BMC Methods

EDITORIAL

Open Access

Artificial intelligence for omics data analysis

Zeeshan Ahmed[1,2*†], Shibiao Wan[3†], Fan Zhang[4,5†] and Wen Zhong[6†]

Abstract

Recent technological advancements have vastly improved access to high-throughput biological instrumentation, sparking an unparalleled surge in omics data generation. The implementation of artificial intelligence techniques is revolutionizing omics data interpretation. The *BMC Methods* Collection "Artificial intelligence for omics data analysis" will feature novel artificial intelligence approaches leveraging multi-omics data to accelerate discoveries in personalized medicine, disease diagnostics, drug development, and biological pathway elucidation.

Main

In recent years, technological advancements have significantly boosted the accessibility of high-throughput biological instrumentation for researchers. This surge has led to an unprecedented rate of biological data generation, marking the dawn of the big data era [1]. Driven by the aspiration for a comprehensive understanding of biological systems, researchers now routinely conduct omics studies, encompassing genomics, transcriptomics, epigenomics, proteomics, and metabolomics, which generate vast amounts of data that hold crucial information about biological processes and disease mechanisms. However, single omics data alone may sometimes struggle to fully elucidate the complexities of biological phenomena [2]. Therefore, integrating data from multiple omics sources can offer a more comprehensive understanding of biological systems by capturing interactions between different molecular layers. As a result, multi-omics approaches are gaining popularity due to their potential to provide a more holistic view of biological mechanisms or diseases by extracting, analyzing, and interpreting hidden information that single technologies cannot reveal [3].

Artificial intelligence approaches in omics analysis

Traditional statistical modeling has long been the default choice for analyzing and interpreting big data. However, in recent years, artificial intelligence (AI) technology has gained popularity across various fields [1]. This surge in popularity can be attributed to the evolution of data types from traditional structured data to non-structured, semi-structured, and heterogeneous architectures with diverse characteristics. Furthermore, the demand for novel insights into biological mechanisms has raised the standards and requirements for the depth and accuracy of omics analysis.

AI was formally defined at the Dartmouth conference in 1956 [4]. After that, it developed rapidly and it now encompasses a range of techniques, including machine learning (ML) and deep learning (DL), that enable computers to learn from data and make predictions or decisions. Specifically, ML focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit programming. Algorithm selection

[†]Zeeshan Ahmed, Shibiao Wan, Fan Zhang and Wen Zhong contributed equally to this work.

*Correspondence:
Zeeshan Ahmed
zahmed@ifh.rutgers.edu
[1] Department of Medicine / Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School, Rutgers Health, 125 Paterson St, New Brunswick, NJ, USA
[2] Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers Health, 112 Paterson St, New Brunswick, NJ, USA
[3] Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA
[4] Division of Rheumatology, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[5] Center for Health Artificial Intelligence, Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[6] Science for Life Laboratory, Department of Biomedical and Clinical Sciences (BKV), Linköping University, Linköping, Sweden

is therefore pivotal, and they can be categorized into supervised, semi-supervised, and unsupervised [5]. DL, a subset of ML, employs neural networks composed of hidden layers that perform various operations to uncover intricate representations of the data. This approach has significantly improved the performance of classifiers, surpassing that of traditional ML algorithms, particularly in scenarios involving large-scale datasets with high dimensionality [5].

The implementation of AI techniques has certainly revolutionized the way researchers derive insights from omics data. For example, the recent developed genomic language model (gLM), trained on millions of metagenomic scaffolds to learn the latent functional and regulatory relationships between genes, has proven to be a potent and promising method to close the gap between genomic-context and gene sequence-structure–function [6]. In another recent study, MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumors), a ML model based on the XGBoost (eXtreme Gradient Boosting) library, has been effective in differentiating pathological subtypes of renal tumors, using DNA methylation markers identified in large tissue datasets [7].

As previously mentioned, the interpretation of single omics data often falls short in explaining complex biological phenomena comprehensively, making it challenging to meet the growing research expectations. However, by integrating multiple omics datasets, researchers can gain a more comprehensive understanding of biological systems. AI techniques have become instrumental in this regard, allowing researchers to manage the high dimensionality and heterogeneity of multi-omics data. This approach not only uncovers hidden patterns but also facilitates the prediction of biological outcomes, thereby accelerating biomedical research and paving the way for personalized medicine [8]. For example, the recently implemented Molecular Twin, a novel AI platform integrating multi-omics data, has proven to be effective in predicting outcomes for pancreatic adenocarcinoma patients [9].

Challenges and perspectives

The accumulation of a large amount of biomedical data and the integration of multi-omics through AI will inevitably bring huge benefits to research, eventually leading to personalized medicine. However, despite the progress made by AI in various biomedical realms, numerous challenges remain [1]. They include but are not limited to the management and integration of high volume and heterogeneous multi-omics data, the expertise required for implementing AI approaches and interpreting AI-driven insights, and the critical task of maintaining data quality and achieving reliable generalization. More details are provided below.

