AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems



| Computational Biology | Commentary

# Decoding cancer prognosis with deep learning: the ASD-cancer framework for tumor microenvironment analysis

**Ziyuan Huang,**[1,2] **Yunzhan Li,**[3] **Vanni Bucci,**[2,4] **John P. Haran**[1,2,4]

**AUTHOR AFFILIATIONS** See affiliation list on p. 4.

**ABSTRACT** Deep learning is revolutionizing biomedical research by facilitating the integration of multi-omics data sets while bridging classical bioinformatics with existing knowledge. Building on this powerful potential, Zhang et al. proposed a semi-supervised learning framework called Autoencoder-Based Subtypes Detector for Cancer (ASD-cancer) to improve the multi-omics data analysis (H. Zhang, X. Xiong, M. Cheng, et al., 2024, mSystems 9:e01395-24, https://doi.org/10.1128/msystems.01395-24). By utilizing autoencoders pre-trained on The Cancer Genome Atlas data, the ASD-cancer framework outperforms the baseline model. This approach also makes the framework scalable, enabling it to process new data sets through transfer learning without retraining. This commentary explores the methodological innovations and scalability of ASD-cancer while suggesting future directions, such as the incorporation of additional data layers and the development of adaptive AI models through continuous learning. Notably, integrating large language models into ASD-cancer could enhance its interpretability, providing more profound insights into oncological research and increasing its influence in cancer subtyping and further analysis.

**KEYWORDS** deep learning, autoencoder, transfer learning, multi-omics, cancer prognosis

Deep learning is transforming cancer data analytics by facilitating the integration of multi-omics data, thereby significantly improving diagnosis, classification, and personalized treatment strategies (1–4). The ASD-cancer framework exemplifies this progress by utilizing autoencoders to extract survival-related features, categorize various cancer subtypes, and identify potential biomarkers (5). It combines gene expression data from The Cancer Genome Atlas (TCGA) with tumor microbiome profiles using Poore et al.'s microbiome profiling technique (6). The framework is validated through external cohort testing with colon cancer data from Qatar and liver cancer data from China (7, 8). However, integrating multi-omics data, particularly tumor microbiome and transcriptome profiles, remains a significant challenge in cancer research (9–12). As discussed in the ASD-cancer study, conventional methods like principal component analysis (PCA) could not effectively capture the complexity of multi-omics data sets (13, 14). This limitation hinders the discovery of intricate biological patterns, highlighting the need for more advanced analytical approaches. ASD-cancer overcomes these limitations, enabling meaningful feature extraction and deeper biological insights. By identifying survival subtypes across 20 cancer types, ASD-cancer enhances risk stratification, informs clinical decisions, and advances personalized oncology, marking a significant step toward precision medicine.

**Editor** Neha Garg, Georgia Institute of Technology, Atlanta, Georgia, USA

Address correspondence to Ziyuan Huang, Ziyuan.Huang2@umassmed.edu.

The authors declare no conflict of interest.

*The views expressed in this article do not necessarily reflect the views of the journal or of ASM.*

See the original article at https://doi.org/10.1128/msystems.01395-24.

**Published** 16 April 2025

## ASD-CANCER FRAMEWORK: A DEEP LEARNING APPROACH FOR TUMOR ANALYSIS

The ASD-cancer framework follows a structured, multiphase data processing pipeline. This method integrates autoencoders with random forest and various bioinformatics techniques, forming a semi-supervised framework for cancer subtype stratification and possible biomarker identification (see Fig. 1). This pipeline can be categorized into the following phases: data input and preprocessing, feature extraction and selection, subtype detection and survival analysis, subtype and clinical stage prediction, gene expression and microbial abundance association, and biological pathway identification.

The pipeline integrates RNA-seq and tumor microbiome data as new or test data. Data preprocessing ensures consistency and quality across multi-omics data sets through normalization.

Subsequently, autoencoders extract features from transcriptomic and microbiome data, capturing nonlinear relationships and reducing noise. Incorporating TCGA data sets enhances feature extraction effectiveness and improves model generalizability for downstream analyses without retraining. The resulting 3,000 latent features per cancer type provide a richer biological context compared to traditional dimensionality reduction techniques like PCA. Cox proportional hazards regression filters out non-informative features, ensuring that only statistically significant survival-associated biomarkers contribute to subtype detection.

