applied sciences

MDPI

Editorial

Deep Learning and Machine Learning Applications in Biomedicine

Peiyi Yan [1], Yaojia Liu [1], Yuran Jia [1] and Tianyi Zhao [2,*]

[1] Institute for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150040, China; 23s136222@stu.hit.edu.cn (P.Y.); 120l021211@stu.hit.edu.cn (Y.L.); 23b903057@stu.hit.edu.cn (Y.J.)

[2] School of Medicine and Health, Harbin Institute of Technology, Harbin 150040, China

* Correspondence: zty2009@hit.edu.cn

The rise of omics research, spanning genomics, transcriptomics, proteomics, and epigenomics, has revolutionized our understanding of biological systems. While the development of these technologies offers immense opportunities for exploring biological complexity, the sheer volume and complexity of multi-omics data present significant analytical challenges. In this context, Artificial Intelligence (AI), with its advantages in data processing and learning capabilities, has become a key tool for multi-omics data analysis. The applications of AI have expanded to include various fields such as disease diagnosis, precision medicine, drug discovery, and elucidating pathogenic mechanisms. This paper delves into the latest advancements of AI in the life sciences sector, with a particular emphasis on its application in crucial areas such as genomics, transcriptomics, and proteomics. By analyzing these successful cases, we aim to demonstrate the potential of AI in handling and applying multi-omics data and provide valuable insights and guidance to researchers.

Genomics, as a scientific field studying the genetic blueprint of organisms, is dedicated to decoding the genomes of living entities. The core objectives of this field include understanding genetic variations, gene functions within the genome, and their impact on an organism's morphology, physiology, and disease occurrence. Particularly, deep learning (DL), a branch of Artificial Intelligence, has demonstrated significant potential in deciphering gene regulation, exploring genome structure, and analyzing variation effects. Tools like DeepVariant [1] and Clairvoyante [2] utilize deep learning to analyze DNA sequence data for variation detection, including identifying single nucleotide polymorphisms (SNPs) and structural variations. Compared to traditional methods, deep learning excels in capturing complex dependencies between sequencing reads, thereby enhancing accuracy and efficiency and enabling more precise genetic analysis. In the realm of cancer genomics, deep learning tools are instrumental in assessing the pathogenicity of variations, elucidating the specific effects of particular variations and informing treatment strategies and prognosis [3–5]. This application is particularly pivotal considering the unique genetic landscapes presented by various cancers. The journey towards understanding gene function has often been hindered by the inefficiency of experimental annotation. Researchers are increasingly employing Artificial Intelligence to develop computational tools in functional genomics. This includes the prediction of gene functions [6] and regulatory elements like enhancers and promoters [7,8]. Additionally, deep learning is utilized to mine and predict the functional impact of non-coding variations from large-scale genomic data, aiding in unraveling the complexities of gene regulatory networks and the potential causal mechanisms behind genetic variations [9–11]. Beyond these applications, deep learning also extends its influence into the realm of epigenomics. For instance, tools like DeepCpG [12] and DeepHistone [13] analyze DNA methylation and histone modification patterns, contributing to a more profound understanding of how these epigenetic factors influence gene expression and disease development.

check for updates

Citation: Yan, P.; Liu, Y.; Jia, Y.; Zhao, T. Deep Learning and Machine Learning Applications in Biomedicine. *Appl. Sci.* **2024**, *14*, 307. https://doi.org/10.3390/app14010307

Received: 13 December 2023
Accepted: 28 December 2023
Published: 29 December 2023

With advancements in high-throughput sequencing technologies, the field of transcriptomics has broadened to include both single-cell and spatial resolution studies. The massive scale and complexity of raw transcriptomic data necessitate sophisticated computational algorithms and tools for preprocessing (quality control, dimensionality reduction, and clustering) and downstream analysis. Various renowned software packages, including Seurat [14] and Scanpy [15], offer comprehensive solutions for transcriptomic data analysis and are adept at tasks like data dimensionality reduction, cell clustering, and differential expression analysis. DL efficiently extracts rich, compact features from noisy, heterogeneous, and high-dimensional scRNA-seq data, thus enhancing downstream analysis. Unsupervised learning, employed for data mining and pattern identification in unlabeled data, is widely applied in scRNA-seq for dimensionality reduction and cell clustering [16–19]. In scRNA-seq, a low RNA capture rate frequently leads to dropout issues. Researchers utilize neural network algorithms for data imputation in scRNA-seq, effectively mitigating noise in gene expression profiles [20–22]. It is noteworthy that a significant advantage of DL in scRNA-seq data analysis is its capacity to handle nonlinear relationships between genes. In tasks like batch effect correction [23,24], cell type identification [25], and gene regulatory network [26,27] analysis, DL methods outperform traditional ones in terms of flexibility and efficiency. Deep learning also finds significant application in the field of spatial transcriptomics. Owing to sequencing technology limitations, emerging spatial transcriptomics has not yet achieved single-cell resolution in gene expression detection. Deep learning can be employed to synergistically analyze single-cell and spatial transcriptomic data, addressing this challenge [28,29]. Additionally, deep learning is utilized for spatial domain identification [30], cell–cell communication [31], 3D reconstruction [32], and detecting spatially variable genes [33]. Furthermore, the development of pre-training models such as Genefomer [34] is paving the way for more sophisticated analyses of specific downstream tasks.

