Rev. Colomb. Cienc. Quím. Farm., **54**(3), 621–642 (2025)

Review article

Drug design and discovery with bioinformatics tools

Amjad I. Oraibi[1]*, Ahmed Mohammed Zheoat[1], Hayder Naji Sameer[2], Mohammed Ayad Alboreadi[1] & Hayder M. Abdulhamza[3]

[1] Department of Pharmacy, Al-Manara College for Medical Sciences University, Misan, Iraq
[2] Ministry of Health and Environment, Thiqar Health Directorate, Nasiriya, Iraq
[3] Department of Pharmacy, Al-Kut University College, Alkut, Wasit, Iraq, 52001

*Corresponding author E-mail: amjadibrahim@uomanara.edu.iq, Orcid ID: https://orcid.org/0009-0009-0879-8380

Received: January 20, 2025
Corrected: June 13, 2025
Accepted: June 18, 2025

https://doi.org/10.15446/rcciquifa.v54n3.118428

SUMMARY

Context: Bioinformatics is a combination of different fields including computer science, biology, information technology and statistics which is used to analyze and interpret the biological data. It is used to design and discover novel drugs through biological data analysis and potential target identification. With increasing drug resistance among bacterial species, there is a need to develop new drugs. **Aims**: However, the drug discovery procedure is laborious, expensive and time-consuming. The new drug identification consists of various steps including target identification, target protein structure analysis, potential drug candidate identification, detecting the safety and efficacy of drug, optimizing them and finally validating the drug. Bioinformatics has a vital role in all these steps. Bioinformatics has emerged as a powerful tool in the field of drug design and discovery, enabling the rapid identification of potential drug targets, optimization of lead compounds, and prediction of drug interactions. For instance, analysis of protein sequences and genetic data enables target identification. **Methods**: Once the target protein is recognized, its structure can be investigated using bioinformatics tools. Identification of potential ligand binding sites allows for screening of compound databases to find drug candidates. A review of the relevant literature highlights that the identification of potential ligand binding sites allows for screening of compound databases to find drug candidates. Simulations of target protein and biomolecule interactions aid in predicting drug safety and efficacy. **Results**: Bioinformatics is utilized for drug optimization to improve safety and efficacy. Recently pharmacophore and molecular docking process are used for screening of thousands of candidate molecules to a few promising leads. In this paper, drug design and discovery has been done with the use of bioinformatics tool in the field of network pharmacology. This paper reviews the application of bioinformatics tools within the field of network pharmacology, focusing on methodologies for drug design and discovery. It aims to clarify existing approaches and propose future directions in the context of drug development, rather than suggesting that original drug design and discovery work has been conducted by the authors.

Keywords: bioinformatics; target protein; ligand binding site; biomolecule interaction; pharmacophore; docking; pharmacology.

RESUMEN

Diseño y descubrimiento de fármacos con herramientas bioinformáticas

Contexto: La bioinformática es una combinación de diferentes campos, como la informática, la biología, las tecnologías de la información y la estadística, que se utiliza para analizar e interpretar datos biológicos. Se utiliza para diseñar y descubrir nuevos fármacos mediante el análisis de datos biológicos y la identificación de posibles dianas. Debido al aumento de la resistencia a los fármacos en las especies bacterianas, surge la necesidad de desarrollar nuevos fármacos. **Objetivos:** Sin embargo, el proceso de descubrimiento de fármacos es laborioso, costoso y requiere mucho tiempo. La identificación de nuevos fármacos consta de varios pasos: la identificación de la diana, el análisis de la estructura de la proteína diana, la identificación de posibles candidatos a fármacos, la detección de la seguridad y la eficacia de los fármacos, su optimización y, finalmente, la validación del fármaco. La bioinformática desempeña un papel fundamental en todos estos pasos. La bioinformática se ha convertido en una herramienta poderosa en el campo del diseño y descubrimiento de fármacos, permitiendo la rápida identificación de posibles dianas farmacológicas, la optimización de compuestos clave y la predicción de interacciones farmacológicas. Por ejemplo, el análisis de secuencias de proteínas y datos genéticos permite la identificación de la diana. **Métodos:** Una vez reconocida la proteína diana, se puede investigar su estructura mediante herramientas bioinformáticas. La identificación de posibles sitios de unión de ligandos permite el cribado de bases de datos de compuestos para encontrar fármacos candidatos. Una revisión de la literatura relevante destaca que la identificación de posibles sitios de unión de ligandos permite el cribado de bases de datos de compuestos para encontrar fármacos candidatos. Las simulaciones de proteínas diana e interacciones biomoleculares ayudan a predecir la seguridad y eficacia de los fármacos. **Resultados:** La bioinformática se utiliza para la optimización de fármacos con el fin de mejorar la seguridad y eficacia. Recientemente, el farmacóforo y el proceso de acoplamiento molecular se han utilizado para el cribado de miles de moléculas candidatas hasta encontrar unas pocas líneas prometedoras. En este artículo, se ha realizado el diseño y descubrimiento de fármacos mediante herramientas bioinformáticas en el campo de la farmacología de redes. Este artículo revisa la aplicación de herramientas bioinformáticas en el campo de la farmacología de redes, centrándose en las metodologías para el diseño y descubrimiento de fármacos. Su objetivo es aclarar los enfoques existentes y proponer futuras direcciones en el contexto del desarrollo de fármacos, en lugar de sugerir que el trabajo original de diseño y descubrimiento de fármacos haya sido realizado por los autores.

Palabras clave: bioinformática; proteínas diana; sitio de unión de ligando; interacción biomolecular; farmacóforo; acoplamiento; farmacología.

