenome
IDI	0.072	0.019	0.117
NRI	0.401	0.296	0.514

Abbreviations: IDI, integrated discrimination improvement; LUAD, lung adenocarcinoma; NRI, net reclassification improvement.

used a gene expression matrix integrating multi-omics information to identify immune subtypes in LUAD. The *ssGSEA* algorithm and immune gene sets were used to calculate the abundance of 28 TIIC types. Subsequently, the NMF clustering method was applied to the immune cell abundance matrix, resulting in the identification of two immune subtypes in LUAD: MOGANTPM\_S1 ( $n = 117$ ) and MOGANTPM\_S2 ( $n = 188$ ). The optimal number of cluster selections is shown in Figure 3a. The two LUAD subtypes can be clearly distinguished in the three-dimensional principal component analysis (PCA) scatter plot (Figure 3b). There was a significant difference in patient prognosis for these two subtypes, with the overall survival rate of MOGANTPM\_S1 patients being significantly better than that of MOGANTPM\_S2 patients (Figure 3c, *Log-rank* test,  $p = 0.041$ ).

Furthermore, to compare the clustering results of the original gene expression data with those of the threeomics fused gene expression data, we performed a similar analysis on the original gene expression matrix. This also identified two subtypes: T\_S1 and T\_S2. The PCA results for T\_S1 and T\_S2 are shown in Supporting Information S1: Figure S1. A Venn diagram illustrating the clustering differences between the two matrices is shown in Figure 4a. Among the 305 samples, the sample classification consistency rate between the original gene expression matrix and three omics data fusion matrix was 71.1% (217/305). The classifications of the remaining 88 samples were different. In the original gene expression matrix, 50 samples belonged to cluster T\_S1 but were classified as MOGANTPM\_S2 in the three omics data fused matrix (referred to as MOGANTPM\_S2-T\_S1). Similarly, there were 38 samples in the original gene expression matrix belonging to cluster T\_S2, but were classified as MOGANTPM\_S1 in the fused matrix (referred to as MOGANTPM\_S1-T\_S2, Supporting Information S1: Table S1). We believe that the proteome and methylome information impacted the clustering results, making the results of some original gene expression data inconsistent with those of the fusion data. Next, PCA dimensionality reduction and visualization were performed on the classification results of MOGANTPM\_S2-T\_S1, MOGANTPM\_S1-T\_S2, MOGANTPM\_S1-T\_S1, and MOGANTPM\_S2-T\_S2 with the protein expression matrix and methylation signal value matrix (Figure 4b,c). In the protein expression matrix clustering results, the discrimination between MOGANTPM\_S2-T\_S1 and MOGANTPM\_S1-T\_S1 was higher than that between MOGANTPM\_S2-T\_S1 and MOGANTPM\_S2-T\_S2. The discrimination between MOGANTPM\_S1-T\_S2 and MOGANTPM\_S1-T\_S1 was higher than that between MOGANTPM\_S1-T\_S2 and MOGANTPM\_S2-T\_S2 (Figure 4b). To some extent, this indicated the successful integration of the proteomic data. In the methylation signal value matrix PCA results (Figure 4c), the



**FIGURE 3** | Identification of LUAD molecular subtypes using the multi-omics fused data. (a) Relationships between the Coefficient, Dispersion, and Silhouette Coefficients and the number of clusters. (b) Principal component analysis of the MOGANTPM\_S1 and MOGANTPM\_S2 LUAD subtypes. (c) Kaplan–Meier survival curves showing the overall survival of patients in the MOGANTPM\_S1 and MOGANTPM\_S2 subtypes. LUAD, lung adenocarcinoma; MOGAN, multi-omics-based generative adversarial network.

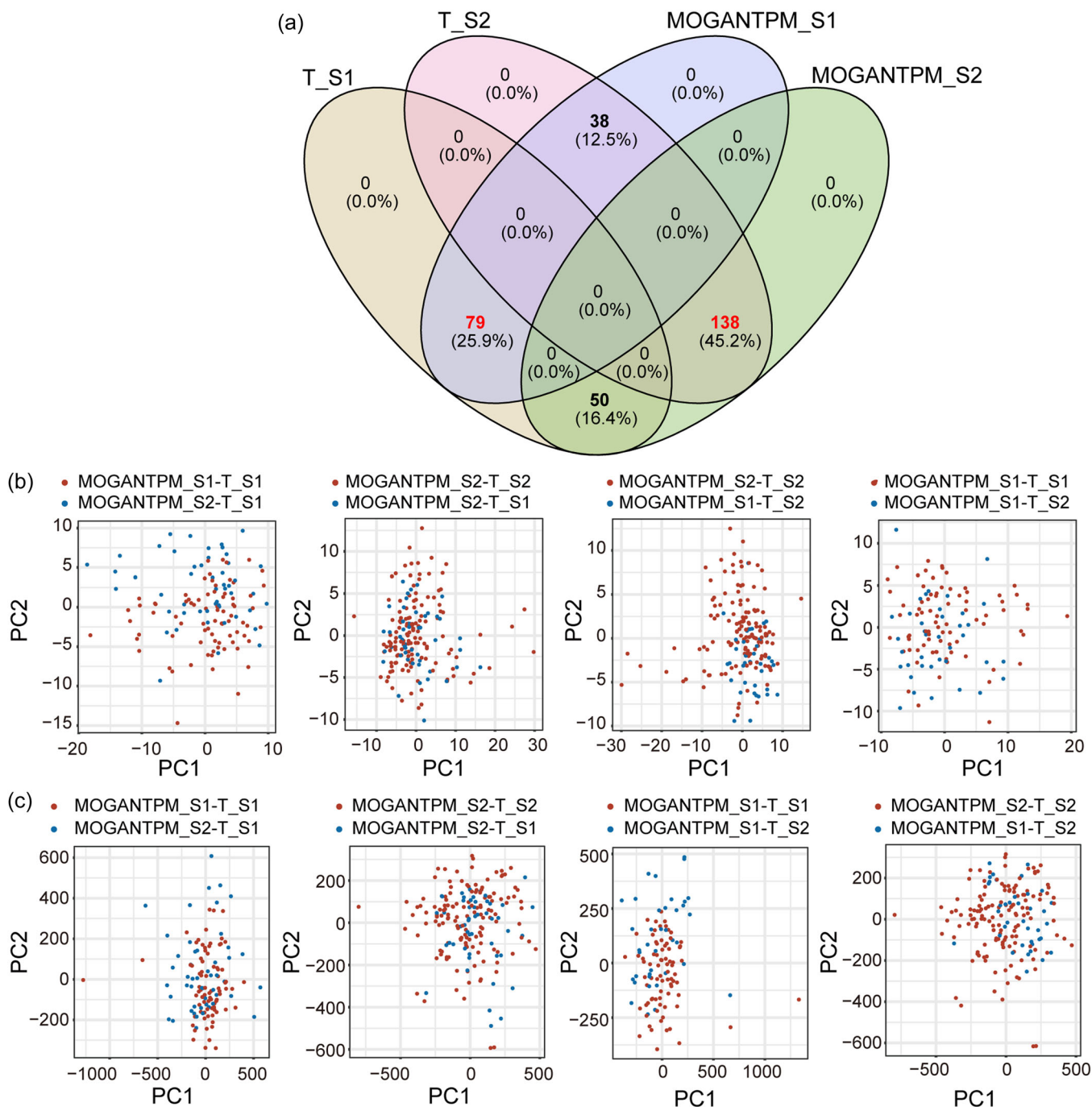
discrimination between MOGANTPM\_S2-T\_S1 and MOGANTPM\_S1-T\_S1 was higher than that between MOGANTPM\_S2-T\_S1 and MOGANTPM\_S2-T\_S2. The discrimination between MOGANTPM\_S1-T\_S2 and MOGANTPM\_S1-T\_S1 was higher than that between MOGANTPM\_S1-T\_S2 and MOGANTPM\_S2-T\_S2. Therefore, we speculated that the proteomic information played a more important role than methylation information in the multi-omics data integration.