Heterogeneity, outliers and missing data imputation

Multi-omics data from different high-throughput sources are usually heterogeneous and noisy. Some omics are more prone to generate sparse data than others and some datasets lack a large number of values, which hinders the integration of multiple datasets [10, 11]. Data preprocessing steps, such as normalization, batch correction, missing value imputation, and outliers detection are crucial for ensuring the quality and reliability of omics data analysis results [1].

Interpretability and explainability

AI models, particularly DL models, are often regarded as "black boxes" due to their complex architectures and lack of interpretability [1]. A transparent and explainable AI algorithm is essential to its final clinical translation and application. On March 15, 2024 the Food and Drug Administration (FDA) published the "Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP are Working Together," which represents the FDA's coordinated approach to AI. This paper is intended to complement the "AI/ML Software as a Medical Device Action Plan" and represents a commitment between the FDA's Center for Biologics Evaluation and Research (CBER), the Center for Drug Evaluation and Research (CDER), and the Center for Devices and Radiological Health (CDRH), and the Office of Combination Products (OCP), to drive alignment and share learnings applicable to AI in medical products more broadly (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device).

Overfitting and generalization

Overfitting, where a model performs well on the training data but fails to generalize to new, unseen data, is a common challenge in AI-driven omics data analysis. Techniques such as cross-validation, regularization, and ensemble learning are used to mitigate the risk of overfitting and improve the generalization performance of AI models [5].

Curse of dimensionality

Most multi-omics datasets suffer from the classical 'curse of dimensionality' problem, i.e. having much fewer observation samples than multi-omics features [11].

Computational and storage cost

The use of AI for multi-omics analysis comes with computational and data storage costs. Most algorithms require high computation power and large volumes of storage capacity to save the logs, results, and analysis [1].

Addressing these issues necessitates Findable, Accessible, Intelligent, and Reproducible (FAIR) solutions, designed for users with and without computational background [12]. These solutions should facilitate biomarker discovery and disease prediction with high precision by leveraging both existing and newly generated multi-omics data alongside demographic and clinical information, uncovering insights often overlooked by traditional statistical and bioinformatics methods. For example, the recent introduction of SLIDE (Significant Latent Factor Interaction Discovery and Exploration), an interpretable latent factor regression-based machine learning approach implemented for ubiquitous biological discovery from high-dimensional multi-omics datasets, overcame some of the previous challenges. While most current methods, such as black-box DL approaches or classification/regression techniques, focus primarily on prediction, preventing them from offering insights into actual mechanisms of complex molecular, cellular or organismal phenotype, SLIDE incorporated nonlinear relationships and came with rigorous guarantees regarding identifiability of the latent factors and corresponding inference [13].

Using AI to address research and clinical needs

However, a critical question remains: which AI approach or algorithm is most suitable to address a specific research question or clinical need? The choice of the appropriate AI approach profoundly influences outcome prediction accuracy, biomarker discovery, and stratification of patient heterogeneity. By applying suitable AI techniques, avenues can be opened for broader biomedical research, ultimately leading to personalized interventions and identification of novel treatment targets [3]. The widespread adoption of these advancements holds immense potential for enhancing public health initiatives worldwide.

Acknowledging the importance of this field, the *BMC Methods* Collection "Artificial intelligence for omics data analysis" (https://www.biomedcentral.com/collections/aioda), focuses on publishing innovative AI approaches using multi-omics data to accelerate discoveries in areas like personalized medicine, disease diagnostics, drug development, and biological pathway elucidation. We invite researchers to submit their work in these areas, contributing to the advancement of AI-driven omics data analysis and its applications in various fields of biological and medical research.

Abbreviations

AI	Artificial intelligence
ML	Machine learning
DL	Deep learning
gLM	Genomic language model
CBER	Center for Biologics Evaluation and Research
CDER	Center for Drug Evaluation and Research
CDRH	Center for Devices and Radiological Health
OCP	Office of Combination Products
FAIR	Findable, Accessible, Intelligent, and Reproducible
SLIDE	Significant Latent Factor Interaction Discovery and Exploration

Acknowledgements

Not applicable.

Authors' contributions

ZA, SW, FZ and WZ conceived and drafted the Editorial. All authors read and approved the final manuscript.

Authors' information

Z.A. is an Assistant Professor at the Department of Medicine / Cardiovascular Disease and Hypertension, Rutgers Robert Wood Johnson Medical School (RWJMS) and Core Faculty Member at the Institute for Health, Health Care Policy and Aging Research (IFH), Rutgers, The State University of New Jersey. Dr Ahmed's lab at Rutgers is focused on implementing Artificial Intelligence (AI), Machine Learning (ML), and standard bioinformatics approaches to multi-omics/genomic and phenotypic data for the identification of patterns revealing predictive biomarkers and risk factors to support earlier diagnosis of patients with complex traits.