Proteomics is one of the leading application domains for AI. Research includes predicting proteins' three-dimensional structures and functions from primary sequences, studying protein interactions, and designing peptides [35]. Natural Language Processing (NLP) and Computer Vision (CV) methods, such as Transformers and Convolutional Neural Networks (CNNs), play a crucial role in the field of proteomics, particularly in protein residue modeling. A notable example is AlphaFold [36], which uses CNNs and RNNs to accurately predict protein spatial structures. DL also excels in identifying proteins' biological functions based on amino acid sequences, aiding in both general and specific protein recognition [37–44]. In peptide research, it has revolutionized traditional methods, such as mass spectrometry, for peptide identification [45–47]. For protein sequence design, the ProtGPT2 model by Ferruz et al. [48] demonstrates DL 's capability in generating biologically consistent sequences. Analyzing post-translational modification (PTM) sites is another critical area where DL, particularly Transformer-based models, effectively classifies and predicts PTMs [49,50]. In the field of pharmacoinformatics, Artificial Intelligence has shown potential in predicting drug targets and drug–protein affinity [51,52]. Lastly, DL has significantly advanced single-cell proteomics analysis, improving proteome coverage and aiding in cell type/state identification from bulk tissue profiles [53].

Advancements in computing and algorithmic technology are broadening the scope of AI in life sciences. This evolution has drastically improved the processing and analysis of biological data, uncovering complex nonlinear correlations within biological systems and offering innovative approaches for disease research and drug development. AI's role in precision medicine is increasingly critical, especially in biomarker discovery, sample classification, and interpreting disease processes. Despite these advances, DL in life sciences faces challenges related to dataset type and size, which affect its effectiveness and present uncertainties. Large-scale datasets demand greater computing power, while factors like model interpretability, data availability, and quality are pivotal, especially in vital areas like medical diagnosis and drug development, where understanding the influence of input features on predictions is crucial for trust in decision making. Future research might focus

on improving algorithm efficiency and model interpretability to overcome these challenges. We are optimistic about the potential of DL in omics data analysis, anticipating that its ongoing development will yield new insights to propel bioinformatics and life science research forward. These advancements are anticipated to deepen our understanding of disease mechanisms and lay essential biological and computational groundwork for the development of future treatments and preventive measures.