### 3.3 | Benchmark Between the Different Algorithms

The robust cancer subtype clustering results are expected to be significantly correlated with patient prognosis and clinical characteristics. To assess our subtype identification results, our method was compared with seven previously published subtype identification methods (Figure 5). iClusterPlus [33], LRAcluster [34], moCluster [35], SNF [36], Subtype-DCC [37], Subtype-GAN [38], and Subtype-WELSR [39] were employed as benchmark algorithms

against MOGAN. Among them, iClusterPlus, LRAcluster, moCluster, and SNF have been commonly used for integrating multi-omics data to cluster cancer subtypes. Subtype-DCC, Subtype-GAN, and Subtype-WELSR are usually used for deep learning techniques. Subtype-WELSR adopted a weighted ensemble strategy, integrating multiple subtype clustering methods. Subtype-DCC achieved joint optimization of feature expression and clustering by introducing contrastive learning in both the sample space and clustering space. Subtype-GAN used an encoder to reduce the dimensionality of multi-omics data, a decoder to reconstruct multi-omics inputs, and a discriminator to ensure that the posterior distribution of the shared layer vectors conformed to a Gaussian distribution.

Here, we found that the  $-\log_{10} p$  of the MOGAN method was significantly superior to the iClusterPlus, LRAcluster, moCluster, SNF, and Subtype-WELSR methods, but inferior to the Subtype-DCC and Subtype-GAN methods (Figure 5a). For identifying clinically significant parameters associated with subtypes, MOGAN has four (the same as LRAcluster). This is less than the five for SNF and Subtype-WELSR, but more than

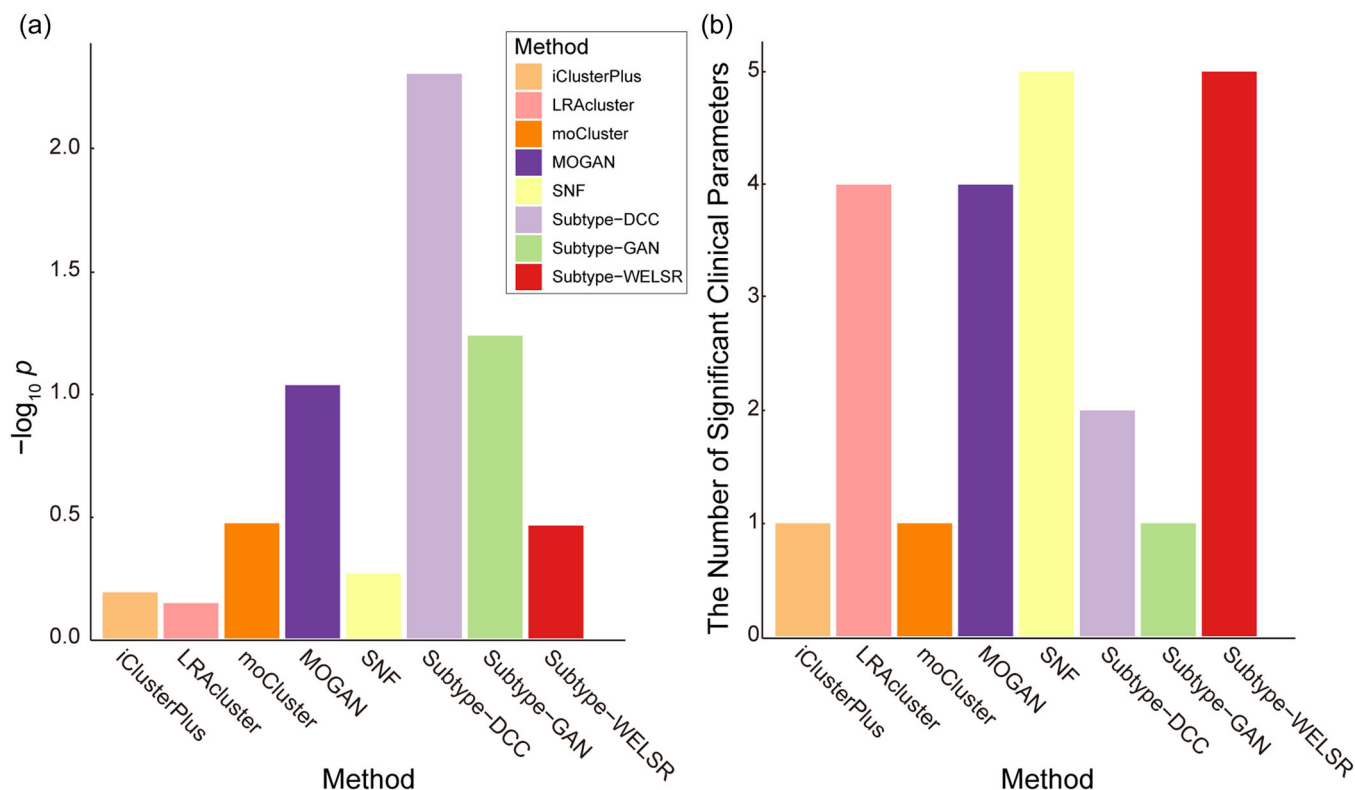


**FIGURE 4** | Clustering results comparison between the original and integrated gene expression data. (a) Venn diagram illustrating the overlap and differences in sample clustering between the original gene expression matrix and fusion matrix of threeomics data. (b) Dimensionality reduction visualization of clustered different samples at the proteomic data level. (c) Dimensionality reduction visualization of clustered different samples at the methylation data level. T\_S1 represents the clustering result of the original gene expression matrix as S1, T\_S2 represents the clustering result of the original gene expression matrix as S2, MOGANTPM\_S1 represents the clustering result of the threeomics fusion matrix as S1, and MOGANTPM\_S2 represents the clustering result of the threeomics fusion matrix as S2. MOGAN, multi-omics-based generative adversarial network.

iClusterPlus (one), moCluster (one), Subtype-GAN (one), and Subtype-GAN (two) (Figure 5b).

Subtype-DCC and Subtype-GAN outperformed the MOGAN method in the  $-\log_{10} p$  evaluation but had fewer significant clinical indicators. Similarly, SNF and Subtype-WELSR had

certain advantages when assessing significant clinical indicators, but their  $-\log_{10} p$  were significantly lower than that of the MOGAN method. In both evaluation metrics, MOGAN ranked above average and achieved a certain balance, indicating that the LUAD subtypes identified by the MOGAN method have a certain degree of clinical significance.



**FIGURE 5** | The performance of MOGAN and seven other clustering methods. MOGAN was compared with seven other multi-omics data clustering methods, including iClusterPlus, LRAcluster, moCluster, SNF, Subtype-DCC, Subtype-GAN, and Subtype-WELSR. (a) The  $-\log_{10} p$  of the eight methods. (b) The number of significant parameters of the eight methods. MOGAN, multi-omics-based generative adversarial network.