RESUMO

Projeto e descoberta de fármacos com ferramentas de bioinformática

Contexto: A bioinformática é uma combinação de diferentes áreas, incluindo ciência da computação, biologia, tecnologia da informação e estatística, utilizada para analisar e interpretar dados biológicos. É utilizada para projetar e descobrir novos fármacos por meio da análise de dados biológicos e da identificação de potenciais alvos. Com o aumento da resistência a fármacos entre espécies bacterianas, surge a necessidade de desenvolver novos fármacos. **Objetivos:** No entanto, o procedimento de descoberta de fármacos é trabalhoso, caro e demorado. A identificação de novos fármacos consiste em várias etapas, incluindo a identificação do alvo, a análise da estrutura da proteína alvo, a identificação de potenciais candidatos a fármacos, a detecção da segurança e eficácia dos fármacos, a otimização dos mesmos e, finalmente, a validação do fármaco. A bioinformática desempenha um papel vital em todas essas etapas. A bioinformática emergiu como uma ferramenta poderosa no campo do projeto e descoberta de fármacos, permitindo a rápida identificação de potenciais alvos, a otimização de compostos líderes e a previsão de interações medicamentosas. Por exemplo, a análise de sequências de proteínas e dados genéticos permite a identificação do alvo. **Métodos:** Uma vez reconhecida a proteína alvo, sua estrutura pode ser investigada utilizando ferramentas de bioinformática. A identificação de potenciais sítios de ligação de ligantes permite a triagem de bancos de dados de compostos para encontrar candidatos a fármacos. Uma revisão da literatura relevante destaca que a identificação de potenciais sítios de ligação de ligantes

permite a triagem de bancos de dados de compostos para encontrar candidatos a fármacos. Simulações de interações entre proteínas-alvo e biomoléculas auxiliam na previsão da segurança e eficácia de fármacos. **Resultados:** A bioinformática é utilizada na otimização de fármacos para melhorar a segurança e a eficácia. Recentemente, o farmacóforo e o processo de docking molecular têm sido utilizados para a triagem de milhares de moléculas candidatas a algumas pistas promissoras. Neste artigo, o projeto e a descoberta de fármacos foram realizados com o uso de ferramentas de bioinformática na área de farmacologia de redes. Este artigo analisa a aplicação de ferramentas de bioinformática na área de farmacologia de redes, com foco em metodologias para projeto e descoberta de fármacos. O objetivo é esclarecer as abordagens existentes e propor direções futuras no contexto do desenvolvimento de fármacos, em vez de sugerir que o trabalho original de projeto e descoberta de fármacos foi conduzido pelos autores.

Palavras-chave: bioinformática; proteína-alvo; sítio de ligação de ligantes; interação com biomoléculas; farmacóforo; docking; farmacologia.

1. INTRODUCTION

Bioinformatics is defined as the process of computational technique to evaluate the data related to the biomolecule. It is the technique of storing, fetching, organizing and evaluating as well as interpreting and usage of information from sequences and biomolecules [1]. The three main reason for incorporating Bioinformatics tools are as follows [2], 1) It organizes data in the form that the researchers can fetch information and submit the new data entries they produced. The protein data bank is the best example for storing 3D macromolecular structures. 2) It develops resources as well as tools which helps in evaluating the data. The selected protein is sequenced and analyzed with the existing sequences. Example such as FASTA and PSI BLAST programs [3] will analyze the resemblance of biological data. It requires both the knowledge of computational theory as well as biology. 3) Finally, it makes use of tools to evaluate the biological data and interpret the outcomes in a biologically significant manner. The biological systems investigate an individual in broad view, and compare with the less that are similar. In the tool of bioinformatics, the data analysis is performed globally with the existing data and it signifies the features that are identical.

The two main principles underlying in the analysis of bioinformatics tools are as follows [4], 1) Comparing and evaluating the data in specification with the biological similarities. 2) Analyzing single mode of data to interpret and evaluation for another set of data.

Margaret Dayhoff referred to the “mother as well as father of bioinformatics” pioneered the bioinformatics application in the biochemistry field. She created a new software “COSMOPROTEIN” working associated with Robert. S. Ledley. They analyzed the biological problems through computational methodology. In the application, the input as well as output data are the sequences of amino acid and is represented by three letter abbreviation. For example, Lysine-Lys, and Serine-Ser. Dayhoff simplified the methodology and transformed it to one-letter code which is used today [5].

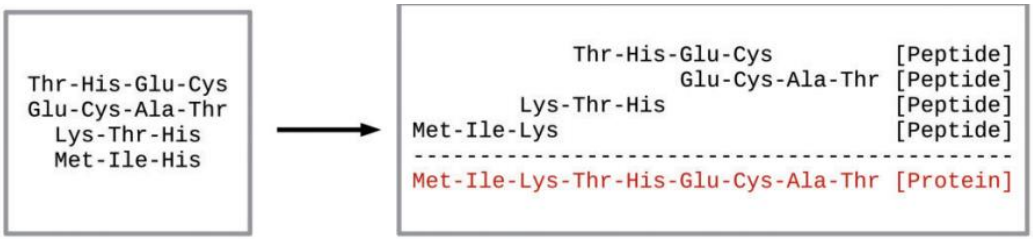

Figure 1. Example of COSMOPROTEIN input and output of protein sequences.

In Figure 1, resembles the input and output data of the protein sequence of COSMOPROTEIN software. The input is the Edman sequence of peptide and it is abbreviated by the three letter words. The output is nothing but the consensus sequence of protein and it is also represented by three letter words. The software referred to as "IBM 7090 – a complete computer application" and it is constructed to demonstrate the primary structure of protein. It is evaluated by the utilization of Edman peptide sequencing. It is coded in FORTRAN which is called as de novo assembler of sequences.

Life scientists are not involving directly in accessing the bioinformatics tool and require adequate skill in understanding the significance of bioinformatics to prevent misuse and interpretation error in the outcome. This leads to significant increase of availability of user friendly application such as Galaxy and the communities such as Bio Star and SEQ answers. It also require skilled bio-informatician to operate this application. Hence, the university adopt bioinformatics in their curriculum of biological science departments [6].

