Guest Editorial Advanced Machine Learning and Artificial Intelligence Tools for Computational Biology: Methodologies and Challenges

IN RECENT years, the management and analysis of biological data have experienced exponential growth propelled by the relentless advancement of machine learning (ML) and artificial intelligence (AI) technologies. This is driven mainly by the remarkable ability and potentials of AI-based systems to craft sophisticated, yet effective, algorithms and analytical models tailored for the interpretation of biological information; thus, assist in making accurate predictions and/or decisions [1]. The surge in AI adoption is not unfounded; it's a response to the overwhelming increase in both the volume and acquisition rates of biological data.

Within the computational biology and bioinformatics research communities, conventional analysis strategies lack the strong potential to grapple with the demands of Big Data, often resulting in flawed practices and inaccurate conclusions, i.e., incorrect practices. Advanced AI tools, on the other hand, possess high potential to develop algorithms and analytical models for interpreting biological information. Particularly, advanced deep learning algorithms, which delve into biological data with unparalleled depth, unraveling hidden structures and features that elude conventional methodologies. These tools serve as catalysts for breakthroughs, furnishing researchers with profound insights crucial for accurate predictions and informed decision-making.

The primary objective of this special issue is to serve as a beacon of illumination, offering researchers worldwide access to cutting-edge research work for the best utilization of AI tools in computational biology and bioinformatics research. The focus is on (1) showcasing compelling and recent AI-based research addressing emerging challenges in computational biology, and (2) presenting original and applicable innovative work/studies aimed at understanding, visualizing, and interpreting biomedical data for biological insights. This research topic consists of eight meticulously curated manuscripts, each is a testament to the optimal utilization of AI tools in computational biology and bioinformatics research. A brief synopsis of each article's contribution, findings as well as limitations are provided below.

Naseem, et al. [2] introduces K-PathVQA, a pioneering approach in Pathology Visual Question Answering (PathVQA) that integrates external medical knowledge to enhance answer inference from pathology images. Unlike previous methods relying solely on image analysis, K-PathVQA leverages a medical knowledge graph (KG) to augment question representations and learn knowledge-image-question interactions jointly. Through experiments on a PathVQA dataset, K-PathVQA outperformed baseline methods, achieving a 4.15% increase in overall accuracy, with notable improvements in both open-ended and closed-ended question types. The study introduces a novel end-to-end trainable method that seamlessly integrates image, question, and knowledge representations, showcasing superior performance and generalizability across multiple medical VQA datasets. However, further research is warranted to assess its robustness across diverse pathology imaging datasets and clinical scenarios, as well as to explore the scalability and computational efficiency for broader deployment in clinical settings. The third paper by Sharma et al. [3] introduces a framework, integrating deep learning algorithms through ensemble techniques, to address the challenge of identifying low hemolytic therapeutic peptides, which is critical for drug development. The framework, trained and tested on recent datasets, significantly outperforms existing in-silico tools, particularly in accurately classifying peptides with N/C terminal modifications. Additionally, the model offers advanced functionalities like mutation analysis and residue scans, enhancing researchers' capabilities for peptide analysis. Ablation studies reveal the essential contributions of ensemble algorithms and handcrafted features alongside deep learning-based features in achieving high performance, providing valuable insights for further optimization. Furthermore, the deployment of the developed model as a web server enhances accessibility and usability for researchers, facilitating broader adoption and application of the proposed methodology. However, the proposed framework's static nature poses a limitation, as it may become outdated with the rapid increase in peptide data. Future improvements could explore the concept of continual learning to make the framework adaptive and resilient to evolving datasets. Additionally, while the framework excels in classifying N/C terminal-modified peptides, further research may be needed to investigate other peptide modifications beyond N-terminal acetylation and C-terminal amidation, thus expanding its applicability in peptide analysis.