### 3.4 | Clinical Features of the LUAD Immune Subtypes

In addition, the age, gender, stage, and smoking status distributions were compared between the two immune subtypes (MOGANTPM\_S1 and MOGANTPM\_S2). The results showed that MOGANTPM\_S1 was predominantly composed of elderly individuals (age  $\geq 65$  years), whereas MOGANTPM\_S2 had a higher proportion of young people ( $p = 0.008$ ). There were also significant differences in the gender distribution and smoking status between the two subtypes, with a higher proportion of male patients and smokers in MOGANTPM\_S2 ( $p = 0.018$  and  $0.022$ , respectively). For detailed patient information, please refer to Table 2.

### 3.5 | Tumor Genomic Variations in the LUAD Immune Subtypes

To investigate the mutational differences between the two LUAD immune subtypes, significant mutation genes (SMG) were analyzed using the TCGA-LUAD whole-exome sequencing data (Figure 6a). Genes such as *MUC16* and *TTN* had a higher mutation frequency in MOGANTPM\_S2 compared with MOGANTPM\_S1 (*Chi-square* test,  $p < 0.001$ ), with missense mutations being the main mutation type in LUAD patients. To further investigate the gene mutation differences between the LUAD immune subtypes, bar charts were used to display the top 20 genes with the highest mutation frequencies between MOGANTPM\_S1 (Figure 6b) and MOGANTPM\_S2 (Figure 6c). Among them, SMGs exhibited a higher

mutation frequency in MOGANTPM\_S2 compared with MOGANTPM\_S1 (Figure 6c). Notably, the driver gene *EGFR* in LUAD had a high mutation frequency in MOGANTPM\_S1 (20.15%, Figure 6b), while its mutation frequency in MOGANTPM\_S2 was significantly lower (9.74%; *Chi-square* test,  $p = 0.008$ ). However, common LUAD driver genes, such as *ALK*, *HER2*, and *ROS1*, had low mutation frequencies in both subtypes. Moreover, MOGANTPM\_S2 exhibited a significantly higher TMB compared with MOGANTPM\_S1 (Figure 6d, *Mann-Whitney U* test,  $p = 3.3 \times 10^{-7}$ ), which may be associated with the enhanced immune response of the patients. However, there was no significant difference in survival between the high-TMB and low-TMB groups (Supporting Information S1: Figure S2, *Log-rank* test,  $p = 0.33$ ), indicating that TMB was not an independent prognostic marker, which was consistent with previous research findings [40].

In addition, the CNVs in MOGANTPM\_S2 were significantly higher than those in MOGANTPM\_S1 (*Mann-Whitney U* test,  $p = 1.1 \times 10^{-5}$ ). Analysis of local chromosomal CNVs revealed significant amplifications in the 14q13.3 region in MOGANTPM\_S1 and the 1q21.3, 1q22, 14q13.1, and 14q13.3 regions in MOGANTPM\_S2 (false discovery rate [FDR]  $< 0.05$ ). Conversely, significant deletions were observed in the 1p13.2, 9p21.3, and 13q12.11 regions in MOGANTPM\_S1, as well as in the 9p21.3 region in MOGANTPM\_S2 (Figure 6e,f; FDR  $< 0.05$ ). One of the significantly amplified loci in MOGANTPM\_S2, 1q21.3 (*SETDB1*) is known to be involved in tumor immune evasion processes [41], potentially contributing to the sub-optimal response to immunotherapy in MOGANTPM\_S2 patients.



**TABLE 2** | Clinical characteristic differences in LUAD patients.

Clinical information	MOGANTPM_S1	MOGANTPM_S2	Total	<i>p</i> <sup>a</sup>
Age (year)				0.008
≥ 65	69	81	150	
< 65	46	102	148	
Unknown	2	5	7	
Gender				0.018
Male	43	95	138	
Female	74	93	167	
Pathologic stage				0.339
I	69	92	161	
II	27	50	77	
III	16	37	53	
IV	4	8	12	
Unknown	1	1	2	
Smoke status				0.022
Smoker	72	139	211	
Nonsmoker	45	49	94	

Abbreviations: LUAD, lung adenocarcinoma; MOGAN, multi-omics-based generative adversarial network.

<sup>a</sup>Pearson's chi-squared test.

### 3.6 | Immune Microenvironment Differences in the LUAD Immune Subtypes

The ESTIMATE algorithm was used to assess the immune cell infiltration and tumor purity heterogeneity of the two subtypes. MOGANTPM\_S1 had significantly higher stromal and immune scores than MOGANTPM\_S2 (Figure 7a,b, *Mann-Whitney U* test,  $p < 2.2 \times 10^{-16}$ ). Additionally, the abundance of infiltrating immune cells in the two subtypes was evaluated using an immune gene set and the ssGSEA algorithm. The ssGSEA results revealed that MOGANTPM\_S1 exhibited higher levels of immune cell infiltration (Figure 7c), suggesting that the MOGANTPM\_S1 subtype might respond better to immunotherapy. However, the infiltration levels of *Gamma delta T cells*, *Memory B cells*, and *CD56dim natural killer cells* were higher in MOGANTPM\_S2 compared with MOGANTPM\_S1. To further evaluate the prognostic value of TIICs, we conducted a survival analysis using different immune cell types. The results showed that higher infiltration levels of Activated B cells (*Log-rank* test,  $p = 0.022$ ) and Macrophages (*Log-rank* test,  $p < 10^{-4}$ ) were associated with a better prognosis. In contrast, higher infiltration levels of Effector memory CD4 cells (*Log-rank* test,  $p = 0.026$ ) and Gamma delta T cells (*Log-rank* test,  $p = 0.002$ ) were associated with a worse prognosis (Figure 7d).

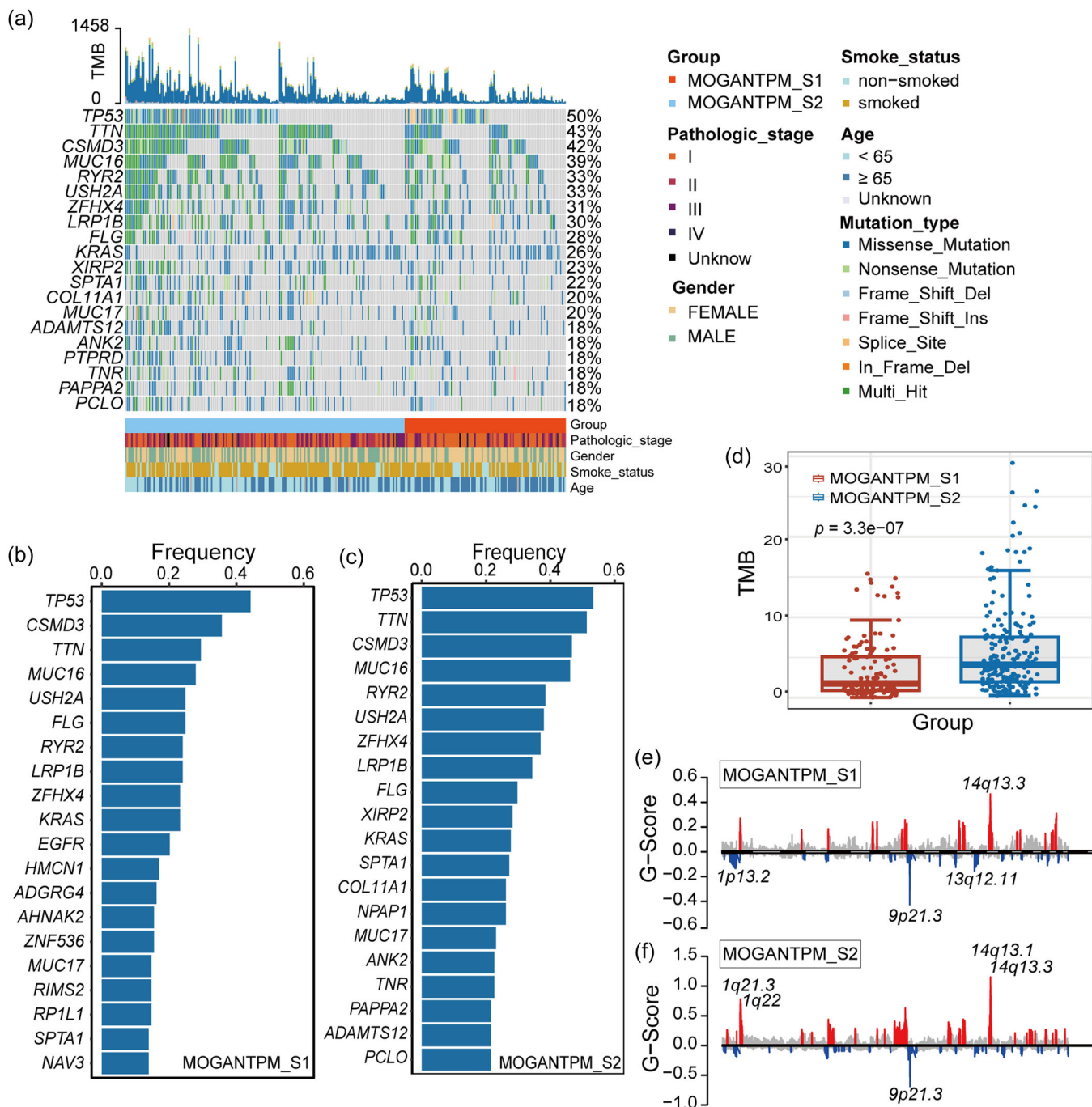
### 3.7 | Building a Diagnostic Model for the LUAD Immune Subtypes