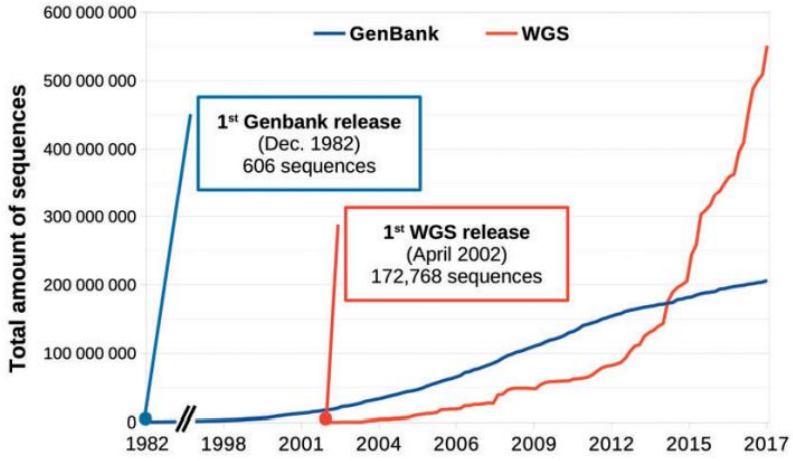

Figure 2. Databases of NCBI Gen Bank, Whole Genome Shotgun (WGS) genome sequences correlate with the time.

In the year 2008, Moore's law became less accurate in determining the DNA sequencing. The magnitude of the outcome falls rapidly in emerging technologies. Therefore, it resulted in the drastic increase of sequences in the databases of GenBank as well as WGS-Whole-Genome Shotgun. It has produced data beyond the level of Exabyte - 10^8. The computational resources are needed for the proper storage and organization for their ease of access and usability [7]. Bioinformatics is the intertwined field that comprises of computer in the biology field. In the Brown's theory of hypothetical future, biology as well as bioinformatics are inter-related with one another and commonly referred to as simply biology. There is no distinction of two fields. The graph is the easiest method of explaining a complex system that comprises of collection of objects and their associations. As the machine learning is succeeded in the bioinformatics system, its outcome is represented by the Euclidean domain. Therefore for complex data analysis, graphical methodology is used to compress the graph in low-dimension which occupies less space preserving node properties and topology.

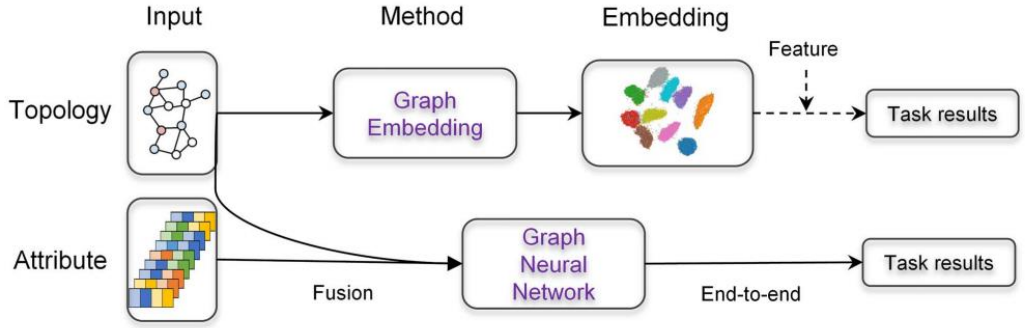

Figure 3. Graphical embedded and neural methods comparative analysis.

There are two major classification in the graphical methodologies of bioinformatics process. They are listed as follows, 1) Graphical Embedding Methods (GEM), 2) Graphical Neural Networks (GNN)

The graphical embedding aims to observe low-dimensional significance of nodes that preserve graphical topology. GNN not only observes the embedding but also completes the work on graphs at the end process. These techniques are widely used for analysis in bioinformatics tool. Figure 3 represents the comparison of graphical embedding methods and neural networks which had its significance in the field of bioinformatics. The GEM produce node representation which can be linked together with the machine learning methodologies to operate downstream task. GNN combines the topology as well as attributes to work on end to end tasks [8].

The development of physiology, molecular biology and other fields leads to the awareness of biomolecule process and their role in performing biochemical activities [9]. The modelling and evaluating of the biological graphs to resolve the issues of complication can be classified into three phases. They are, 1) Bipartite graph [10] and it is significant in the drug indication process such as drug-target association graph [11]. 2) Multi relational graph [12] and it is significant for analyzing drug discovery and its remedies. It focusses on the drug-disease, drug-gene, and drug-drug associations. 3) Biomedical graph also referred to as knowledge graph [13] and it is used in the database of PharmGKB, Gene Ontology and DrugBank for treatment of diseases [14].

1.1. Drug Discovery On The Basis Of Computational Tools

In traditional method, drug design and discovery is tedious as well as time consuming technique and also an expensive process. It will take an average of ten to fifteen years for the drug to reach market. It leads to the generation of CADD (Computer Aided Drug Design) systems which are ligand-based, structure-based as well as system based. The process is carried out in five stages [15].

Stage 1. Identification of target. It reveals the target protein which identify the diseases at the first stage of process.

Stage 2. At the second stage, the screening of identical compound that inhibits the target protein. These compounds are segregated.

Stage 3. At the third stage, the transformation of identical compounds to lead compound will take place. The optimization of the lead compound used for the treatment of disease will be identified.

Stage 4. Next, the evaluation of pharmacokinetic properties of the optimized drug compound. It will provide the complete information regarding the drug and target association.

Stage 5. Finally, the adverse effects of the drug can be determined by the process of ADMET – Adsorption, Distribution, Metabolism, Excretion and Toxicity. This process is carried out in the initial stage of drug development to increase the success rate as well as minimization of time consuming process.