Identifying pivotal T-cell markers that is crucial for predicting disease progression and immune state modulation in the context of Rheumatoid Arthritis (RA) was studied in [4] by Talitckii et al. By leveraging machine learning and feature selection algorithms on a comprehensive dataset obtained from a mouse model of collagen-induced arthritis, the study aims to distil a lower-dimensional subset of T cell markers that accurately forecast disease outcomes and treatment efficacy. The introduced algorithms offer a robust and versatile framework applicable

Date of current version 5 April 2024.
Digital Object Identifier 10.1109/JBHI.2024.3372662

to similar datasets, facilitating insights into the self-nonself determination process vital in autoimmune diseases. Notably, the identified markers shed light on the complex dynamics of the immune response and represent valuable assets in tailoring personalized immunotherapy approaches. While the paper acknowledges potential avenues for further exploration, such as the inclusion of additional biomarkers, its open-source algorithms foster collaboration and propel advancements in understanding immune system dynamics and therapeutic interventions. This research underscores the significance of harnessing machine learning methodologies in unravelling the complexities of autoimmune diseases and advancing precision medicine strategies in immunotherapy.

To predict methylation sites using graphs, Gu et al. developed a hybrid Bayesian optimization-based graphical approach [5]. The paper addresses the crucial need for accurate prediction of protein methylation sites (PMSP) to advance understanding of gene expression regulation and disease treatments. Motivated by the limited use of topological information in existing computational methods, the study introduces GraphMethySite, an innovative framework leveraging graph convolutional networks and Bayesian Optimization (BO) to predict methylation sites based on three-dimensional spatial protein structures. Furthermore, the extension of GraphMethySite, termed GraphMethySite+, incorporates hybrid Bayesian Optimization to automatically optimize discrete and continuous parameters defining the graphical structure, enhancing predictive accuracy. Empirical results on benchmark datasets demonstrate the superiority of GraphMethySite over existing methods, highlighting the potential of incorporating topological information in PMSP prediction. However, the study acknowledges limitations, including the computational complexity of predicting protein structures and the incomplete understanding of the biological functions of identified methylation sites, emphasizing the importance of integrating computational predictions with experimental validations to obtain comprehensive insights into protein methylation biology.

In the context of proteins interactions, Lin et al. [6] studied the challenging task of predicting interactions between proteins by bridging chemical and geometric features in protein surface analysis. Recognizing the importance of hierarchical relationships among atoms and the interaction between chemical and geometric features, the study introduces Hierarchical Chemical and Geometric Feature Interaction Network (HCGNet). HCGNet utilizes deep learning techniques to learn chemical and geometric features in a hierarchical and interactive manner, capturing multiscale relationships and interactions crucial for effective protein surface analysis. Experimental results demonstrate that HCGNet outperforms prior state-of-the-art methods in both site prediction and interaction matching tasks, achieving a significant improvement of 2.3% and 3.2%, respectively. The key contributions of the paper include highlighting the significance of hierarchical relationships and interactions in protein surface learning, proposing HCGNet as a novel framework for effective protein surface analysis, and demonstrating its superior performance over existing methods. However, the study acknowledges that the hyperparameters and model design of HCGNet may require specific tuning for different downstream applications, and further research is needed to develop a unified framework and collect extensive data for general biomolecular surface learning, representing a limitation of the current work.