As described earlier, we used transcriptome-proteome-methylation integrated data to classify the LUAD subtypes, finding that gene expression and protein expression contributed greatly to the classification identification. We then planned to use these two omics data types to construct a

LUAD subtype diagnostic model for clinical subtype classification. The MOGANTPM\_S1 and MOGANTPM\_S2 subtypes were compared at the gene and protein levels, with the results shown in Figure 8a. Generally, gene expression was positively correlated with protein abundance, so significantly different genes and proteins were concentrated in the third and seventh quadrants ( $\log_{2}FC$  of transcript  $> 0.25$  and  $\log_{2}FC$  of protein  $> 0.5$ ;  $\log_{2}FC$  of transcript  $< -0.25$  and  $\log_{2}FC$  of protein  $< -0.5$ ). Five genes were ultimately identified: *PRKCB*, *MYH11*, *CCNB1*, *CCNE1*, and *FOXM1* (Figure 8a). A subtype diagnostic model using logistic regression was constructed using these five genes to predict the subtype of patients. To enhance the robustness of this model, 40% of the samples were used for training and 60% for model validation. The AUC value of the model in the validation set was 0.8562 (Figure 8b), with an accuracy value of 0.8099 and an F1-score of 0.7578, allowing it to effectively distinguish the LUAD immune subtypes. The confusion matrix of the model in the validation set is shown in Supporting Information S1: Table S2.

### 3.8 | Predicting the ICB Responses of the LUAD Immune Subtypes

ICI therapy is considered to be one of the most successful approaches in cancer treatment in recent years. However, only a minority of eligible candidates respond to ICB therapy [42]. The TIDE algorithm can be used to simulate the tumor immune escape mechanism and predict patient responses to ICIs. The TIDE algorithm was used with the TCGA-LUAD cohort data for ICB response prediction, with the MOGANTPM\_S2 TIDE score found to be significantly higher than that of MOGANTPM\_S1 (Figure 9a, *Mann-Whitney U* test,  $p < 2.2 \times 10^{-16}$ ). Similarly, we

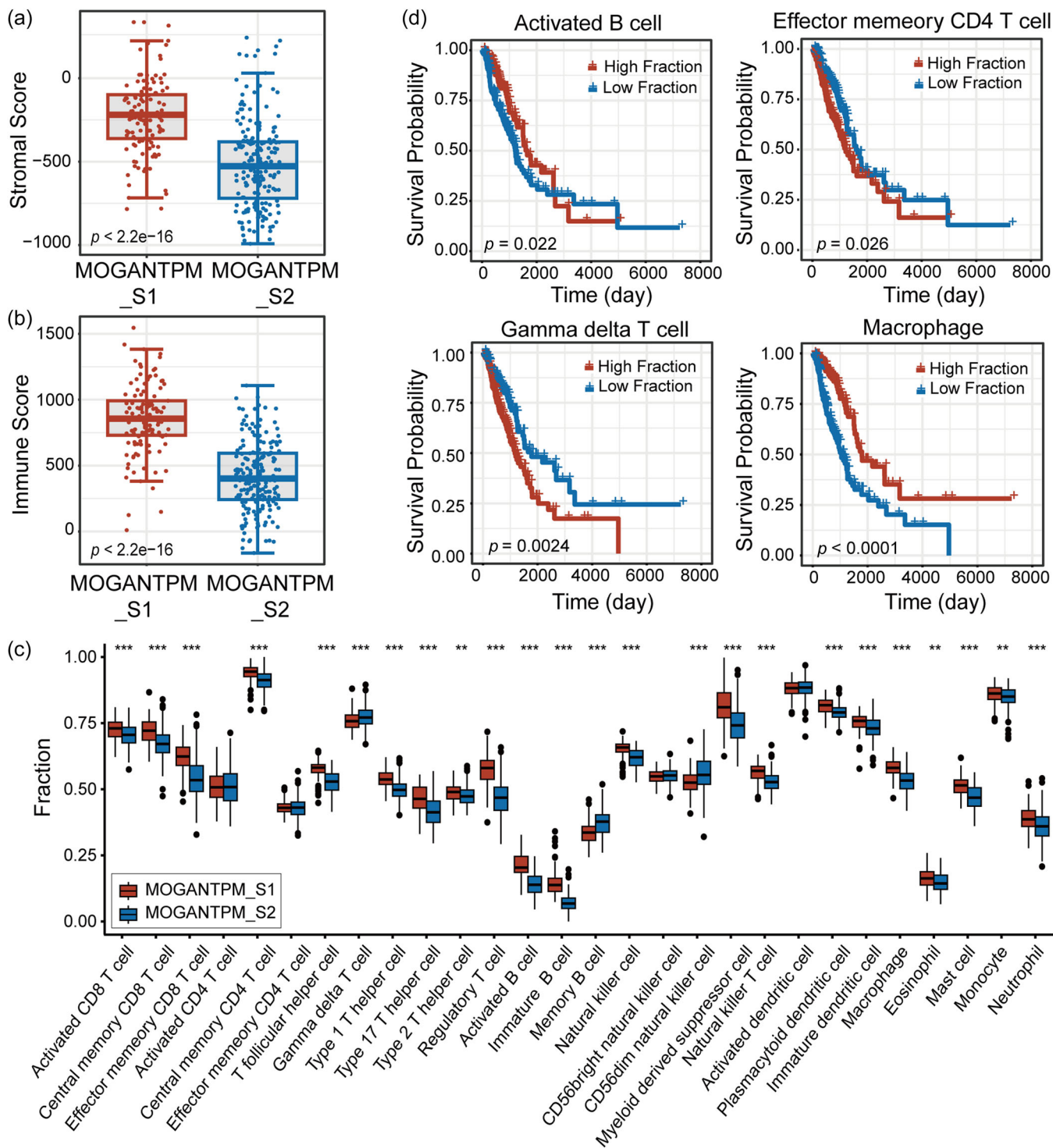


**FIGURE 6** | Mutation landscape of the two LUAD immune subtypes. (a) Overall mutation landscape of the LUAD immune subtypes. (b, c) Genes with the highest mutation frequencies (TOP 20) in MOGANTPM\_S1 and MOGANTPM\_S2. (d) The difference in tumor mutation burden between the MOGANTPM\_S1 and MOGANTPM\_S2 subtypes. The Mann-Whitney  $U$  test was used for statistical comparisons between the two subtypes. (e, f) Significant amplifications (in red) or deletions (in blue) of chromosomal regions in MOGANTPM\_S1 and MOGANTPM\_S2. LUAD, lung adenocarcinoma; MOGAN, multi-omics-based generative adversarial network.