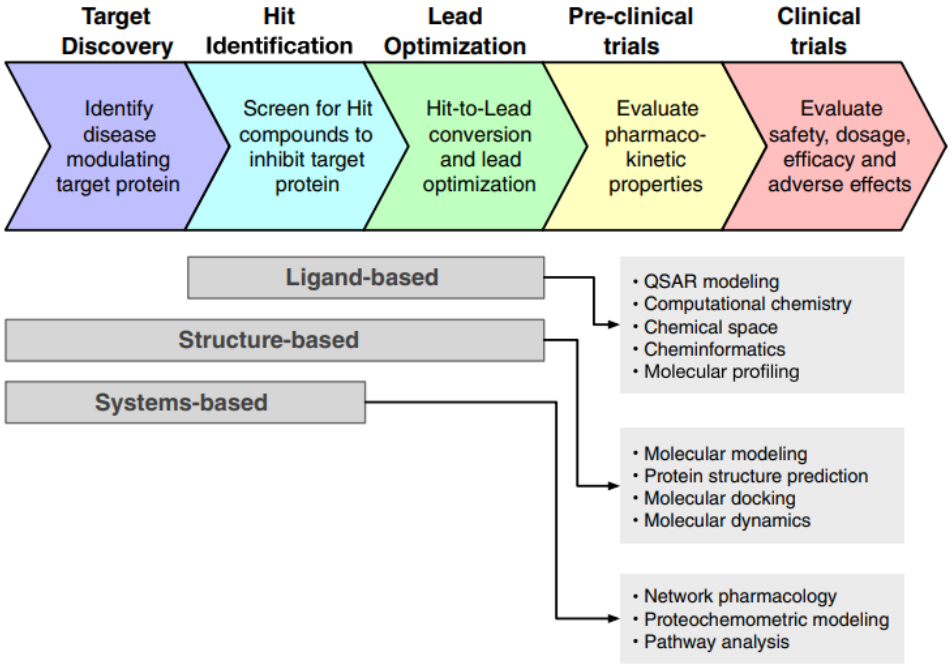

Figure 4. Drug discover methodology through computational tools.

Figure 4 demonstrates the complete drug discovery process, which is categorized into five critical stages namely the Target Discovery, Hit Identification, Lead Optimization, Pre-clinical Trials, and Clinical Trials. In the initial stage, Target Discovery used to recognize the disease moderating target proteins that is utilized in numerous diseases. The next phase, Lead Optimization, is utilized in converting identified hits into lead compounds by rigorous refinement to enrich their pharmacological properties. The Pre-clinical Trials, where the pharmacokinetic properties of optimized leads are estimated, evaluating their ADME (absorption, distribution, metabolism and excretion) features. Finally, Clinical Trials are conducted to evaluate the safety, dosage, efficacy and adverse effects of the drug in humans. Additionally, there are three approaches namely Ligand-based methods, which used the techniques such as QSAR (Quantitative Structure-Activity Relationship) modeling and molecular profiling, Structure-based methods, which mainly involves the molecular modeling and Systems-based approaches that incorporate network pharmacology and which is used to understand the biological associations of drug candidates. This structured framework guarantees a systematic transition from identifying targets to the development of secure and efficient pharmaceuticals**.**

The implementation of CADD in the system at the stage of target identification through experiments in wet-lab and followed by HTS-high throughput screening [16]. Due to huge requirement of resources with the extensive cost leads to the generation of molecular docking process [17]. It makes use of known target screened against the virtual library. This process is a highly efficient one but it requires crystal structure of the compounds. If the crystal structure is not available, de novo prediction technique will be used for screening process. It will determine the compound with the high degree of binding affinity (Figure 4).

The huge number of bioactivity sets will be analyzed by QSAR- Quantitative Structure Activity Relationship [18]. It predicts the activity of compounds through their structures. Those predictions are used in the process of target identification, ADMET-Adsorption and Metabolism prediction and drug repurposing process. In the available computational tools, it is important to focus on the quality of data rather than the quantity. Many analysis has been performed to illustrate the quality of data but the outcome is the prediction of the compounds which provide false positive result during analysis [19]. Therefore, proper workflows as well as pipelines are needed for ensuring the computational research to be more accurate.

Due to high level of data sets, the utilization of shared as well as public in the form of cloud or HPC-High performance Computing infrastructure of computational tool is required [20]. It added next level of complexity for the reproducibility of computational analysis. To obtain the portability between the systems, tools such as virtual machines and also virtual containers are utilized. High level of automation can be obtained through connection of containers with the virtual machines. It facilitates highly reliable as well as replicable services such as deploying models acts as a services compared to networks.

1.2. Problem Statement

In drug design and discovery, merging bioinformatics tools offers a hopeful strategy to speed up the finding of new drug targets, improving lead compounds, and forecasting drug interactions. Nevertheless, even with the progress made in bioinformatics, there are still obstacles and deficiencies that must be resolved in order to improve the efficiency and effectiveness of drug discovery processes. The complexity of biological systems makes it difficult to accurately model drug interactions and predict drug effectiveness. Bioinformatics tools must consider the intricacy of biological processes in order to enhance the precision of drug design and discovery. The wealth of data from genomics, proteomics, and other omics areas requires reliable tools for data integration and analysis to uncover valuable insights for drug discovery.

Bioinformatics tools need to efficiently manage and analyze various data sets in order to support well-informed decision-making. The designing of new drugs involves the development of innovative medication that communicates with the target. It comprises of proteomics, genomics and bioinformatics process resulted in the discovery of novel drugs through mechanism of action. The drug discovery process is crucial step and it is a time consuming process. It depends on several factor such as detection of drug targets, 3-D structure of protein and computational power in bioinformatics tool to detect protein and drug interactions. In modern world, the innovation in technologies has provide the pathway for drug designing process to be simpler in the pharmaceutical industry.

Forecasting drug-target interactions is essential in order to discover possible drug candidates with accuracy. Bioinformatics tools must improve their predictive abilities to simplify the lead optimization and drug development process. Integration of multiple omics data poses a considerable challenge when attempting to gain a thorough understanding of disease mechanisms and drug responses. Bioinformatics tools must create strong methods for combining and examining multi-omics data to support precision medicine efforts in drug discovery. Thus, the current study targets to improve drug development processes by using bioinformatics tools to tackle existing challenges and gaps in the field of drug design and discovery.

