Corresponding Author:
dr.syedrizwan@kiu.edu.pk

DARSGAH-E-AHLEBAIT
Journal of biotechnological sciences [ISSN: 2308-0043]

Chemo-informatics Challenges: Tackling Big Data for Drug Discovery

Syeda Sana Urooj[1] and Syed Rizwan Abbas[2]

[1]Department of Animal Sciences; Karakorum International University; Gilgit
[2]Department of Plant Sciences; Karakorum International University; Gilgit.

ABSTRACT

The combination of cutting-edge data analytics and revolutionary bioinformatics has tremendously transformed the drug discovery landscape in the recent past. This has fundamentally impacted the ability to capitalize on unique opportunities to boost the speed and efficiency of developing new treatments. However, the overflow in data generation has brought together major challenges in the data management sector i.e. collection, storage, processing, and interpretation of data. This review aims to explore the inherent challenges in bioinformatics specifically focusing on dealing with large-scale data sets for drug development. Furthermore, we would also dig into the present-day advances to address these issues such as the application of Artificial intelligence (AI), as a means. Utilizing AI technologies, researchers can easily and effectively navigate the rising problems of big data, eventually cracking its ultimate potential to speed up the drug discovery process. Thus, this review overall highlights the link between bioinformatics and artificial intelligence in revolutionizing drug development, emphasizing the necessity for groundbreaking solutions to harness the wealth of data available in the modern biomedical landscape. Through an inclusive analysis, we aim to shed light on the present state of bioinformatics in drug discovery and suggest pathways for future research and development in this rapidly growing field.

KEYWORDS: Bioinformatics Challenges; Big data; Tackling; Drug discovery

INTRODUCTION

Bioinformatics is an interdisciplinary science that has substantially transformed biological data handling, analysis, and interpretation by blending computer science, statistics, and biology (Iqbal & Kumar, 2023). Bioinformatics has advanced greatly since its inception in the 1980s, especially in the healthcare industry, where it is essential for comprehending genomic data and making it available for designed medication and early for disease detection (Tran, Thi,

& Chu, 2024). Bioinformatics aims to study molecular data to help with personalized medicine and early disease diagnosis. This comprises finding genetic changes, comprehending disease causes, and evolving treatment advances (Ogunjobi *et al.,* 2024; Viveka *et al.,* 2024).

Bioinformatics has advanced remarkably because of the complexity of biological systems and the necessity to manage enormous volumes of biological data (Viveka *et al.,* 2024). Two major turning points in the journey have been the invention of DNA sequence technology and the completion of the human genome project in 2003 (Taylor *et al.,* 2024). Bioinformatics covers a broad spectrum of fields, from proteomics and genomics to clinical information and artificial intelligence (AI) (Chafai, Bonizzi, Botti, & Badaoui, 2024). Bioinformatics tools and techniques are essential to the drug discovery process because they make it possible to identify new drug candidates, make it easier to navigate vast and complex data landscapes, and streamline the drug development process (Tripathi, 2024).

New directions in drug development have been made possible by the emergence of big data, which is produced by high-throughput technologies like mass spectrometry and next-generation sequencing (NGS) (Domrazek & Jurka, 2024). But to fully utilize big data in drug development, quite a few bioinformatics issues must be resolved at every stage of the data lifecycle, from collection to interpretation (Carreras-Puigvert & Spjuth, 2024). Big data from multi-omics experiments, such as genomics, transcriptomics, proteomics, metabolomics, phenomics, glycomics, and epigenomics, is referred to as biological big data. This type of data has the potential to significantly improve our understanding of biology and lead to the development of novel disease treatments (Mukherjee, Abraham, Singh, Balaji, & Mukunthan, 2024). Nonetheless, these data pose significant challenges for storage and interpretation due to their enormous volume and complexity (Chaudhari, Pant, Jha, Pathak, & Singh, 2024).

Regardless of these hurdles, the remarkable developments in artificial intelligence and biological sciences have proven crucial in finding solutions to several difficult problems, including those in healthcare. Due to the exponential growth of data, advances in data storage and retrieval technologies, and computer power, several fields of Artificial intelligence (AI) have been demonstrably impacted (Zhou, Shen, He, Weng, & Chen, 2024). These developments have sparked the creation of data-driven AI-based methods like deep learning (DL) that are revolutionizing the sector. Data advancement necessitates an understanding of its limitations, storage, problems, and applications. For this reason, bioinformatics is turning into a vital aspect of present-day research and healthcare (Nguyen, Tran, & Khanh Le, 2024).