applied the above diagnostic model to the three GEO cohort datasets and compared the TIDE scores between the two subtypes. In the GSE30219, GSE42127, and GSE31210 datasets, the MOGANTPM\_S2 TIDE scores were all higher than those of MOGANTPM\_S1 (Figure 9b-d, Mann-Whitney  $U$  test,  $p < 0.01$ ). This indicates that LUAD patients with the MOGANTPM\_S2 subtype had a higher risk of immune dysfunction and immune rejection, suggesting that they may be insensitive to ICB therapy.

#### 4 | Discussion

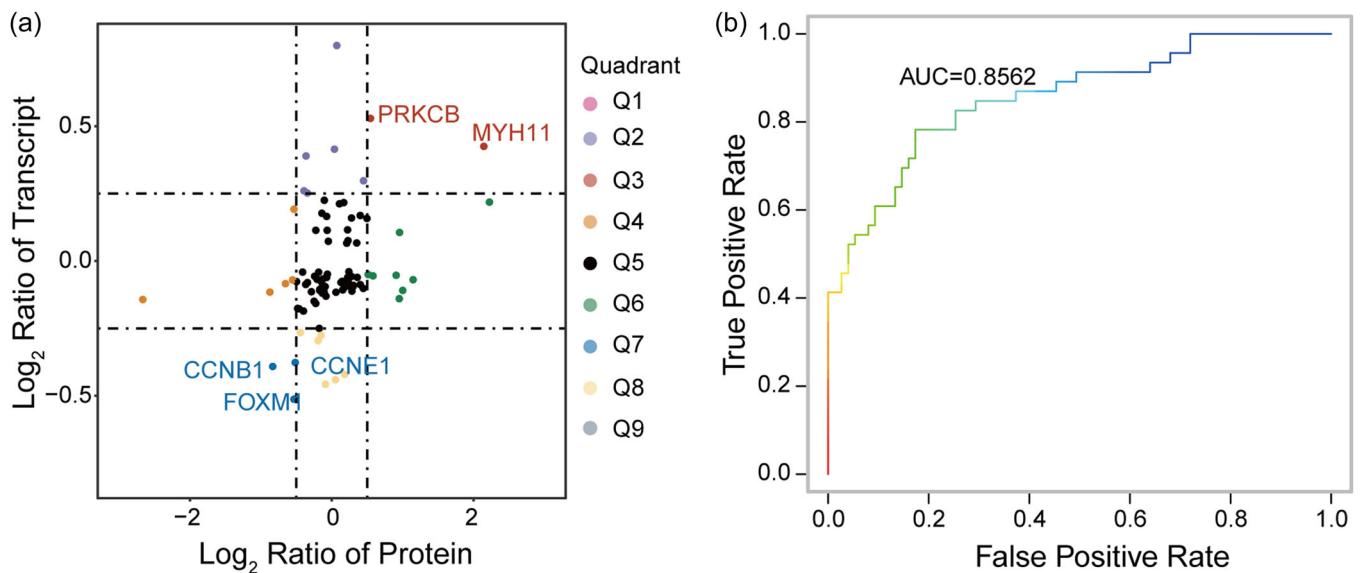
As LUAD is a heterogeneous tumor type, it is difficult to accurately and comprehensively characterize its molecular features at a single-omics level. Some researchers have used deep learning methods to identify subtypes [43, 44] but did not fully consider the interactions among different omics. Therefore, we used an improved GAN method to integrate transcriptomic, proteomic, and epigenomic data, as well as the



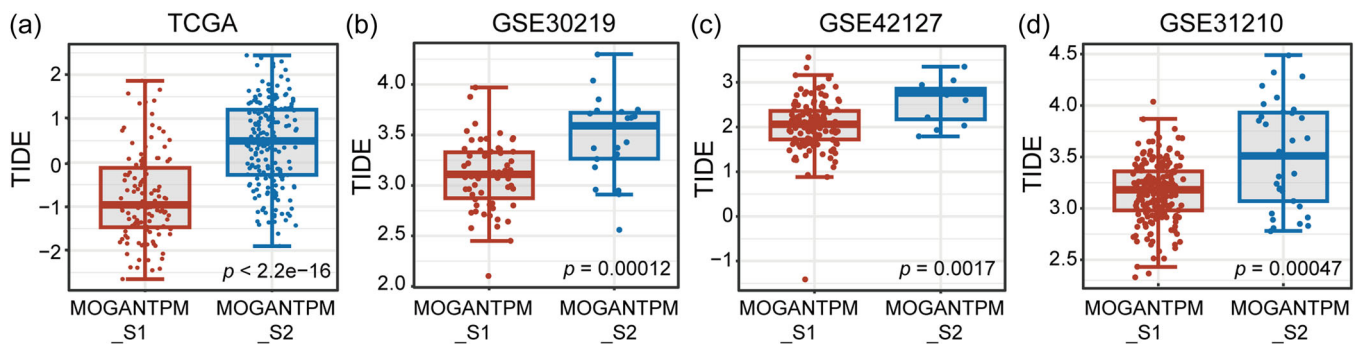
**FIGURE 7** | Immune infiltration landscape of the LUAD immune subtypes. (a) The stromal scores of the LUAD immune subtypes using the ESTIMATE algorithm were statistically significantly different using the Mann–Whitney  $U$  test. (b) The immune scores of the LUAD immune subtypes using the ESTIMATE algorithm were statistically significantly different using the Mann–Whitney  $U$  test. (c) Scores of 22 TIIC types in the MOGANTPM\_S1 and MOGANTPM\_S2 subtypes. Statistical differences between the two subtypes were assessed using the Mann–Whitney  $U$  test (\*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ). (d) The Kaplan–Meier survival curves of TIICs were significantly associated with patient survival rates. LUAD, lung adenocarcinoma; MOGAN, multi-omics-based generative adversarial network; TIIC, tumor-infiltrating immune cell.

interactions between these omics, to identify LUAD molecular subtypes. In this study, we constructed an interaction network between genes and methylation sites and between genes and proteins for multi-omics data integration. Subsequently, an improved GAN (MOGAN) was employed to fuse the multi-

omics data and interaction network, enabling the representation of disease multi-omics landscapes at the molecular level. We first used the new integrated data set to identify LUAD immune subtypes, then used the raw gene expression data set to cluster the subtypes and compared the differences among the



**FIGURE 8** | Diagnostic model construction for the LUAD immune subtypes. (a) Nine-quadrant plot of differential expression information of the transcriptome and proteome. The horizontal axis represents the differentially expressed multiplicity of proteins and the vertical axis represents the differentially expressed multiplicity of genes. The right side shows the high-expressed proteins, the left side shows the low-expressed proteins, the upper portion shows the high-expressed genes, and the lower portion shows the low-expressed genes. (b) Working characterization curves of subjects in the validation set for the LUAD subtype diagnostic model. LUAD, lung adenocarcinoma.