The significance of this research lies in its potential to revolutionize the drug discovery process by providing a more efficient and effective approach to identifying new drug targets,

optimizing lead compounds, and predicting drug interactions. By addressing the existing challenges and gaps in the field, this study hopes to contribute to the development of innovative and effective drugs that can improve patient outcomes and enhance the overall quality of life.

1.3. Research Objectives

The main aim of the study is to investigate the process of drug designing and discovery with the support of bioinformatics tool

The study has been instigated with the following objectives such as, 1) Analyze the challenges faced in the designing process of novel drug compounds. 2) Discussion on the existing technologies to overcome the challenges in drug discovery process. 3) Proposing the framework to avail the production of new drug compounds in the bioinformatics tool

1.4. Paper Organization

The paper is organized in the following manner in which Section 1 provides the brief introduction about the bioinformatics tool and the process involved in the drug discovery. The existing research works associated to the present study is presented in Section 2. Various process involved in the drug designing and its discovery is elucidated in Section 3. Finally, in Section 4, the paper is concluded with the suggestion to overcome the challenges faced by the pharmaceutical industry in the drug designing process.

2. LITERATURE REVIEW

The below section enumerates the review of different existing studies, stating the various bioinformatics tools in the analysis of drug design and development in the field of pharmacology.

In the early nineties, efforts are put on to construct the software which will evaluate the aspects in the process of drug discovery such as protein-ligand interactions, QSAR – Quantitative Structure Activity Relationship models, pharmacophore and molecular dynamics. The existing study analyses the various computational tools involved in the process of drug discovery [21]. CADD method is used for target identification, which is developed from single thread to multiple threads to maximize the computational power. Due to the complexity of analyzing large volume of data, the HPC – High Performance Computing infrastructure is used for the designing purposes. It not only analyses large number of data but also requires only less computing time. HPC helped at the time of pandemic situation in finding the solution to the infectious diseases. It makes use of diverse field such as grid and cloud computing. Cloud computing is considered to be more efficient and cost effective for providing solutions to access quickly and perform research with lower economic resources. Virtual Screening (VS) also gain its importance in the screening of large data sets and extracting smaller datasets. It is a more efficient technique. Machine learning techniques in the VS, leads to the capability of designing new molecules. Deep learning (DL) is also achieving impressive outcomes. Due to inexpensive and short time duration, ML and DL proved to be more efficient when working in cloud or clusters. Web server supports the researchers to perform experiments using database and software more economically.

The generation of new drug requires several number of compounds to be produced and it is time consuming process. It also needed biological evaluations. Therefore, the pharmaceutical industry implemented CADD for drug designing purposes. QSAR is a popular method used in CADD, which can be used for virtual screening of lead drugs and its optimization. It analyses large quantity of compounds, minimizing the time duration and labor cost for the

synthesis of chemical. It improves the efficiency of drug discovery. To cope up the requirement of researchers to handle the technology, several QSAR associated servers such as web-4D-QSAR as well as DPubChem have been established. QSAR modeling cannot be performed appropriately in the existing servers. Hence, the existing study generated a new model of Cloud 3D QSAR [22]. It combines the tasks of generation of molecular structures, configuration and MIF-Molecular Interaction Field analysis and outcome will be evaluated to provide a solution in one step. Since non-expert user find difficult in using QSAR for the drug designing process. This gives the motivation to construct the model as it automates all the existing process and generates prediction function. It facilitates the generation of appropriate QSAR model in the drug designing process.

The development and discovery of the drug requires the combination of various scientific as well as technological fields such as biology, chemistry and information technology. The existing study demonstrate the utilization of various technologies in the drug designing process. CADD techniques is used with numerous work process [23]. VS is one of the most frequent methods of CADD and is used in the drug designing process. It processes the initial stage of drug development. Several requirements of VS can meet up with the improvement of cloud based platforms. Those platforms support the researchers with the applications to make and monitor the process of drug discovery and their development pipeline. It enhances the performance of VS and helps to detect the commercially available drugs and compounds for testing purposes. The process will be classified into various stages such as constructing the chemical data sets, analyzing the libraries and the ligand-based or structure-based can be conducted in the cloud technologies. These type of platforms are best at its utilization in the pharmaceutical industry and R&D process in the generation of re-purposing of drugs with several computational options.

The Big data is defined as a large volume of data comprising both unstructured as well as structured huge data to be resolved by the conventional application software. The outburst of data is due to three main reasons. There are several progressive information collected from wide applications such as mobile sensors, social media and other associated devices. Availability of storage space at the cheaper prices and acquisition of knowledge strategies have been developed with the effective retrieval process. Due to drastic increasing of data in the bioinformatics, there is a demand in the data storage for the vast information. The conventional study suggests that the cloud computing will resolve the problem of big data issue in bioinformatics and other challenges in the bioinformatics sector [24]. It is a cost effective technique to develop and collect the large quantity of data with the support from parallel processing tools. The framework so called “Hadoop parallel” programming is frequently used. This pave the way to incorporation of the economical Bioinformatics clouds for the Big data [25]. The existing paper suggest the perspective of cloud computing services in supplementation with the Bioinformatics tools. Fog server will compare the trust values of various cloud providers. For future improvements, efforts are being made to build an effective cloud data storage system to ensure security in the bioinformatics system so that various pharmaceutical industries are benefited by this methodology.

Structure oriented design as well as computational chemistry has accelerated the process of drug discovery. It does not act as driving force of the discovery of small molecule. The existing study suggests the improvement of computational tool in the drug discovery process from the last decades [26]. The development of technology in the accurate prediction of drug from the initial process of potency to final stage of dissolving characteristics. Next, in terms of enhancement in the artificial intelligence as well as deep learning methods and finally sudden

increase of computational tool with the emergence of GPU and also cloud computing. It leads to the capability to identify and predict accurately the chemical compound similar to drugs in silico. Due to the generation of cryogenic electron microscopy as well as prediction of protein structure through the computational tool in the field of structural biology, leads to the access of several three dimensional structures of the de novo complexes of drug and receptors. The continual advancement in the sector of computational chemistry structural biology and also machine learning with the accelerated computational power leads to the safe, novel and effective medicine generation for the patients.