Funding: Natural Science Foundation of China (62102116); Interdisciplinary Research Foundation of HIT; Key R&D Program in Heilongjiang Province (2022ZX02C21); National Key R&D Program of China (2022YFC3321103).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987. [CrossRef] [PubMed]
2. Luo, R.; Sedlazeck, F.J.; Lam, T.-W.; Schatz, M.C. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* **2019**, *10*, 998. [CrossRef] [PubMed]
3. Korvigo, I.; Afanasyev, A.; Romashchenko, N.; Skoblov, M.J. Generalising better: Applying deep learning to integrate deleteriousness prediction scores for whole-exome SNV studies. *PLoS ONE* **2018**, *13*, e0192829. [CrossRef] [PubMed]
4. Quang, D.; Chen, Y.; Xie, X.J.B. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **2014**, *31*, 761–763. [CrossRef] [PubMed]
5. Yousefi, S.; Amrollahi, F.; Amgad, M.; Dong, C.; Lewis, J.E.; Song, C.; Gutman, D.A.; Halani, S.H.; Velazquez Vega, J.E.; Brat, D.J.J.; et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **2017**, *7*, 11707. [CrossRef] [PubMed]
6. Peng, J.; Xue, H.; Wei, Z.; Tuncali, I.; Hao, J.; Shang, X.J. Integrating multi-network topology for gene function prediction using deep neural networks. *Brief. Bioinform.* **2021**, *22*, 2096–2105. [CrossRef] [PubMed]
7. Zhang, T.; Li, L.; Sun, H.; Xu, D.; Wang, G. DeepICSH: A complex deep learning framework for identifying cell-specific silencers and their strength from the human genome. *Brief. Bioinform.* **2023**, *24*, bbad316. [CrossRef] [PubMed]
8. Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.-C. iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **2016**, *32*, 362–369. [CrossRef]
9. De La Vega, F.M.; Chowdhury, S.; Moore, B.; Frise, E.; McCarthy, J.; Hernandez, E.J.; Wong, T.; James, K.; Guidugli, L.; Agrawal, P.B.; et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. *Genome Med.* **2021**, *13*, 153. [CrossRef]
10. Wong, A.K.; Sealfon, R.S.; Theesfeld, C.L.; Troyanskaya, O.G. Decoding disease: From genomes to networks to phenotypes. *Nat. Rev. Genet.* **2021**, *22*, 774–790. [CrossRef]
11. Xiao, Y.; Wang, J.; Li, J.; Zhang, P.; Li, J.; Zhou, Y.; Zhou, Q.; Chen, M.; Sheng, X.; Liu, Z.; et al. An analytical framework for decoding cell type-specific genetic variation of gene regulation. *Nat. Commun.* **2023**, *14*, 3884. [CrossRef] [PubMed]
12. Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67.
13. Yin, Q.; Wu, M.; Liu, Q.; Lv, H.; Jiang, R. DeepHistone: A deep learning approach to predicting histone modifications. *BMC Genom.* **2019**, *20*, 3884. [CrossRef] [PubMed]
14. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive integration of single-cell data. *Cell* **2019**, *177*, 1888–1902.e21. [CrossRef] [PubMed]
15. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 15. [CrossRef]
16. Deng, Y.; Bao, F.; Dai, Q.; Wu, L.F.; Altschuler, S.J. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat. Methods* **2019**, *16*, 311–314. [CrossRef]
17. Huh, R.; Yang, Y.; Jiang, Y.; Shen, Y.; Li, Y. Same-clustering: Single-cell aggregated clustering via mixture model ensemble. *Nucleic Acids Res.* **2020**, *48*, 86–95. [CrossRef]
18. Liu, Q.; Wang, D.; Zhou, L.; Li, J.; Wang, G. MTGDC: A multi-scale tensor graph diffusion clustering for single-cell RNA sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3056–3067. [CrossRef]
19. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef]
20. Eraslan, G.; Simon, L.M.; Mircea, M.; Mueller, N.S.; Theis, F.J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **2019**, *10*, 390. [CrossRef]
21. Liu, Q.; Luo, X.; Li, J.; Wang, G. scESI: Evolutionary sparse imputation for single-cell transcriptomes from nearest neighbor cells. *Brief. Bioinform.* **2022**, *23*, bbac144. [CrossRef] [PubMed]