Following feature selection, the Gaussian mixture model (GMM) clusters patient data into biologically distinct survival subtypes, optimizing clustering stability through silhouette scoring. Then, Kaplan-Meier survival curves, log-rank tests, and Cox regression perform the prognostic relevance analysis of these subtypes, confirming survival differences between these cancer types.

After survival analysis, the chi-square test and random forest classifier are employed to predict cancer subtypes and clinical stages using the features identified in prior

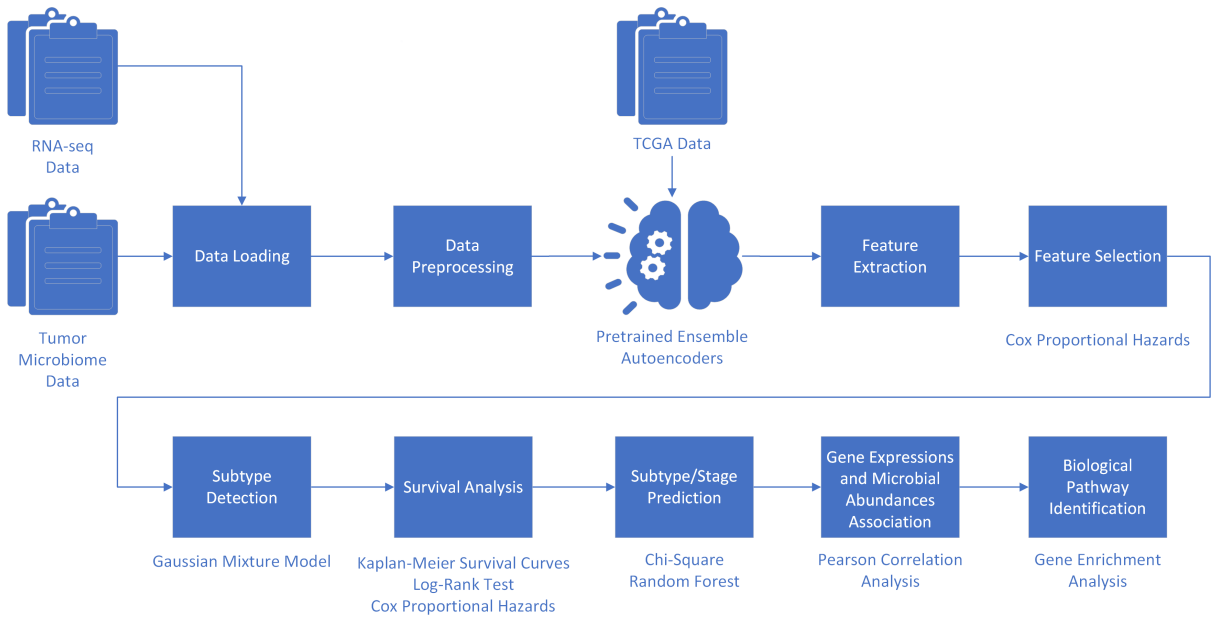


**FIG 1** Data flow diagram of ASD-cancer. The ASD-cancer framework integrates multi-omics data, deep learning, and statistical modeling to identify cancer subtypes and assess survival outcomes. The workflow begins with data loading and preprocessing RNA-seq and tumor microbiome data, incorporating TCGA data sets. Pre-trained ensemble autoencoders extract latent features, which undergo feature selection using Cox proportional hazards regression to retain survival-associated biomarkers. Subtype detection is conducted using GMM, followed by survival analysis that includes Kaplan-Meier survival curves, log-rank tests, and Cox regression to validate subtype distinctions. Subtype and stage prediction are conducted using the chi-square test and random forest classifier. Pearson correlation analysis examines associations between gene expression and microbial abundances, while gene enrichment analysis identifies biological pathways and potential biomarkers for cancer prognosis and treatment.

steps. This semi-supervised learning approach integrates the large-scale TCGA data set with smaller, heterogeneous cohorts, enhancing subtype classification generalization by improving feature quality and reducing noise.