**FIGURE 9** | Prediction of immune checkpoint blockade responses for the two LUAD immune subtypes. (a) Box plot of TIDE scores for the MOGANTPM\_S1 and MOGANTPM\_S2 subtypes in the TCGA-LUAD cohort, with statistical differences between the two groups compared using the Mann-Whitney  $U$  test. (b-d) Box plots of TIDE scores for the MOGANTPM\_S1 and MOGANTPM\_S2 subtypes in the GSE30219, GSE42127, and GSE31210 cohorts, with statistical differences between the two groups compared using the Mann-Whitney  $U$  test. LUAD, lung adenocarcinoma; MOGAN, multi-omics-based generative adversarial network.

datasets. The classification effect of the integrated data set (AUC = 0.882, IDI = 0.117, NRI = 0.514) was better than that of other datasets (Figure 2 and Table 1), suggesting a successful integration of the transcriptomics, proteomics, and epigenomics data. Additionally, by integrating the multi-omics information and applying immune cell infiltration analysis, we identified two LUAD immune subtypes: MOGANTPM\_S1 and MOGANTPM\_S2. The prognosis of MOGANTPM\_S1 patients was significantly better than that of MOGANTPM\_S2 patients, with the MOGANTPM\_S1 stromal and immune scores also being higher. Most of the examined immune cell types exhibited higher infiltration levels in MOGANTPM\_S1 than in MOGANTPM\_S2, suggesting that MOGANTPM\_S1 patients might have a more favorable response to immune therapy. However, the infiltration levels of Gamma delta T cells, Memory B cells, and CD56dim natural killer cells were higher in MOGANTPM\_S2, with the reason for this remaining unclear.

Further analysis indicated that the mutation frequency of MOGANTPM\_S1 was lower than that of MOGANTPM\_S2, but the mutation frequency of the LUAD driver gene *EGFR* was higher. We speculated that *EGFR*-targeted drugs might have better efficacy in MOGANTPM\_S1 patients. In addition, MOGANTPM\_S1 exhibited a lower TMB and higher levels of immune cell infiltration, suggesting that there might be interaction mechanisms between TMB and immune cell infiltration in MOGANTPM\_S1. Although MOGANTPM\_S2 demonstrated a higher TMB, certain studies have indicated that some patients may not experience clinical benefits even with a high TMB [45, 46]. Moreover, MOGANTPM\_S2 exhibited a significant amplification of the 1q21.3 locus (*SETDB1*), which is associated with tumor immune evasion processes, which could explain the poor efficacy of immune therapy in MOGANTPM\_S2 patients. Finally, we used the TIDE algorithm to predict ICB responses in the two subtypes, with MOGANTPM\_S1 patients being more

sensitive to ICB therapy. In summary, we identified two LUAD subtypes, MOGANTPM\_S1 and MOGANTPM\_S2, with MOGANTPM\_S1 patients found to have a better prognosis and potentially more sensitivity to immune therapy.

Furthermore, we constructed a subtype diagnostic model using logistic regression with the differentially expressed genes (proteins) between the two subtypes, including *PRKCB*, *MYH11*, *CCNB1*, *CCNE1*, and *FOXM1*. This model reliably predicted the immune therapy response for MOGANTPM\_S1 and MOGANTPM\_S2 patients in different datasets. The majority of genes in the subtype diagnostic model are associated with cancer development. For example, protein kinase C beta (*PRKCB*) belongs to the protein kinase C (PKC) family, a family of serine/threonine-specific protein kinases. PKC family members phosphorylate a wide variety of protein targets and are involved in multiple cellular signaling pathways. They also serve as major receptors for phorbol esters, a class of tumor promoters. *PRKCB* can reportedly affect the prognosis of LUAD patients through methylation and immune infiltration [47]. Myosin heavy chain 11 (*MYH11*) serves as a contractile protein and converts chemical energy into mechanical energy through ATP hydrolysis. Low expression of *MYH11* can promote cancer cell metastasis, proliferation, and invasion [48]. Researchers have found that decreased *MYH11* expression levels are associated with poor prognosis in lung cancer patients [49]. Cyclin E1 (*CCNE1*) promotes cell cycle progression by forming complexes [50] and is associated with poor prognosis in breast cancer, bladder cancer, and colorectal cancer patients [51]. Furthermore, *CCNE1* may serve as a biomarker for immune therapy response in certain types of cancer [52]. Cyclin B1 (*CCNB1*) promotes the transition of cells from the G2 phase to the M phase of the cell cycle [53]. Overexpression of *CCNB1* promotes LUAD progression, while miR-139-5 negatively regulates *CCNB1* in LUAD, thereby inhibiting cell proliferation, migration, invasion, and cell cycle progression [54]. Forkhead box protein M1 (*FOXM1*) plays roles in cell cycle progression, cancer therapy resistance, and metastasis of lung cancer. High *FOXM1* mRNA expression patterns may serve as an independent biomarker for poor prognosis in LUAD patients [55]. Additionally, *FOXM1* regulates the expression of *PD-L1*, induces cell death, and may serve as a favorable option for improving LUAD patient prognosis through ICIs [56].

## 5 | Conclusions

Overall, this study optimized the previously developed GAN and successfully integrated three omics data types using the MOGAN method to identify two LUAD immune subtypes. Additionally, a subtype diagnostic model was constructed for clinical subtype detection, with the reliability of the model being validated at multiple levels. However, this study still had some limitations. This study included 305 patients from the TCGA database and 444 patients from the GEO database for LUAD subtype diagnosis and prognosis prediction. The inclusion of data solely from public databases may have introduced

inherent biases, so further validation using prospective clinical data is necessary in the future.

---

### Author Contributions

**Haibin He:** data curation (lead), formal analysis (lead), writing – original draft (equal). **Longxing Wang:** formal analysis (supporting), validation (supporting). **Mingyue Ma:** funding acquisition (lead), project administration (lead), writing – review and editing (lead).

### Acknowledgments

The authors are very grateful for the public availability of the various databases.

### Ethics Statement

The authors have nothing to report.

### Consent

The authors have nothing to report.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

All data generated or analyzed during this study are included in this published article and its supplementary information files. The scripts were uploaded to <https://github.com/Hehb123/MOGAN>.

### References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians* 68, no. 6 (2018): 394–424, <https://doi.org/10.3322/caac.21492>.
2. C. Xia, X. Dong, H. Li, et al., “Cancer Statistics in China and United States, 2022: Profiles, Trends, and Determinants,” *Chinese Medical Journal* 135, no. 5 (2022): 584–590, <https://doi.org/10.1097/CM9.0000000000002108>.
3. L. Li, Y. Chen, W. Liao, et al., “Associations of IFT20 and GM130 Protein Expressions With Clinicopathological Features and Survival of Patients With Lung Adenocarcinoma,” *BMC Cancer* 22, no. 1 (2022): 809, <https://doi.org/10.1186/s12885-022-09905-6>.
4. T. T. Nguyen, H. S. Lee, B. M. Burt, et al., “A Lepidic Gene Signature Predicts Patient Prognosis and Sensitivity to Immunotherapy in Lung Adenocarcinoma,” *Genome Medicine* 14, no. 1 (2022): 5, <https://doi.org/10.1186/s13073-021-01010-w>.
5. C. Yuan, H. Chen, S. Tu, et al., “A Systematic Dissection of the Epigenomic Heterogeneity of Lung Adenocarcinoma Reveals Two Different Subclasses With Distinct Prognosis and Core Regulatory Networks,” *Genome Biology* 22, no. 1 (2021): 156, <https://doi.org/10.1186/s13059-021-02376-1>.
6. E. A. Collisson, P. Bailey, D. K. Chang, and A. V. Biankin, “Molecular Subtypes of Pancreatic Cancer,” *Nature Reviews Gastroenterology & Hepatology* 16, no. 4 (2019): 207–220, <https://doi.org/10.1038/s41575-019-0109-y>.
7. S. Shi, M. Xu, and Y. Xi, “Molecular Subtypes Based on DNA Promoter Methylation Predict Prognosis in Lung Adenocarcinoma Patients,” *Aging* 12, no. 23 (2020): 23917–23930, <https://doi.org/10.18632/aging.104062>.