CHALLENGES IN MANAGING BIOLOGICAL BIG DATA FOR DRUG DISCOVERY

DATA STORAGE AND VISUALIZATION

The storage and visualization of this huge amount of data creates another trouble in the rapidly evolving landscape of cutting-edge technologies (Shukla, Muhuri, & Abraham, 2020). The volume of data for storage, processing, and analysis is expected to expand significantly. This necessitates specialized analysis software for studying and comparing data and tools for modeling, visualizing, exploring, and interpreting data (S. V. Singh, 2024). These resources support the analysis, structural, and functional characterization of biomolecules, contributing to advancing fields such as Genomics, Proteomics, Transcriptomics, and metabolomics. Big data produces a huge volume of unstructured data that is multiplex and larger than what traditional databases can handle (Magro, Venezia, & Balistreri, 2024). A change in computing infrastructure

is essential to manage the demands of data storage and the intensive server processing required for the safe analysis of enormous volumes of data. Furthermore, for researchers and clinicians to make correct decisions, data visualization tools that can successfully convey complex information to them are essential (S. V. Singh, 2024).

DATA GENERATION AND TRANSFER

Generating and transferring biomedical data or "big data" causes significant hurdles, but is becoming more and more accessible. For example, next-generation sequencing (NGS) technologies produce enormous amounts of data that need advanced computational tools to analyze and manage safely (Demirbaga, Aujla, Jindal, & Kalyon, 2024). This includes the requirement for more sophisticated encryption and de-identification techniques in security systems to protect patient privacy when data is transferred—currently, this is mainly accomplished by mailing external drives comprising the data (Demirbaga *et al*., 2024).

DATA VOLUME AND STORAGE

One of the main hurdles in drug discovery is the massive amount of data produced by high-throughput technologies like mass spectroscopy and next-generation sequencing (NGS) (Schmidt & Hildebrandt, 2017). Biological big data includes a wide range of data from multi-omics experiments, including transcriptomics, metabolomics, glycomics, genomics, proteomics, epigenomics, phenomics, and other omics. This data is crucial for studying biological processes, gaining insight into the functioning of living systems, developing new disease treatments, and understanding the causes of data on various conditions (Fondi & Liò, 2015). However, storing and analyzing this huge amount of data presents challenges due to its complexity and sheer size, requiring substantial storage space and processing power. Furthermore, the complexity of multi-omics data limits insights that can be derived. Nevertheless, biological big data holds great promise for advancing our understanding of biology and developing new treatments for diseases (Odenkirk, Reif, & Baker, 2021). However, the field of big data research is swiftly growing with numerous applications, as the volume of data continues to grow, it is important to understand storage, functionality, limitations, and challenges. Due to exponential growth in data volume drug discovery applications also require scalable and reliable storage systems. This gigantic volume of data must be stored efficiently, necessitating sophisticated data management systems that ensure data integrity, accessibility, and long-term preservation (Chaudhari *et al.,* 2024).

DATA INTEGRATION AND SECURITY

Firmly integrating clinical data with omics data, administrative data, and financial information poses another significant problem. To ensure the privacy and security of patient information during data storage and transfer, an integrated system that can manage the complexity of these many kinds must be developed (Shah & Khan, 2020). There are issues with data security and privacy associated with cloud computing platforms like Amazon Web Services, although they can offer the elastic scalability required for DNA sequencing (Tyagi, Sachan, & Kumari, 2024).

DATA QUALITY AND NOISE

Upholding high data quality is critical for accurate data analysis and interpretation in bioinformatics and drug development (Tantoso *et al*., 2019). While high-throughput methods offer the capacity to process an enormous array of data, they also introduce noise and errors such as sample contamination, batch effects, and sequencing errors (Čuklina *et al*., 2021). These

factors can considerably impact data integrity. Substandard data quality can have a wide range of implications, impeding the accurate identification of potential therapeutic options, destabilizing the reproducibility of research findings, and leading to inaccurate and potentially detrimental conclusions. Consequently, stringent quality control measures are essential throughout the data analysis process to address these challenges and ensure the reliability and validity of the results (Guo *et al*., 2024).