Beyond subtype classification, Pearson correlation analysis is applied to identify associations between gene expression profiles and microbial abundances, offering a deeper understanding of potential tumor-microbiome interactions. This step aims to identify molecular signatures associated with cancer progression.

Finally, gene enrichment analysis is performed to identify molecular pathways and functional annotations associated with detected subtypes. This final step reveals potential biomarkers and mechanistic insights that may inform cancer prognosis and therapeutic strategies.

## FUTURE DIRECTIONS FOR DEEP LEARNING IN ONCOLOGY

The ASD-cancer framework has effectively showcased the potential of transfer learning implementation for enhanced multi-omics cancer data analysis. Future research could involve additional data layers, such as proteomics, metabolomics, medical imaging, and clinical data, to further test the effectiveness of transfer learning with even more data modalities. Expanding data integration in this way offers an endless opportunity to deepen our understanding of cancer dynamics, further addressing critical gaps and uncovering new insights.

In addition to data integration, another crucial area to enhance this study is the incorporation of adaptive deep learning frameworks for continuous learning. Building on the foundations of ASD-cancer, these continuous learning modules connect live clinical and multimodal data for real-time predictive analysis in alignment with the latest cancer research and treatment investigations.

Enhancing interpretability is another critical direction for ASD-cancer improvement. While the framework has shown impressive performance in survival subtyping and classification, the intermediate and final outputs require greater explainability. One promising approach is integrating large language models and ASD-cancer to produce natural language explanations of these outputs (15). This integration would help to connect the gap between complex data sets, AI-driven insights, and clinical decision-making, potentially making results more intuitive and translating ASD-cancer's findings into actionable interventions.

## CONCLUSION

The ASD-cancer framework integrates deep learning, machine learning, and bioinformatics while utilizing TCGA data through transfer learning to investigate multi-omics cancer data, aiming to enhance predictive accuracy and result reliability. Its semi-supervised architecture extracts survival-related features from such complex data sets, identifying distinct subtypes across 20 cancer types. The framework outperforms the baseline model in cancer risk stratification and provides biologically meaningful insights into host gene expression-microbiome interactions. The ASD-cancer framework offers a robust foundation for understanding tumor microenvironment dynamics and enhancing the reliability of risk stratification. By leveraging its artificial neural network architecture to integrate diverse multi-omics data, ASD-cancer serves as a valuable and scalable platform for understanding tumor microenvironment dynamics, improving risk stratification, and guiding future investigations into personalized treatment strategies.

## ACKNOWLEDGMENTS

We acknowledge the valuable feedback and discussions from our colleagues and collaborators, which helped shape the perspectives presented in this commentary.

## AUTHOR AFFILIATIONS

[1]Department of Emergency Medicine, UMass Chan Medical School, Worcester, Massachusetts, USA
[2]Department of Microbiology, UMass Chan Medical School, Worcester, Massachusetts, USA
[3]Department of Cellular and Molecular Physiology, Penn State College of Medicine, Hershey, Pennsylvania, USA
[4]Program in Microbiome Dynamics, UMass Chan Medical School, Worcester, Massachusetts, USA

## AUTHOR ORCIDs

Ziyuan Huang http://orcid.org/0000-0002-2215-2473
John P. Haran http://orcid.org/0000-0001-7311-1121

## AUTHOR CONTRIBUTIONS

Ziyuan Huang, Conceptualization, Visualization, Writing – original draft, Writing – review and editing | Yunzhan Li, Conceptualization, Writing – review and editing | Vanni Bucci, Conceptualization, Writing – review and editing | John P. Haran, Conceptualization, Writing – review and editing