8. Q. Wang, M. Li, M. Yang, et al., "Analysis of Immune-Related Signatures of Lung Adenocarcinoma Identified Two Distinct Subtypes: Implications for Immune Checkpoint Blockade Therapy," *Aging* 12, no. 4 (2020): 3312–3339, <https://doi.org/10.18632/aging.102814>.
9. M. Zhang, J. Ma, Q. Guo, S. Ding, Y. Wang, and H. Pu, "CD8<sup>+</sup>," *Frontiers in Immunology* 13 (2022): 806877, <https://doi.org/10.3389/fimmu.2022.806877>.
10. X. He, X. Liu, F. Zuo, H. Shi, and J. Jing, "Artificial Intelligence-Based Multi-Omics Analysis Fuels Cancer Precision Medicine," *Seminars in Cancer Biology* 88 (2023): 187–200, <https://doi.org/10.1016/j.semcancer.2022.12.009>.
11. J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and H. Dawood, "A Hierarchical Integration Deep Flexible Neural Forest Framework for Cancer Subtype Classification by Integrating Multi-Omics Data," *BMC Bioinformatics* 20, no. 1 (2019): 527, <https://doi.org/10.1186/s12859-019-3116-7>.
12. R. G. W. Verhaak, K. A. Hoadley, E. Purdom, et al., "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in Pdgfra, IDH1, Egfr, and Nf1," *Cancer Cell* 17, no. 1 (2010): 98–110, <https://doi.org/10.1016/j.ccr.2009.12.020>.
13. P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, "Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review," *Biotechnology Advances* 49 (2021): 107739, <https://doi.org/10.1016/j.biotechadv.2021.107739>.
14. K. T. Ahmed, J. Sun, S. Cheng, J. Yong, and W. Zhang, "Multi-Omics Data Integration by Generative Adversarial Network," *Bioinformatics* 38, no. 1 (2021): 179–186, <https://doi.org/10.1093/bioinformatics/btab608>.
15. J. Peng, W. Hui, Q. Li, et al., "A Learning-Based Framework for miRNA-Disease Association Identification Using Neural Networks," *Bioinformatics* 35, no. 21 (2019): 4364–4371, <https://doi.org/10.1093/bioinformatics/btz254>.
16. T. Y. Lee, K. Y. Huang, C. H. Chuang, C. Y. Lee, and T. H. Chang, "Incorporating Deep Learning and Multi-Omics Autoencoding for Analysis of Lung Adenocarcinoma Prognostication," *Computational Biology and Chemistry* 87 (2020): 107277, <https://doi.org/10.1016/j.combiolchem.2020.107277>.
17. J. C. Hegenbarth, G. Lezsoche, L. J. De Windt, and M. Stoll, "Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics," *Frontiers in Molecular Medicine* 2 (2022): 1–14, <https://doi.org/10.3389/fmmed.2022.839338>.
18. L. D. Moore, T. Le, and G. Fan, "DNA Methylation and Its Basic Function," *Neuropsychopharmacology* 38, no. 1 (2013): 23–38, <https://doi.org/10.1038/npp.2012.112>.
19. B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool, "Proteomics: Technologies and Their Applications," *Journal of Chromatographic Science* 55, no. 2 (2017): 182–196, <https://doi.org/10.1093/chromsci/bmw167>.
20. W. C. Cho, "Proteomics Technologies and Challenges," *Genomics, Proteomics & Bioinformatics* 5, no. 2 (2007): 77–85, [https://doi.org/10.1016/S1672-0229\(07\)60018-7](https://doi.org/10.1016/S1672-0229(07)60018-7).
21. K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge," *Contempor Oncol* 19, no. 1A (2015): 68–77, <https://doi.org/10.5114/wo.2014.47136>.
22. B. T. Lee, G. P. Barber, A. Benet-Pagès, et al., "The UCSC Genome Browser Database: 2022 Update," *Nucleic Acids Research* 50, no. D1 (2022): D1115–D1122, <https://doi.org/10.1093/nar/gkab959>.
23. J. Li, Y. Lu, R. Akbani, et al., "TCPA: A Resource for Cancer Functional Proteomics Data," *Nature Methods* 10, no. 11 (2013): 1046–1047, <https://doi.org/10.1038/nmeth.2650>.
24. T. Barrett, S. E. Wilhite, P. Ledoux, et al., "NCBI GEO: Archive for Functional Genomics Data Sets: Update," *Nucleic Acids Research* 41, no. Database issue (2013): 991–995, <https://doi.org/10.1093/nar/gks1193>.
25. C. Mering, "STRING: A Database of Predicted Functional Associations between Proteins," *Nucleic Acids Research* 31, no. 1 (2003): 258–261, <https://doi.org/10.1093/nar/gkg034>.
26. X. Li, W. Zhang, J. Song, X. Zhang, L. Ran, and Y. He, "SLCO4C1 Promoter Methylation Is a Potential Biomarker for Prognosis Associated With Biochemical Recurrence-Free Survival After Radical Prostatectomy," *Clinical Epigenetics* 11, no. 1 (2019): 99, <https://doi.org/10.1186/s13148-019-0693-2>.
27. Y. Lou, J. Xu, Y. Zhang, et al., "Akt Kinase LANCL2 Functions as a Key Driver in EGFR-Mutant Lung Adenocarcinoma Tumorigenesis," *Cell Death & Disease* 12, no. 2 (2021): 170, <https://doi.org/10.1038/s41419-021-03439-8>.
28. Q. Song, S. D. Merajver, and J. Z. Li, "Cancer Classification in the Genomic Era: Five Contemporary Problems," *Human Genomics* 9, no. 1 (2015): 27, <https://doi.org/10.1186/s40246-015-0049-8>.
29. T. M. Therneau, *A package for survival analysis in R. R Package Ver 3.7-0*. 2024. <https://doi.org/10.32614/CRAN.package.survival>.
30. I. Dormuth, T. Liu, J. Xu, M. Yu, M. Pauly, and M. Ditzhaus, "Which Test for Crossing Survival Curves? A User's Guideline," *BMC Medical Research Methodology* 22, no. 1 (2022): 34, <https://doi.org/10.1186/s12874-022-01520-0>.
31. A. J. Gentles, A. M. Newman, C. L. Liu, et al., "The Prognostic Landscape of Genes and Infiltrating Immune Cells Across Human Cancers," *Nature Medicine* 21, no. 8 (2015): 938–945, <https://doi.org/10.1038/nm.3909>.
32. S. T. Paijens, A. Vledder, M. de Bruyn, and H. W. Nijman, "Tumor-Infiltrating Lymphocytes in the Immunotherapy Era," *Cellular & Molecular Immunology* 18, no. 4 (2021): 842–859, <https://doi.org/10.1038/s41423-020-00565-9>.
33. Q. Mo, S. Wang, V. E. Seshan, et al., "Pattern Discovery and Cancer Gene Identification in Integrated Cancer Genomic Data," *Proceedings of the National Academy of Sciences* 110, no. 11 (2013): 4245–4250, <https://doi.org/10.1073/pnas.1208949110>.
34. D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast Dimension Reduction and Integrative Clustering of Multi-Omics Data Using Low-Rank Approximation: Application to Cancer Molecular Classification," *BMC Genomics* 16, no. 1 (2015): 1022, <https://doi.org/10.1186/s12864-015-2223-8>.
35. C. Meng, D. Helm, M. Frejno, and B. Kuster, "Mocluster: Identifying Joint Patterns Across Multiple Omics Data Sets," *Journal of Proteome Research* 15, no. 3 (2016): 755–765, <https://doi.org/10.1021/acs.jproteome.5b00824>.
36. B. Wang, A. M. Mezlini, F. Demir, et al., "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale," *Nature Methods* 11, no. 3 (2014): 333–337, <https://doi.org/10.1038/nmeth.2810>.
37. J. Zhao, B. Zhao, X. Song, et al., "Subtype-DCC: Decoupled Contrastive Clustering Method for Cancer Subtype Identification Based on Multi-Omics Data," *Briefings in Bioinformatics* 24, no. 2 (2023): bbad025, <https://doi.org/10.1093/bib/bbad025>.
38. H. Yang, R. Chen, D. Li, and Z. Wang, "Subtype-Gan: A Deep Learning Approach for Integrative Cancer Subtyping of Multi-Omics Data," *Bioinformatics* 37, no. 16 (2021): 2231–2237, <https://doi.org/10.1093/bioinformatics/btab109>.
39. W. Song, W. Wang, and D. Q. Dai, "Subtype-WESLR: Identifying Cancer Subtype With Weighted Ensemble Sparse Latent Representation of Multi-View Data," *Briefings in Bioinformatics* 23, no. 1 (2022): bbab398, <https://doi.org/10.1093/bib/bbab398>.
40. C. Valero, M. Lee, D. Hoen, et al., "The Association Between Tumor Mutational Burden and Prognosis Is Dependent on Treatment Context,"