S.W. is an Assistant Professor and the Assistant Director for Bioinformatics and Systems Biology Core at University of Nebraska Medical Center (UNMC). With more than 13 years of experience in bioinformatics and machine learning, Dr Wan has published > 40 articles in prestigious journals. He is a TPC member for > 20 machine learning related international conferences including IEEE ICTAI and IEEE IAICT and is an IEEE Senior Member.

F.Z. and her lab are at the intersection of the Department of Medicine Division of Rheumatology and the Department of Biomedical Informatics Center for Health Artificial Intelligence at the University of Colorado Anschutz Medical Campus. The Zhang Lab focuses on developing advanced computational AI and statistical methods for single-cell omics to study immune-mediated inflammatory disease pathogenesis for translational medicine.

W.Z. is an Assistant Professor at Linköping University. Her research mainly focuses on the integration of multi-omics, the interplay between genetics and phenotypes, and the development of data-driven strategies/tools for precision medicine. The aim is to investigate the molecular biomarkers for the estimation of disease risks, early diagnosis of disease, stratification of drug treatment response, disease progression monitoring, and the stratification of patients.

Funding

Z.A.: is supported in part by the National Institute on Aging of the National Institutes of Health under award number R33AG068931, and the National Institute on Drug Abuse under award number U01 DA056472-01A1. The funding sources had no role in the design, or the decision to submit the manuscript.

S.W: is supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA036727, the American Cancer Society under award number IRG-22–146-07-IRG, and by the Buffett Cancer Center, which is supported by the National Cancer Institute under award number CA036727, in collaboration with the UNMC/Children's Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group. S.W. also receives funding from the National Institute on Alcohol Abuse and Alcoholism (P50AA030407-5126, Pilot Core grant) and the Nebraska EPSCoR FIRST Award (OIA-2044049). The content is solely the responsibility of the authors and does not necessarily represent the official views from the funding organizations.

F.Z. is supported by the PhRMA Foundation grant, the Arthritis National Research Foundation award, and the University of Colorado Translational Research Scholars Program award.

W.Z. appreciates the support from the Science for Life Laboratory and the Department of Biomedical and Clinical Sciences (BKV) at Linköping University in Sweden. Their work is supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239).

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 26 April 2024 Accepted: 30 April 2024
Published online: 03 May 2024

References

1. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021;49:107739. https://doi.org/10.1016/j.biotechadv.2021.107739.
2. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8):467–84. https://doi.org/10.1038/s41576-019-0127-1.
3. Chen C, Wang J, Pan D, et al. Applications of multi-omics analysis in human diseases. MedComm (2020). 2023;4(4):e315. https://doi.org/10.1002/mco2.315. Published 2023 Jul 31.
4. McCarthy J, Minsky M, Rochester N, Shannon CE. A proposal for the dartmouth summer research project on artificial intelligence. AI Mag. 2006;27(4):12–4.
5. Li R, Li L, Xu Y, Yang J. Machine learning meets omics: applications and perspectives. Brief Bioinform. 2022;23(1):bbab460.
6. Hwang Y, Cornman AL, Kellogg EH, et al. Genomic language model predicts protein co-regulation and function. Nat Commun. 2024;15(1):2880.
7. Rossi SH, Newsham I, Pita S, et al. Accurate detection of benign and malignant renal tumor subtypes with MethylBoostER: an epigenetic marker-driven learning framework. Sci Adv. 2022;8(39):eabn9828.
8. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: tools, advances, and future approaches. J Mol Endocrinol. 2019;2018. https://doi.org/10.1530/JME-18-0055. Published online July 13.
9. Osipov A, Nikolic O, Gertych A, et al. The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. Nature Cancer. 2024;5(2):299–314.
10. Song M, Greenbaum J, Luttrell J IV, Zhou W, Wu C, Shen H, Gong P, Zhang C, Deng H-W. A review of integrative imputation for multi-omics datasets. Front Genet. 2020;11:570255. https://doi.org/10.3389/fgene.2020.570255.
11. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;1(19):3735–46.
12. Ahmed Z. Precision medicine with multi-omics strategies, deep phenotyping, and predictive analysis. Prog Mol Biol Transl Sci. 2022;190:101–25.
13. Rahimikollu J, Xiao H, Rosengart A, et al. SLIDE: significant latent factor interaction discovery and exploration across biological domains. Nat Methods. 2024. https://doi.org/10.1038/s41592-024-02175-z. Advance online publication.

Publisher's Note