algorithms

MDPI

Editorial

Advanced Research on Machine Learning Algorithms in Bioinformatics

Roberto Pagliarini * and Carla Piazza

Department of Mathematics, Computer Science, and Physics, University of Udine, Via delle Scienze, 206, 33100 Udine, Italy; carla.piazza@uniud.it
* Correspondence: roberto.pagliarini@uniud.it

Epigenetic variation and somatic mutations represent molecular components of biodiversity that directly link the genome to the environment. Epigenetics emerged as a promising aspect for the diagnosis of several disorders [1]. It could become an opportunity to uncover new mechanisms as well as therapeutic targets for cancer and analyze their links with metabolic dysregulation [2]. The application of Machine Learning (ML) and Automated Reasoning (AR) techniques to mutational studies composed of huge amounts of multi-omics data could significantly boost discovery and therapy development [3–6].

This Special Issue contains the latest research on the development and application of Machine Learning and artificial intelligence methods to this kind of problem. In the paper entitled *"TMP-M2Align: A Topology-Aware Multiobjective Approach to the Multiple Sequence Alignment of Transmembrane Proteins"*, the authors introduce a multiobjective evolutionary aligner that explicitly models transmembrane protein topology. It uses predicted topology to apply region-specific substitution matrices and stronger gap penalties within TM segments, and optimizes a topology-aware Sum-of-Pairs score plus an Aligned Regions score that rewards topological consistency. On BAliBASE RefSet 7 and GPCR families, the method improves SP and Total Column scores and preserves helical continuity better than several standard and TM-aware aligners. The main caveat is its dependence on the topology-prediction accuracy and comparisons are limited to only some TM tools due to access constraints. The paper *"Enhancing Ferroptosis-Related Protein Prediction Through Multimodal Feature Integration and Pre-Trained Language Model Embeddings"* proposes a multi-modal classification pipeline for predicting ferroptosis-related proteins by combining sequence-derived features, physicochemical descriptors, and embeddings from pre-trained protein language models. A feature-fusion strategy and downstream classifier yield improved predictive performance over baseline feature sets. The study demonstrates that language-model embeddings add complementary information that boosts sensitivity and specificity on curated ferroptosis datasets, but model interpretability and generalization to unseen species remain open issues noted by the authors. The authors of the manuscript entitled *"Enhanced Viral Genome Classification Using Large Language Model"* apply Large Language Model (LLM) embeddings and downstream classifiers to the problem of viral genome classification, showing that sub-word/token-level representations from LLMs, adapted to nucleotide sequences, achieve higher accuracy and robustness to sequencing errors than traditional *k*-mer or alignment-based features. They evaluate these classifiers on multiple viral datasets and report superior macro/micro F1 scores, while also discussing the computational cost and the need for careful tokenization and length management for long genomes as practical limitations. The paper *"A Comparative Study of Machine Learning Techniques for Cell Annotation of scRNA-Seq Data"* is a benchmarking study that evaluates multiple ML approaches for single-cell RNA-sequencing cell-type annotation. The work

check for updates

Received: 8 December 2025
Accepted: 10 December 2025
Published: 16 December 2025

Citation: Pagliarini, R.; Piazza, C. Advanced Research on Machine Learning Algorithms in Bioinformatics. *Algorithms* **2025**, *18*, 794. https://doi.org/10.3390/a18120794

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

highlights that ensemble and transfer-learning approaches often outperform single models, with the performance depending strongly on the reference dataset quality and batch effects. Simpler models can match complex ones when feature selection and normalization are well done. The authors stress reproducible benchmarking and careful pre-processing as key determinants of real-world performance. *"A Novel Model for Noninvasive Haemoglobin Detection Based on Visibility Network and Clustering Network for Multi-Wavelength PPG Signals"* introduces a two-stage model using visibility-graph features and a clustering network applied to multi-wavelength photoplethysmography (PPG) signals to non-passively estimate haemoglobin concentrations. The method combines graph-based signal descriptors with a learned clustering representation to improve robustness to noise and physiological variability, showing promising RMSE and bias compared to baseline PPG estimators; the limitations include the dataset size and the need for broader clinical validation. The approach developed in *"CSpredR: A Multi-Site mRNA Subcellular Localization Prediction Method Based on Fusion Encoding and Hybrid Neural Networks"* integrates multiple encodings with a hybrid neural network architecture to predict localization across several cellular compartments. Evaluations on multi-site localization benchmarks show improved precision and recall relative to prior methods, and the authors discuss how feature fusion and attention mechanisms help capture signals for multi-compartment targeting, while acknowledging the remaining challenges in dynamic localization and condition-specific behavior. The authors of *"Three-Way Alignment Improves Multiple Sequence Alignment of Highly Diverged Sequences"* demonstrate that three-way alignment strategies can substantially improve alignment accuracy for highly diverged sequence sets where pairwise heuristics fail. Experiments indicate gains in alignment quality metrics for divergent protein families and show that three-way guided progressive schemes better preserve conserved motifs. Trade-offs include extra computational cost and the need to integrate three-way scores into existing progressive frameworks.