The traditional research has attempted to explore the recent applications of computational drug discovery methods specifically, virtual screening (VS) [27]. It employs deep learning and machine learning approaches to forecast drug-target communication for detecting new drugs. The research has been described the main elements of VS techniques incorporating compound and protein descriptors/features, libraries and toolkits for VS. Further, bioactivity databases and gold standard datasets are also used for benchmarking and training. The results have been detected that deep learning approaches especially DNN have expressed higher performance in forecasting drug-target interactions related to other techniques. DNNs are perfect fit for investigating the association among compounds and targets. DL is applied to other drug-discovery based tasks like aqueous solubility prediction, therapeutic effect prediction and compound toxicity prediction. Finally, it has been detected that DL has significant potential to develop the field of computational drug discovery.

The conventional study has aimed to offer a summary of how bioinformatics tools can speed up the process of identifying drug targets, analyzing and adjusting drug candidates, and categorizing potential side effects and drug resistance during drug discovery [28]. It has investigated the present scenario of utilizing bioinformatics in the field of drug design and discovery. They investigated how different high-throughput data types such as genomics, epigenetics, transcriptomics, proteomics, etc. can be utilized for drug repurposing and mechanism-based drug development. The research also examined how improvements in modeling protein and RNA structures, databases of small molecules, and computational methods such as protein-ligand docking and virtual screening are facilitating the process of drug discovery through computer simulations.

Further, the outcomes have been identified that bioinformatics tools play a key role in speeding up drug discovery via facilitating target identification, lead optimization, and predicting side effects. Data from multiple omic disciplines offers valuable data for identifying novel drug targets and reusing current medications. Additionally, the outcomes have been detected that, progress in predicting protein structures and utilizing small molecule databases, along with methods like docking and virtual screening, are improving the efficacy of virtual screening for potential drug candidates. Nevertheless, existing computational techniques still have constraints and need cycles of designing, producing, and examining to identify the best drug compounds.

Deep learning has its advantages of image classification and speech recognition over machine learning. As the DL theory focused on the Euclidean data and had its better outcomes. The existing study concentrates on the non-Euclidean data such as graphical representation of data [29]. More researchers are performed in processing the graphical data to determine complex relationship among objects. GEM can be performed for mapping the graphical data to simple depictions. Although, this method loses topological characteristics in the initial stage thereby affecting the final outcomes. GNN a sub division of DL in non-Euclidean data perform well in various operation which process data in graphical form. GNN plays a vital role in the

field of bioinformatics. GNN are classified into three main models such as classification of nodes, link identification and generation of graphs. It has its wide application in the process of identification of diseases, drug discovery and finally biomedical imaging. GNN has attained excellent outcomes in many biological process yet still faces issues regarding poor quality of data processing, procedure and interpretability. Despite of many challenges, GNN proves to be highly efficient in solving numerous problems in the research of bioinformatics.

Deep neural network is the general and successful methodology of machine learning process. Graph comprises of edges and nodes. Entities is represented by the nodes and interaction as well as relationship among entities is represented by the edges. The essential part of the graph is edges. GCN - graph convolutional networks and GAT- graph attention networks are some of the neural models which make use of utilized features of edges. It also includes edges of multi dimension. The proposed framework in the conventional study comprises of both the models of neural network such as GCN as well as GAT [30]. This proposed framework has the following innovative criteria such as 1. It use particularly stochastic graph edge as an alternative to the generally used normalization methodology in GNN. 2. It builds new formulas for the process in each layer to handle the edge of multi-dimensional characteristics. 3. In the layers of network, there is adaptability of edge features. 4. It encode the directions in multi-dimensional features of edges. The outcome will be the exploitation of the sources of information regarding the graph edges. The new model can be applied to the classification of graphical nodes and whole graphs and finally regression on datasets. The proposed model provide enhanced performance when comparing with the traditional model to exploit the edge features in GNN. It eradicates the limitation of GAT as it handle only one binary edge and also GCN which can capable to handle one dimensional edge. The model is effective in learning directed graphs.

3. DESIGNING AND DEVELOPMENT OF DRUGS

3.1. Graphical representation of learning in the field of pharmaceuticals

In the modern pharmaceutical sector, finding and creating new medications is a difficult and intricate procedure. The procedure requires substantial financial commitments, repetitive stages, and considerable chances of not succeeding. In order to tackle these obstacles, scientists have resorted to using visual representation as an effective method to speed up the process of finding and repurposing drugs. Using visual aids like graphs, charts, and diagrams is a way to represent complicated data clearly and succinctly. Graphical representation can be utilized in pharmaceutical research to merge chemical information of compounds with their interactions with targets. This enables researchers to visualize the connections between various compounds and their possible therapeutic benefits. Using visuals in pharmaceutical research has multiple advantages. First and foremost, it allows scientists to rapidly pinpoint patterns and trends in vast datasets, which can speed up the process of finding new drugs. Additionally, it enables to observe how compounds interact with targets, aiding in the detection of possible side effects and improving the development of new drugs.

Graphical representation can be utilized not only to speed up drug discovery and repositioning, but also to evaluate side effects of current drugs. By visualizing how different compounds are related and their possible side effects, it can pinpoint potential dangers and enhance the development of new drugs to reduce these risks. In general, the utilization of visual aids in pharmaceutical research could greatly transform the process of finding and creating

new medications. The use of visual aids in representing intricate data can speed up the identification of new medications, improve the development of new drugs, and reduce the dangers linked to adverse reactions [31].

3.1.1. Prediction of drug-target interaction

The connection between medications and targets is a vital element of the process of discovering and developing drugs. Scientists have come up with a new method to forecast these interactions by using tripartite graphs that illustrate the connections between drugs, targets, and diseases [32]. This method entails examining the correlations among drug-target connections and drug-protein pairs by utilizing Graph Convolutional Networks (GCN) to outline the characteristics of drug and protein pairs [33].