*Nature Genetics* 53, no. 1 (2021): 11–15, <https://doi.org/10.1038/s41588-020-00752-4>.

41. G. K. Griffin, J. Wu, A. Iracheta-Vellve, et al., “Epigenetic Silencing by SETDB1 Suppresses Tumour Intrinsic Immunogenicity,” *Nature* 595, no. 7866 (2021): 309–314, <https://doi.org/10.1038/s41586-021-03520-4>.

42. D. Pan, A. Y. Hu, S. J. Antonia, and C. Y. Li, “A Gene Mutation Signature Predicting Immunotherapy Benefits in Patients With NSCLC,” *Journal of Thoracic Oncology* 16, no. 3 (2021): 419–427, <https://doi.org/10.1016/j.jtho.2020.11.021>.

43. K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer,” *Clinical Cancer Research* 24, no. 6 (2018): 1248–1259, <https://doi.org/10.1158/1078-0432.CCR-17-0853>.

44. L. Y. Guo, A. H. Wu, Y. Wang, L. Zhang, H. Chai, and X. F. Liang, “Deep Learning-Based Ovarian Cancer Subtypes Identification Using Multi-Omics Data,” *BioData Mining* 13, no. 1 (2020): 10, <https://doi.org/10.1186/s13040-020-00222-x>.

45. D. Marinelli, M. Mazzotta, S. Scalera, et al., “KEAP1-driven Co-Mutations in Lung Adenocarcinoma Unresponsive to Immunotherapy Despite High Tumor Mutational Burden,” *Annals of Oncology* 31, no. 12 (2020): 1746–1754, <https://doi.org/10.1016/j.annonc.2020.08.2105>.

46. A. Ribas and J. D. Wolchok, “Cancer Immunotherapy Using Checkpoint Blockade,” *Science* 359, no. 6382 (2018): 1350–1355, <https://doi.org/10.1126/science.aar4060>.

47. J. Wang, M. Shi, H. Zhang, et al., “PRKCB Is Relevant to Prognosis of Lung Adenocarcinoma through Methylation and Immune Infiltration,” *Thoracic Cancer* 13, no. 12 (2022): 1837–1849, <https://doi.org/10.1111/1759-7714.14466>.

48. J. Wang, P. Xu, Y. Hao, et al., “Interaction Between DNMT3B and MYH11 via Hypermethylation Regulates Gastric Cancer Progression,” *BMC Cancer* 21, no. 1 (2021): 914, <https://doi.org/10.1186/s12885-021-08653-3>.

49. M. J. Nie, X. T. Pan, H. Y. Tao, et al., “Clinical and Prognostic Significance of MYH11 in Lung Cancer,” *Oncology Letters* 19, no. 6 (2020): 3899–3906, <https://doi.org/10.3892/ol.2020.11478>.

50. E. Y. Kang, A. Weir, N. S. Meagher, et al., “CCNE1 and Survival of Patients With Tubo-Ovarian High-Grade Serous Carcinoma: An Ovarian Tumor Tissue Analysis Consortium Study,” *Cancer* 129, no. 5 (2023): 697–713, <https://doi.org/10.1002/cncr.34582>.

51. N. Nakayama, K. Nakayama, Y. Shamima, et al., “Gene Amplification CCNE1 Is Related to Poor Survival and Potential Therapeutic Target in Ovarian Cancer,” *Cancer* 116, no. 11 (2010): 2621–2634, <https://doi.org/10.1002/cncr.24987>.

52. X. Zheng, L. Chen, W. Liu, et al., “CCNE1 Is a Predictive and Immunotherapeutic Indicator in Various Cancers Including UCEC: A Pan-Cancer Analysis,” *Hereditas* 160, no. 1 (2023): 13, <https://doi.org/10.1186/s41065-023-00273-0>.

53. M. Fischer, M. Quaas, L. Steiner, and K. Engeland, “The p53-p21-DREAM-CDE/CHR Pathway Regulates G2/M Cell Cycle Genes,” *Nucleic Acids Research* 44, no. 1 (2016): 164–174, <https://doi.org/10.1093/nar/gkv927>.

54. B. Bao, X. Yu, and W. Zheng, “MiR-139-5p Targeting CCNB1 Modulates Proliferation, Migration, Invasion and Cell Cycle in Lung Adenocarcinoma,” *Molecular Biotechnology* 64, no. 8 (2022): 852–860, <https://doi.org/10.1007/s12033-022-00465-5>.

55. H. Xiao, Z. Jiang, X. Fu, et al., “High Expression of Forkhead Box M1 (FOXM1) Is a Poor Prognostic Biomarker in Lung Adenocarcinoma,” *Translational Cancer Research* 9, no. 10 (2020): 6331–6343, <https://doi.org/10.21037/tcr-20-1103>.

56. H. Madhi, J. S. Lee, Y. E. Choi, et al., “FOXM1 Inhibition Enhances the Therapeutic Outcome of Lung Cancer Immunotherapy by Modulating PD-L1 Expression and Cell Proliferation,” *Advanced Science*

(Weinheim, Baden-Württemberg, Germany) 9, no. 29 (2022): e2202702, <https://doi.org/10.1002/advs.202202702>.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.