22. Wu, X.; Zhou, Y. GE-Impute: Graph embedding-based imputation for single-cell RNA-seq data. *Brief. Bioinform.* **2022**, *23*, bbac313. [CrossRef] [PubMed]
23. Xiong, L.; Tian, K.; Li, Y.; Ning, W.; Gao, X.; Zhang, Q.C. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **2022**, *13*, 6118. [CrossRef] [PubMed]
24. Xu, Y.; Das, P.; McCord, R.P. SMILE: Mutual information learning for integration of single-cell omics data. *Bioinformatics* **2022**, *38*, 476–486. [CrossRef] [PubMed]
25. Hu, J.; Li, X.; Hu, G.; Lyu, Y.; Susztak, K.; Li, M. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.* **2020**, *2*, 607–618. [CrossRef] [PubMed]
26. Xu, J.; Zhang, A.; Liu, F.; Zhang, X.J. STGRNS: An interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics* **2023**, *39*, btad165. [CrossRef] [PubMed]
27. Zhao, M.; He, W.; Tang, J.; Zou, Q.; Guo, F. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Brief. Bioinform.* **2022**, *23*, bbab568. [CrossRef] [PubMed]
28. Dong, R.; Yuan, G.-C. SpatialDWLS: Accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **2021**, *22*, 145. [CrossRef]
29. Ma, Y.; Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.* **2022**, *40*, 1349–1359. [CrossRef]
30. Dong, K.; Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **2022**, *13*, 1739. [CrossRef]
31. Jin, S.; Guerrero-Juarez, C.F.; Zhang, L.; Chang, I.; Ramos, R.; Kuan, C.-H.; Myung, P.; Plikus, M.V.; Nie, Q. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **2021**, *12*, 1088. [CrossRef] [PubMed]
32. Wang, G.; Zhao, J.; Yan, Y.; Wang, Y.; Wu, A.R.; Yang, C. Construction of a 3D whole organism spatial atlas by joint modelling of multiple slices with deep neural networks. *Nat. Mach. Intell.* **2023**, *5*, 1200–1213. [CrossRef]
33. Hu, J.; Li, X.; Coleman, K.; Schroeder, A.; Ma, N.; Irwin, D.J.; Lee, E.B.; Shinohara, R.T.; Li, M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **2021**, *18*, 1342–1351. [CrossRef] [PubMed]
34. Theodoris, C.V.; Xiao, L.; Chopra, A.; Chaffin, M.D.; Al Sayed, Z.R.; Hill, M.C.; Mantineo, H.; Brydon, E.M.; Zeng, Z.; Liu, X.S.; et al. Transfer learning enables predictions in network biology. *Nature* **2023**, *618*, 616–624. [CrossRef] [PubMed]
35. Zhao, T.; Liu, J.; Zeng, X.; Wang, W.; Li, S.; Zang, T.; Peng, J.; Yang, Y. Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* **2021**, *22*, bbab014. [CrossRef]
36. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef]
37. Zhang, T.; Jia, Y.; Li, H.; Xu, D.; Zhou, J.; Wang, G. CRISPRCasStack: A stacking strategy-based ensemble learning framework for accurate identification of Cas proteins. *Brief. Bioinform.* **2022**, *23*, bbac335. [CrossRef]
38. Kulmanov, M.; Hoehndorf, R. DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics* **2020**, *36*, 422–429. [CrossRef]
39. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2019**, *35*, 2757–2765. [CrossRef]
40. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227. [CrossRef]
41. Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Computing D: Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* **2018**, *117*, 212–217. [CrossRef]
42. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2017**, *384*, 135–144. [CrossRef]
43. Guo, X.; Tiwari, P.; Zhang, Y.; Han, S.; Wang, Y.; Ding, Y. Medicine: Random Fourier features-based sparse representation classifier for identifying DNA-binding proteins. *Comput. Biol. Med.* **2022**, *151*, 106268. [CrossRef] [PubMed]
44. Liu, Y.; Guan, S.; Jiang, T.; Fu, Q.; Ma, J.; Cui, Z.; Ding, Y.; Wu, H. Medicine: DNA protein binding recognition based on lifelong learning. *Comput. Biol. Med.* **2023**, *16*, 107094.
45. Gao, Y.; Gao, Y.; Fan, Y.; Zhu, C.; Wei, Z.; Zhou, C.; Chuai, G.; Chen, Q.; Zhang, H.; Liu, Q. Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nat. Mach. Intell.* **2023**, *5*, 236–249. [CrossRef]
46. Liu, K.; Ye, Y.; Li, S.; Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nat. Commun.* **2023**, *14*, 7974. [CrossRef]
47. Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007–4016. [CrossRef]
48. Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **2022**, *13*, 4348. [CrossRef]
49. Li, W.; Li, G.; Sun, Y.; Zhang, L.; Cui, X.; Jia, Y.; Zhao, T. Prediction of SARS-CoV-2 Infection Phosphorylation Sites and Associations of these Modifications with Lung Cancer Development. *Curr. Gene Ther.* **2023**. [CrossRef]
50. Yu, K.; Zhang, Q.; Liu, Z.; Du, Y.; Gao, X.; Zhao, Q.; Cheng, H.; Li, X.; Liu, Z.-X. Deep learning based prediction of reversible HAT/HDAC-specific lysine acetylation. *Brief. Bioinform.* **2020**, *21*, 1798–1805. [CrossRef]

51. Ding, Y.; Tang, J.; Guo, F.; Zou, Q. Identification of drug–target interactions via multiple kernel-based triple collaborative matrix factorization. *Brief. Bioinform.* **2022**, *23*, bbab582. [CrossRef] [PubMed]
52. Li, Y.; Qiao, G.; Wang, K.; Wang, G. Drug–target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* **2022**, *23*, bbab346. [CrossRef] [PubMed]
53. Wang, F.; Yang, F.; Huang, L.; Li, W.; Song, J.; Gasser, R.B.; Aebersold, R.; Wang, G.; Yao, J. Deep domain adversarial neural network for the deconvolution of cell type mixtures in tissue proteome profiling. *Nat. Mach. Intell.* **2023**, *5*, 1236–1249. [CrossRef]