The GCN method is employed for assessing the relationship between drugs and targets, utilizing end-to-end learning approach to forecast the binding affinity of drugs and targets [34]. This method uses visual nodes to study drugs and targets, assessing the resemblances between target–target, drug–drug, and drug–target pairs [35]. Graph DTA is a new approach that represents drugs graphically and utilizes GNN to assess the binding affinity between drugs and targets. This method utilizes graph neural networks to understand the intricate connections between drugs and targets, allowing for the anticipation of drug-target interactions [36].

Tripartite graphs and graph neural networks can aid in making precise forecasts about the relationships between drugs and targets. It can enhance the identification of new drugs and the improvement of current ones. This method could completely transform drug discovery by allowing the creation of better and more precise treatments for different illnesses.

3.1.2. Prediction of drug–disease association

The identification of novel applications for current drugs, known as drug repositioning, has become a more significant approach in the field of drug development. Successful drug repositioning relies on precise identification of drug-disease associations [37]. Graphical techniques have become a strong tool for recognizing these connections. Methods like Deepwalk, SDNE, LINE, and HOPE are employed for the purpose of depicting drugs and diseases in a graph by using graph embedding techniques. The graph that is produced assists in grasping the complex relationships between diseases and drugs, aiding in understanding their embedding. Heterogeneous graphs are built to depict similarities between pairs of drugs and diseases [38].

Graph Convolutional Networks (GCNs) are used to understand the characteristics of diseases and drugs, enabling the discovery of potential connections between them. The GCN layer considers the graph's structural information, enabling better detection of drug-disease connections [39]. Through the use of graph embedding and GCNs, scientists are able to quickly discover fresh uses for current medications, speeding up the process of drug repositioning [40]. This method could greatly lessen the time and expenses linked to conventional drug discovery techniques, ultimately resulting in the creation of better treatments for different diseases [41].

3.1.3. Prediction of drug–drug interaction

Identifying drug-drug associations is crucial in preventing harm and minimizing costs by avoiding potential adverse effects caused by the interaction of different medications [42]. By utilizing a knowledge graph approach, data is extracted from databases such as Drugbank,

KEGG, and PharmGKB to improve the comprehension of drug interactions and possible negative effects [43]. Precise forecasting of drug interactions helps to prevent negative reactions and also supports the creation of novel medications for treating different illnesses [44].

Visualization of learning acts as a connection between biomedical information and advanced machine learning methods, enabling investigation from molecular level interactions to healthcare systems [45].

This method is crucial in the field of bioinformatics as it provides understanding of intricate connections in biological systems. The main types of Graph Embedding Models (GEM) - homogeneous and heterogeneous - facilitate network embedding to enhance understanding of drug interactions and disease mechanisms [46]. The use of visual learning tools extends to molecules, pharmaceuticals, genomics, and healthcare systems, providing a detailed understanding of intricate biological processes [47]. The existence of open resource platforms and libraries for graph representation learning boosts its usefulness, encouraging progress in this field and supporting creativity in biomedical research. This comprehensive method not only boosts our comprehension of drug interactions but also drives biomedical research towards better treatments and healthcare results [48].

3.2. Concept of digital twin

A digital twin is referred to a virtual demonstration of an object or else the system that spans its lifecycle. It is upgraded from real time data. It uses simulation, machine learning, deep learning and reasoning to support decision making process [49]. It gains more interest for its capability and impact in the fields of manufacturing, medicine and health cares. The concept of DT has been established in the year 2002 for the presentation of PLM-Product Lifecycle Management by Michael Grieves. The model has been referred to as “Grieves’ DT model” [50]. It comprises of three elements. They are, 1) Physical object in the real place, 2) Virtual object in the virtual place, and 3) Link connecting the (a) data flow in the real space and virtual space, (b) information flow from virtual and real space. It is considered to be the enabler for the data transformation and allows the synchronization and convergence of both physical and virtual systems [51]. Digital twins (DT) are intelligent systems but are not autonomous totally. Both the DT and AI related application requires several human intervention regarding the validation of new characteristics and alteration of physical assets. It also exploited to provide detection and treatment of diseases [52]. In the medicinal field, the combination of both DT and AI or DL will be more efficient and DL boost up the analyses, identification and suggestion of DT. For this reason, the combination of DT and DL is incorporated in the various medicinal researches and added its advantageous properties in the process of drug discovery [53]. DT is varied from the CADD model in obtaining the simulations [54].

The digital and physical twins are equipped with the devices to provide a seamless connection as well as progressive data exchange through physical communication or indirectly by means of cloud computing. DT progressively receive twin data that describe twin status of physical space and varied with respect to time and also its lifecycle. The environmental data refers to the status of the environment. It continuously go back to its physical twin, to the expert of the domain, and to the other DT for the optimization [49].

Three main types of communication process involved in the system are as follows [55], 1) Interconnecting physical and virtual twin, 2) Interrelating DT and several DT in the environment, and 3) Among the DT and expert domain which will interact and work with the DT through the interfaces.

It has the capability to handle high-dimensional data and includes efficient decoding and analysis process for the high dimensional data. It comprises of data fusion process for combining the several data sources and generate accurate, consistent and useful information compared to individual data provided. DT technology comprises of progressive improvement in AI which includes learning algorithms of either supervised or unsupervised [56]. It has high predicting capability which obtain sensed data from the physical twin and environment.

The parameters of digital twin technology are described as follows, Digital twins combine information from different sources such as electronic health records, wearable devices, and imaging technologies to develop a complete virtual model of a drug, process, or system. Digital twin has the capacity to monitor the physical component in real-time, allowing researchers to observe how a drug behaves in the body and forecast its effectiveness and safety. Simulation and Digital twins have the ability to imitate how a drug acts in the body, predict its effectiveness and safety, and enhance drug formulations and dosages before clinical trials. Digital twins enable the development of individualized patient models, leading to customized treatment and better patient results. Quality of data is crucial for accurate performance of digital twins. Low-quality data may result in incorrect forecasts and modeling.