Taken together, the papers of this Special Issue illustrate a clear trend: specialized, hybrid representations plus task-aware algorithms yield measurable gains across biological sequence and biomedical signal tasks. In sequence analysis, embedding domain constraints into alignment objective functions or alignment strategies improves biological plausibility and standard accuracy metrics. In predictive modelling, the combination of modern representation learning with classical features or tailored architectures provides robustness and better discrimination than single-paradigm methods, albeit at higher computational cost and with interpretability/generalization caveats. Finally, graph-informed features fused with learned representations increase robustness to noise. Common limitations are the dependence on upstream predictions or tokenizations, the constrained benchmark breadth, and practical concerns, which all authors consistently acknowledge. We hope this collection of works will inspire further research and contribute to the ongoing application of ML and AR algorithms in bioinformatics.

Funding: Roberto Pagliarini and Carla Piazza are partially supported by the project "National Biodiversity Future Center—NBFC" (project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP G23C22001110007) funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Roberto Pagliarini and Carla Piazza are members of the GNCS group of INdAM.

Acknowledgments: The Guest Editors would like to express their gratitude to the authors who have chosen to publish their articles in this Special Issue of *Algorithms*, as well as to the reviewers whose support, through their evaluation of these manuscripts, allowed us to select only high-quality works.

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

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
AR	Automated Reasoning
LLM	Large Language Model
PPG	Photoplethysmography
RMSE	Root Mean Square Error

List of Contributions

1. Cedeño-Muñoz, J.; Zambrano-Vega, C.; Nebro, A.J. TMP-M2Align: A Topology-Aware Multiobjective Approach to the Multiple Sequence Alignment of Transmembrane Proteins. *Algorithms* **2025**, *18*, 640. https://doi.org/10.3390/a18100640.
2. Zhou, J.; Wang, C. Enhancing Ferroptosis-Related Protein Prediction Through Multimodal Feature Integration and Pre-Trained Language Model Embeddings. *Algorithms* **2025**, *18*, 465. https://doi.org/10.3390/a18080465.
3. Gunasekaran, H.; Wilfred Blessing, N.R.; Sathic, U.; Husain, M.S. Enhanced Viral Genome Classification Using Large Language Models. *Algorithms* **2025**, *18*, 302. https://doi.org/10.3390/a18060302.
4. Wani, S.A.; Quadri, S.; Mir, M.S.; Gulzar, Y. A Comparative Study of Machine Learning Techniques for Cell Annotation of scRNA-Seq Data. *Algorithms* **2025**, *18*, 232. https://doi.org/10.3390/a18040232.
5. Liu, L.; Wang, Z.; Zhang, X.; Zhuang, Y.; Liang, Y. A Novel Model for Noninvasive Haemoglobin Detection Based on Visibility Network and Clustering Network for Multi-Wavelength PPG Signals. *Algorithms* **2025**, *18*, 75. https://doi.org/10.3390/a18020075.
6. Wang, X.; Suo, W.; Wang, R. CSpredR: A Multi-Site mRNA Subcellular Localization Prediction Method Based on Fusion Encoding and Hybrid Neural Networks. *Algorithms* **2025**, *18*, 67. https://doi.org/10.3390/a18020067.
7. Askari Rad, M.; Kruglikov, A.; Xia, X. Three-Way Alignment Improves Multiple Sequence Alignment of Highly Diverged Sequences. *Algorithms* **2024**, *17*, 205. https://doi.org/10.3390/a17050205.

References

1. Rasool, M.; Malik, A.; Naseer, M.I.; Manan, A.; Ansari, S.; Begum, I.; Qazi, M.H.; Pushparaj, P.; Abuzenadah, A.M.; Al-Qahtani, M.H.; et al. The role of epigenetics in personalized medicine: Challenges and opportunities. *BMC Med. Genom.* **2015**, *8*, S5. [CrossRef] [PubMed]
2. Cavalli, G.; Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **2019**, *571*, 489–499. [CrossRef] [PubMed]
3. Liò, P.; Zuliani, P. (Eds.) *Automated Reasoning for Systems Biology and Medicine*; Computational Biology; Springer: New York, NY, USA, 2019; Volume 30. [CrossRef]
4. Dos Martires, P.Z.; Derkinderen, V.; De Raedt, L.; Krantz, M. Automated reasoning in systems biology: A necessity for precision medicine. In Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR '24, Hanoi, Vietnam, 2–8 November 2024. [CrossRef]

5. Pagliarini, R.; Podrini, C. Metabolic Reprogramming and Reconstruction: Integration of Experimental and Computational Studies to Set the Path Forward in ADPKD. *Front. Med.* **2021**, *8*, 740087. [CrossRef] [PubMed]
6. Srinivasa, K.G.; Siddesh, G.M.; Manisekhar, S. *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*; Springer: Berlin/Heidelberg, Germany, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.