DATA INTEGRATION AND HETEROGENEITY

In drug development, bioinformatics data often comes from numerous sources and types of omics research, resulting in considerable heterogeneity (Matthews, Hanison, & Nirmalan, 2016). Integrating this disparate information into a coherent and complete understanding of biological processes and possible therapeutic targets is a big problem (Zitnik et al., 2019). The main causes of data heterogeneity are differences in data formats, experimental methods, and the kinds of biological entities studied. Overcoming these inconsistencies is necessary for effective integration to enable efficient analysis and interpretation in the context of drug discovery (Zech & Winkelmann, 2024).

COMPUTATIONAL RESOURCES AND POWER

Large biological data sets require vast computer resources to analyze and handle drug discovery processes (Schneider, 2018). For bioinformatics research, high–performance computing (HPC) resources are mostly needed to manage the demanding computations and data processing activities. Sufficient computational resources must be available to complete bioinformatics investigations accurately and on time (Banimfreg, 2023).

DATA ANALYSIS AND INTERPRETATION

Ensuring great data quality is critical to accurate analysis and interpretation in bioinformatics and drug discovery (Shi, Perkins, Fang, & Tong, 2008). Numerous kinds of noise and mistakes, including sample contamination, batch effects, and sequencing errors, can be introduced by high-throughput methods. Insufficient data quality can affect the identification of possible therapeutic options by negotiating the reproducibility of study findings and resulting in inaccurate conclusions. (Abdi *et al.,* 2024).

THE ROLE OF AI IN BIOINFORMATICS AND DRUG DISCOVERY

Artificial intelligence is transforming bioinformatics and drug development by solving large data challenges (Keedwell & Narayanan, 2005). AI technologies like machine learning, deep learning, and natural language processing make it easier for researchers to handle, integrate, and analyze large and complicated information (Jing, Bian, Hu, Wang, & Xie, 2018). These major developments accelerate the pace of research and development in the biomedical profession, resulting in more personalized and effective therapies (Zhu, 2020).

AI algorithms can analyze genetic and proteomic data to find novel pharmacological targets, predict the efficacy and safety of new drugs, and optimize drug design (Dara, Dhamercherla, Jadav, Babu, & Ahsan, 2022). AI boosts research accuracy and efficiency, decreasing the cost and time associated with bringing pharmaceuticals to the market, by automating and optimizing multiple phases of the drug development pipeline, AI may also model biological processes and disease systems, offering more insight into drug interactions and possible adverse effects (Khan *et al*., 2024).

AI IN TARGET IDENTIFICATION

AI is particularly great at finding new therapeutic targets by analyzing huge amounts of genomic and proteomic data (Mann, Kumar, Zeng, & Strauss, 2021). Genes or proteins that may be suitable targets for therapeutic intervention can be identified using machine learning algorithms, which can analyze enormous datasets to find connections between genetic variants and disease phenotype (Sharma *et al*., 2024).

AI IN DRUG SCREENING

By predicting a compound's biological action, artificial intelligence (AI) may greatly speed up the drug screening process (Álvarez-Machancoses & Fernández-Martínez, 2019). The discovery of potential drug candidates can be streamlined by using deep learning models that have been trained on large chemical and biological datasets to predict the interactions that novel compounds will have with the target protein (S. Singh, Gupta, Sharma, & Sahi, 2024).

AI IN DRUG DESIGN AND OPTIMIZATION

The use of AI-driven algorithms has the potential to transform the process of drug development by permitting the optimization of drug candidates to improve safety and efficacy (Tiwari, Pal, Chaudhary, & Nath, 2023). Researchers can develop novel compounds with specific biological features by leveraging artificial intelligence (AI) (Melo, Maasch, & de la Fuente-Nunez, 2021). This novel approach predicts which medicine will have the most beneficial therapeutic effects and the fewest negative side effects by modeling molecular interactions (Tatonetti, Ye, Daneshjou, & Altman, 2012). As a result, this strategy reduces the need for costly and time-consuming experimental testing, which speeds up the process of bringing novel pharmaceuticals to market. This study reveals the positive impact of AI on drug development, emphasizing the potential for significant breakthroughs in pharmaceutical research and the discovery of new medicines (Loeffler *et al*., 2024).

CONCLUSION

To sum up, bioinformatics is essential for overcoming the challenges that come with using big data in drug discovery. By addressing issues related to data collection, storage, processing, and understanding, scientists can fully leverage big data to expedite the discovery and creation of innovative treatments. Moving drug development towards targeted therapies and precise medicine will require interdisciplinary teams and technology innovations. Bioinformatics must use artificial intelligence (AI) to overcome the difficulties posed by large data. AI will become more crucial as the discipline develops in terms of expanding our knowledge of biology and enhancing patient outcomes. This study has given a comprehensive overview of the state of Artificial intelligence and drug development, as well as its future directions. It has highlighted certain AI tools and methods that are driving these developments.