For a study of long non-coding RNAs (lnvRNAs), Wadhera et al. [7] presented a framework for protein surface analysis. They introduced a LncRNA-Protein Interactions based on Kernel Combinations and Graph Convolutional Networks (LPI-KCGCN), aimed at predicting LncRNA-Protein Interactions (LPI) with improved accuracy and efficiency. LPI-KCGCN integrates kernel learning and Graph Convolutional Networks (GCN) to extract features from lncRNAs and proteins, leveraging various sequence features, sequence similarity features, expression features, and gene ontology. The framework constructs kernel matrices and similarity matrices to capture the topology of the LPI network, which are then utilized in a two-layer GCN to extract potential representations in the lncRNA and protein space. The predicted matrix is obtained through training the network to produce scoring matrices for lncRNAs and proteins. The numerical results demonstrate the effectiveness of LPI-KCGCN, achieving impressive performance with AUC values of 0.9714 and 0.9907 and AUPR values of 0.9216 and 0.9267 on balanced and highly unbalanced datasets, respectively, outperforming state-of-the-art methods. Additionally, the framework maintains satisfactory results on unbalanced datasets and exhibits low complexity, offering advantages in terms of timesaving and reduced biological experiment requirements. However, a limitation of the work is not explicitly mentioned, such as potential challenges in generalizing the framework to other biological interaction prediction tasks or dealing with noisy or incomplete data. While the proposed framework, LPI-KCGCN, shows promising results, its generalizability to other biological interaction prediction tasks needs further validation. Additionally, reliance on computational methods may introduce biases, and the performance may vary based on data quality and completeness. Interpretation of predicted interactions without experimental validation poses a challenge, emphasizing the need for experimental studies. Moreover, the framework's applicability may be limited by the specificity of the datasets used.

Prediction of Minimum Inhibitory Concentration (MIC) values was investigated in [8] by Sharma et al. The paper introduces ESKAPEE-MICpred, a pioneering model designed to predict Minimum Inhibitory Concentration (MIC) values against the ESKAPEE group of bacteria for Antibacterial Peptides (ABPs), thus filling a crucial gap in antimicrobial research. By leveraging transfer learning and ensemble techniques, the model offers a streamlined approach to rapidly screening potential ABPs, aiding wet-lab researchers in identifying effective antimicrobial agents. The findings demonstrate ESKAPEE-MICpred's efficacy, as evidenced by a high Pearson correlation coefficient exceeding 0.8 between predicted and actual MIC values. This robust performance underscores the model's reliability and utility in guiding experimental efforts. Additionally, the deployment of ESKAPEE-MICpred as a web server enhances accessibility, facilitating widespread use and contributing to efforts to combat antimicrobial resistance. However, the study acknowledges the need for ongoing improvements, particularly in dataset quantity and quality, to enhance the model's accuracy and relevance. Addressing these limitations will further optimize ESKAPEE-MICpred's potential to address the urgent global challenge of antimicrobial resistance.

Finally, a study by Yi et al. [9] for diabetic retinopathy (DR) The introduced a novel approach for segmenting microvascular lesions using a compound scaling encoder-decoder network. The proposed method focuses on addressing the challenges associated with segmenting small and sparse lesions. The method incorporates a lightweight encoder architecture with a compound scaling coefficient and an attention-based decoder with spatial and excitation blocks. It also proposes a compound loss function and employs transfer learning to address class imbalance and data scarcity issues. Overall, the author's pipeline integrating compound scaling encoder-decoder network, attention mechanisms, compound loss, and transfer learning, offers a promising approach for accurately segmenting microvascular lesions in DR. However, the study has some limitations including the reliance on diverse pre-trained models for transfer learning, uncertainties in generalization to varied clinical settings, potential computational resource demands, and challenges in interpreting model decisions. Addressing these limitations will be crucial for advancing the practical applicability and robustness of the proposed method in clinical practice.

Collectively, the eight papers within this Special Issue address diverse yet highly pertinent topics in computational pathology, presenting cutting-edge research aimed at maximizing the effective integration of AI tools in computational biology and bioinformatics research. We anticipate that this Special Issue will significantly enhance awareness within the scientific community by showcasing advancements, novel technologies, and practical applications of AI-based tools in computational biology. We extend our gratitude to all the authors for their valuable contributions to this special issue. Furthermore, we acknowledge the invaluable input of numerous experts who participated in the rigorous review process, offering constructive feedback to enhance the quality and clarity of the articles. Special thanks are extended to the Editor-in-Chief, Professor Dimitrios I. Fotiadis, and the publishing team for their unwavering support and insightful guidance throughout the culmination of this special issue.