## REFERENCES

1. Sarada T, Sruthi A. 2024. Deep learning for forecast, treatment, and diagnosis of cancer. 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC); Coimbatore, India: , p 1790–1795
2. Alharbi F, Vakanski A, Zhang B, Elbashir MK, Mohammed M. 2025. Comparative analysis of multi-omics integration using graph neural networks for cancer classification. IEEE Access 13:37724–37736. https://doi.org/10.1109/ACCESS.2025.3540769
3. Waqas A, Tripathi A, Ahmed S, Mukund A, Farooq H, Johnson J, Stewart P, Naeini M, Schabath MB, Rasool G. 2025. Self-normalizing foundation model for enhanced multi-omics data analysis in oncology. SSRN. https://doi.org/10.2139/ssrn.5055163
4. Sujatha P, Primi N, Menaga D, Ashpin Pabi DJ, Veerakumar S, Kumar BR. 2024. Molecular biomarkers for personalized diagnosis and treatment of gastric cancer using deep learning techniques. 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI); Chennai, India
5. Zhang H, Xiong X, Cheng M, Ji L, Ning K. 2024. Deep learning enabled integration of tumor microenvironment microbial profiles and host gene expressions for interpretable survival subtyping in diverse types of cancers. mSystems 9:e01395-24. https://doi.org/10.1128/msystems.01395-24
6. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature New Biol 579:567–574. https://doi.org/10.1038/s41586-020-2095-1
7. Roelands J, Kuppen PJK, Ahmed EI, Mall R, Masoodi T, Singh P, Monaco G, Raynaud C, de Miranda N, Ferraro L, et al. 2023. An integrated tumor, immune and microbiome atlas of colon cancer. Nat Med 29:1273–1286. https://doi.org/10.1038/s41591-023-02324-5
8. Huang H, Ren Z, Gao X, Hu X, Zhou Y, Jiang J, Lu H, Yin S, Ji J, Zhou L, Zheng S. 2020. Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma. Genome Med 12:102. https://doi.org/10.1186/s13073-020-00796-5
9. Acharya D, Mukhopadhyay A. 2024. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. Brief Funct Genomics 23:549–560. https://doi.org/10.1093/bfgp/elae013
10. Nema R. 2024. An omics-based tumor microenvironment approach and its prospects. Rep Pract Oncol Radiother 29:649–650. https://doi.org/10.5603/rpor.102823
11. Wu J, Chen Z, Xiao S, Liu G, Wu W, Wang S. 2024. DeepMolC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification. BMC Genomics 25:1209. https://doi.org/10.1186/s12864-024-11112-5
12. Lanre S, Ekundayo F, Damilare O. 2024. Cancer diagnosis and prognosis using multi-Omics data: a data science and machine learning approach. World J Adv Res Rev 24:2052–2069. https://doi.org/10.30574/wjarr.2024.24.3.3876
13. Marto Hasugian P, Mawengkang H, Sihombing P, Efendi S. 2023. Review of high-dimensional and complex data visualization. 2023 International Conference of Computer Science and Information Technology (ICOSNIKOM); Binjia, Indonesia
14. Yao Y, Ochoa A. 2023. Limitations of principal components in quantitative genetic association models for human studies. Elife 12:e79238. https://doi.org/10.7554/eLife.79238
15. Mannekote A, Davies A, Pinto JD, Zhang S, Olds D, Schroeder NL, Lehman B, Zapata-Rivera D, Zhai C. 2024. Large language models for whole-learner support: opportunities and challenges. Front Artif Intell 7:1460364. https://doi.org/10.3389/frai.2024.1460364

## AUTHOR BIO

**Ziyuan Huang** is a postdoctoral associate at UMass Chan Medical School, working with the Haran research group and collaborating with the Bucci lab, respectively. His research applies artificial intelligence and deep learning methodologies, particularly large language models (LLMs), to advance the understanding of Alzheimer's disease. By integrating these technologies, he aims to analyze disease progression, predict outcomes, and uncover novel insights to inform therapeutic strategies. Specializing in designing and customizing artificial neural networks (ANNs), Dr. Huang employs advanced data processing techniques to tackle challenges in analyzing complex multi-modal biomedical data. As a member of the Microbiology & Microbiome Dynamics AI Hub within the Department of Microbiology at UMass Chan, he bridges deep learning, computational biology, and clinical neuroscience. His interdisciplinary approach leverages LLMs to synthesize insights from literature, clinical data, laboratory results, and imaging studies, with the goal of developing predictive tools for Alzheimer's disease progression and laying the foundation for personalized treatment strategies. He holds a master of science in analytics (2017) and a Ph.D. in data sciences (2023) from Harrisburg University of Science and Technology.