REFERENCES

Abdi, G., Jain, M., Barwant, M., Tendulkar, R., Tendulkar, M., Tariq, M., & Amir, A. (2024). Unveiling the Dynamic Role of Bioinformatics in Automation for Efficient and Accurate Data Processing and Interpretation Advances in Bioinformatics (pp. 279-319): Springer.

Álvarez-Machancoses, Ó., & Fernández-Martínez, J. L. (2019). Using artificial intelligence methods to speed up drug discovery. Expert opinion on drug discovery, 14(8), 769-777.

Banimfreg, B. H. (2023). A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics. Healthcare Analytics, 100190.

Carreras-Puigvert, J., & Spjuth, O. (2024). Artificial intelligence for high content imaging in drug discovery. Current Opinion in Structural Biology, 87, 102842.

Chafai, N., Bonizzi, L., Botti, S., & Badaoui, B. (2024). Emerging applications of machine learning in genomic medicine and healthcare. Critical Reviews in Clinical Laboratory Sciences, 61(2), 140-163.
Chaudhari, J. K., Pant, S., Jha, R., Pathak, R. K., & Singh, D. B. (2024). Biological big-data sources, problems of storage, computational issues, and applications: a comprehensive review. Knowledge and Information Systems, 1-51.
Čuklina, J., Lee, C. H., Williams, E. G., Sajic, T., Collins, B. C., Rodríguez Martínez, M., . . . Keele, G. R. (2021). Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. Molecular systems biology, 17(8), e10240.
Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: a review. Artificial Intelligence Review, 55(3), 1947-1999.
Demirbaga, Ü., Aujla, G. S., Jindal, A., & Kalyon, O. (2024). Big Data Analytics in Bioinformatics Big Data Analytics (pp. 265-284): Springer.
Domrazek, K., & Jurka, P. (2024). Application of Next-Generation Sequencing (NGS) Techniques for Selected Companion Animals. Animals, 14(11), 1578.
Fondi, M., & Liò, P. (2015). Multi-omics and metabolic modeling pipelines: challenges and tools for systems microbiology. Microbiological research, 171, 52-64.
Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., & Cheng, J. (2024). Diffusion models in bioinformatics and computational biology. Nature Reviews Bioengineering, 2(2), 136-154.
Iqbal, N., & Kumar, P. (2023). From data science to bioscience: the emerging era of bioinformatics applications, tools and challenges. Procedia Computer Science, 218, 1516-1528.
Jing, Y., Bian, Y., Hu, Z., Wang, L., & Xie, X.-Q. S. (2018). Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. The AAPS journal, 20, 1-10.
Keedwell, E., & Narayanan, A. (2005). Intelligent Bioinformatics: The application of artificial intelligence techniques to bioinformatics problems: John Wiley & Sons.
Khan, M. F., Ali, A., Rehman, H. M., Noor Khan, S., Hammad, H. M., Waseem, M., . . . Jabbar, A. (2024). Exploring optimal drug targets through subtractive proteomics analysis and pan genomic insights for tailored drug design in tuberculosis. Scientific Reports, 14(1), 10904.
Loeffler, H. H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L. H., & Engkvist, O. (2024). Reinvent 4: Modern AI–driven generative molecule design. Journal of Cheminformatics, 16(1), 20.
Magro, D., Venezia, M., & Balistreri, C. R. (2024). The omics technologies and liquid biopsies: Advantages, limitations, applications. Medicine in Omics, 100039.
Mann, M., Kumar, C., Zeng, W.-F., & Strauss, M. T. (2021). Artificial intelligence for proteomics and biomarker discovery. Cell Systems, 12(8), 759-770.
Matthews, H., Hanison, J., & Nirmalan, N. (2016). "Omics"-informed drug and biomarker discovery: opportunities, challenges, and future perspectives. Proteomes, 4(3), 28.
Melo, M. C., Maasch, J. R., & de la Fuente-Nunez, C. (2021). Accelerating antibiotic discovery through artificial intelligence. Communications biology, 4(1), 1050.
Mukherjee, A., Abraham, S., Singh, A., Balaji, S., & Mukunthan, K. (2024). From Data to Cure: A Comprehensive Exploration of Multi-omics Data Analysis for Targeted Therapies. Molecular Biotechnology, 1-21.
Nguyen, M. H., Tran, N. D., & Khanh Le, N. Q. (2024). Big Data and Artificial Intelligence in Drug Discovery for Gastric Cancer: Current Applications and Future Perspectives. Current Medicinal Chemistry.
Odenkirk, M. T., Reif, D. M., & Baker, E. S. (2021). Multiomic big data analysis challenges: increasing confidence in the interpretation of artificial intelligence assessments. Analytical chemistry, 93(22), 7763-7773.
Ogunjobi, T. T., Ohaeri, P. N., Akintola, O. T., Atanda, D. O., Orji, F. P., Adebayo, J. O., . . . Shodipe, O. O. (2024). Bioinformatics Applications in Chronic Diseases: A Comprehensive Review of Genomic, Transcriptomics, Proteomic, Metabolomics, and Machine Learning Approaches. Medinformatics.
Schmidt, B., & Hildebrandt, A. (2017). Next-generation sequencing: big data meets high performance computing. Drug discovery today, 22(4), 712-717.
Schneider, G. (2018). Automating drug discovery. Nature reviews drug discovery, 17(2), 97-113.
Shah, S. M., & Khan, R. A. (2020). Secondary use of electronic health record: Opportunities and challenges. IEEE access, 8, 136947-136965.