FAHMI KHALIFA, *Guest Editor*
Electrical and Computer Engineering Department
School of Engineering, Morgan State University
Baltimore MD 21251, USA
e-mail: fahmi.khalifa@morgan.edu

IMRAN RAZZAK, *Guest Editor*
School of Computer Science and Engineering
University of New South Wales
Sydney, 2052, Australia
e-mail: imran.razzak@unsw.edu.au

MOHAMMAD AMJAD KAMAL, *Guest Editor*
Joint Laboratory of Artificial Intelligence in Healthcare
Institutes for Systems Genetics and West China School of Nursing
Frontiers Science Center for Disease-related Molecular Network
West China Hospital
Sichuan University
Chengdu, China

King Fahd Medical Research Center
King Abdulaziz University
Jeddah 21589, Saudi Arabia

Department of Pharmacy
Faculty of Health and Life Sciences
Daffodil International University
Dhaka-1216, Bangladesh

Centre for Global Health Research
Saveetha Medical College and Hospital
Saveetha Institute of Medical and Technical Sciences
Chennai, Tamil Nadu, India

Novel Global Community Educational Foundation
Hebersham, NSW 2770
Australia
e-mail: rrs.USA.au@gmail.com

AHMED SOLIMAN, *Guest Editor*
Faculty of Engineering
Mansoura University
Mansoura 35516, Egypt
e-mail: ahmedsoliman@mans.edu.eg

REFERENCES

[1] I. Razzak, M. K. Khan, G. Xu, and F. Khalifa, "Guest editorial open and interpretable AI in computational pathology," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1657–1660, Apr. 2023.

[2] U. Naseem, M. Khushi, A. G. Dunn, and J. Kim, "K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering," *IEEE J. Biomed. Health Inform.*, early access, Jul. 11, 2023, doi: 10.1109/JBHI.2023.3294249.

[3] R. Sharma, S. Shrivastava, S. K. Singh, A. Kumar, A. K. Singh, and S. Saxena, "EnDL-HemoLyt: Ensemble deep learning-based tool for identifying therapeutic peptides with low hemolytic activity," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1691–1700, 2024.

[4] A. Talitckii, J. L. Mangal, B. K. Colbert, A. P. Acharya, and M. M. Peet, "Employing feature selection algorithms to determine the immune State of a mouse model of rheumatoid arthritis," *IEEE J. Biomed. Health Inform.*, early access, Oct. 26, 2023, doi: 10.1109/JBHI.2023.3327230.

[5] L.-Y. Gu et al., "Hybrid Bayesian optimization-based graphical discovery for methylation sites prediction," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1701–1708, 2024.

[6] Y. Lin, L. Pan, Y. Li, Z. Liu, and X. Li, "Exploiting hierarchical interactions for protein surface learning," *IEEE J. Biomed. Health Inform.*, early access, Jan. 19, 2024, doi: 10.1109/JBHI.2024.3356231.

[7] C. Shen, D. Mao, J. Tang, Z. Liao, and S. Chen, "Prediction of LncRNA-protein interactions based on kernel combinations and graph convolutional networks," *IEEE J. Biomed. Health Inform.*, early access, Jun. 16, 2023, doi: 10.1109/JBHI.2023.3286917.

[8] R. Sharma, S. Shrivastava, S. K. Singh, A. Kumar, A. K. Singh, and S. Saxena, "Artificial intelligence-based model for predicting the minimum inhibitory concentration of antibacterial peptides against ESKAPEE pathogens," *IEEE J. Biomed. Health Inform.*, early access, Apr. 28, 2023, doi: 10.1109/JBHI.2023.3271611.

[9] D. Yi, P. Baltov, Y. Hua, S. Philip, and P. K. Sharma, "Compound scaling encoder-decoder (CoSED) network for diabetic retinopathy related biomarker detection," *IEEE J. Biomed. Health Inform.*, early access, Sep. 11, 2023, doi: 10.1109/JBHI.2023.3313785.