Sharma, V., Singh, A., Chauhan, S., Sharma, P. K., Chaudhary, S., Sharma, A., . . . Fuloria, N. K. (2024). Role of Artificial Intelligence in Drug Discovery and Target Identification in Cancer. Current drug delivery, 21(6), 870-886.

Shi, L., Perkins, R. G., Fang, H., & Tong, W. (2008). Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Current opinion in biotechnology, 19(1), 10-18.

Shukla, A. K., Muhuri, P. K., & Abraham, A. (2020). A bibliometric analysis and cutting-edge overview on fuzzy techniques in Big Data. Engineering Applications of Artificial Intelligence, 92, 103625.

Singh, S., Gupta, H., Sharma, P., & Sahi, S. (2024). Advances in Artificial Intelligence (AI)-assisted approaches in drug screening. Artificial Intelligence Chemistry, 2(1), 100039.

Singh, S. V. (2024). Bioinformatics–Supporting modern life science research, applications, and challenges. Brazilian Journal of Development, 10(2), e67060-e67060.

Tantoso, E., Wong, W.-C., Tay, W. H., Lee, J., Sinha, S., Eisenhaber, B., & Eisenhaber, F. (2019). Hypocrisy around medical patient data: issues of access for biomedical research, data quality, usefulness for the purpose and omics data as game changer. Asian Bioethics Review, 11(2), 189-207.

Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. Science translational medicine, 4(125), 125ra131-125ra131.

Taylor, D. J., Eizenga, J. M., Li, Q., Das, A., Jenike, K. M., Kenny, E. E., . . . Paten, B. (2024). Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Annual Review of Genomics and Human Genetics, 25.

Tiwari, P. C., Pal, R., Chaudhary, M. J., & Nath, R. (2023). Artificial intelligence revolutionizing drug development: Exploring opportunities and challenges. Drug Development Research, 84(8), 1652-1663.

Tran, L. T., Thi, H. V., & Chu, D.-T. (2024). Bioinformatics in Preventive Medicine and Epidemiology Advances in Bioinformatics (pp. 347-358): Springer.

Tripathi, R. K. P. (2024). Next-Gen Pharma: A Roadmap Through Computational Drug Discovery Converging Pharmacy Science and Engineering in Computational Drug Discovery (pp. 28-53): IGI Global.

Viveka, S., Shreevatsa, B., Kavana, C., Dharmashekara, C., Mahesh, K., Kollur, S. P., & Shivamallu, C. (2024). Introduction to Bioinformatics Bioinformatics for Oral Cancer (pp. 1-28): CRC Press.

Zech, M., & Winkelmann, J. (2024). Next-generation sequencing and bioinformatics in rare movement disorders. Nature Reviews Neurology, 1-13.

Zhou, Y., Shen, X., He, Z., Weng, H., & Chen, W. (2024). Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functional Phenotypes. Journal of Theory and Practice of Engineering Science, 4(02), 45-51.

Zhu, H. (2020). Big data and artificial intelligence modeling for drug discovery. Annual review of pharmacology and toxicology, 60, 573-589.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, 71-91.