Digital twins have the potential to cut down the expenses and time linked with conventional drug development methods through the use of in silico testing and simulation. Flexibility in digital twins enables modeling and simulation of diverse scenarios in biological systems and processes. Digital twins require precise data and simulations to offer trustworthy forecasts and understandings, making sure the virtual replica faithfully mirrors the real-world equivalent.

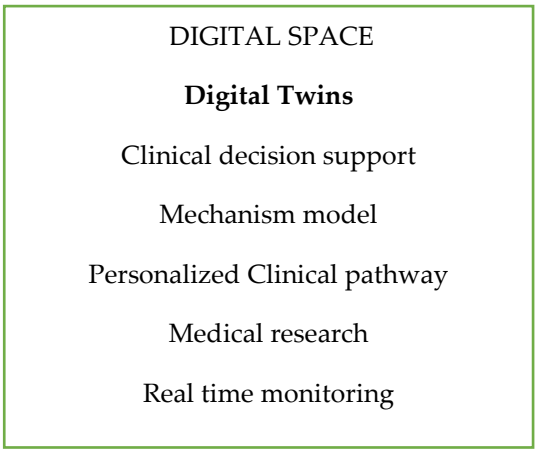

Figure 5. Digital twin technology in the pharmaceutical field.

In the pharmaceutical field, the DT has its wide applications. It had its exploration with the diverse technologies of big data, DL, IoT (Internet of Things). All the nation has spending billions of dollars in the Iot in the field of research in pharmaceutical industry [57]. DT is the only solution for generating the precision drugs and requires integrating as well as processing huge volume of data. IoT acts as a technical support for the data collection process through 2D codes, cards and sensors [58]. It collect the real-time data and feedback the processed data and controls the operation through IT (Information Technology). By integrating the DT of equipment of the industry and auxiliary equipment will provide a new platform for the health care management and services. By incorporating DT as well as big data processing, simulation can be fetched with the help of high resolution systems of patients for the accurate determination of target and their appropriate drugs. It also illustrate the proper treatment methods for the

patients at the aim of attaining precision medical procedures. Thus, the implementation of DT in the pharmaceutical industry, will pave the way of accurate designing of drugs with the limited usage of medical resources [59]. As a field make new evolutions, a more understanding of ability and limitations of DT has been vital for the developed incorporation in the pharmacological workflows.

3.3. Deep learning in the bioinformatics

DL has gain its exponential growth in the application of bioinformatics and became a promising power in determining the complex relationship of huge scale biomedical and biological data [60]. Due to arise of several issues in the Machine Learning (ML) pave the way to emergence of DL in the bioinformatics sector [61]. It is the time to develop the architecture of DL in the scope of drug discovery in the bioinformatics tool [62].

ML is the main contributor of AI. The modernized ML technology lead to the emergence of DL. It is founded of ANN – Artificial Neural Network. It has the capability of operating on any nonlinear function with any level of accuracy. Though, it has solve the issues in the computational task, it is termed as "black boxes" due to lack of interpretability. The performance level also reduced when compared to other ML methods [63].

The improvements in the technology of ANN via DL leads to minimize the issues in the computational task [64]. The complex information derived from the data of modern life has been well handled by ANN compared to the other methods [65]. The computational power has also increased with the reasonable cost and these devices are appropriate for ideal models [66]. Optimized algorithm has proposed to the introduction of deep ANN for the ideal technique of analyzing big data [67].

The challenges faced by the bioinformatics in the drug discovery process are addressed as follows, 1) The model interpretability is essential in the designing process and it support the researcher to aware of the model that solve the biological problems. For Example, DNA-protein binding. 2) The accuracy of the computational model should be in the range of 98-99% and it is much difficult to attain the high level of accuracy

The two major classification of ANN are convolution neural network (CNN) which process imaging data and recurrent neural network (RNN) which process language data. Image is highly approached by DL [68]. CNN are commonly used in the bioinformatics when compared to RNN [69]. As CNN can easily detect local features and solving the basic issues such as identification and sequencing of motifs.

Attention mechanism was first proposed to minimize the issues faced by RNN in the application of bioinformatics [70]. It has capability to handle issues of bio sequence analysis [71], such as RNA sequence evaluation and prediction, protein structure from the sequence of amino acid and detection of EPI- Enhancer-Promoter Interactions [72]. EPI reveals the gene expression thereby identifying the disease occurrences. Though experimental methods of determining EPI is tedious process, time consuming and expensive. EPIVAN has been constructed to identify long-range EPI through DL methods as well as attention mechanism [73].

Reinforcement learning (RL) is the method of successive action to happen while providing the present state of the partial solution to obtain the cumulative reward. The state will be change as the action changes. It will become the guiding action of future paradigms. This type of RL is executed in the DL process referred as Deep reinforcement learning [74]. It has the advantageous of less computational complications and does not require external supervision when compares with supervised approach and genetic algorithm [75].

Numerous data are available in the field of bioinformatics. There exist a case of data scarcity in the sector of biology and also biomedicine. There might be possibility of single data availability in the categories. Standard DL algorithm cannot be used in such cases. Hence, proposed few shot learning model for DL to handle such cases effectively [76]. As the name suggest it make use of few number of data. If there is training of one data, then the process is termed to be on-shot learning. In case of zero shot learning, there will be no sample [77]. This type of process used in the prediction of protein function and process of drug discovery. Due to availability of small biological data it is difficult to predict the novel drug compounds. Therefore the use of few shot learning process in the limited data and accurately determine the precise drug compounds.

The emergence of long short term memory model (LSTM) has attained its excellence in various task and specifically sequential issues. RNN has the limitation of working in the long term dependencies [78]. The LSTM is nothing but the extension of RNN. It can minimize the problem of internal memory and gate networks. It is used in the field of language processing in terms of classification of sentence and language modeling. It helps to construct more complicated neural network for the problems of answering task and translational task [79].

The integration of LSTM and graphical CNN will enhance the learning process of reasonable distance metrics when compared with the small molecules. The continuous refinement of LSTM results in the generation of new related assays which is not similar to those assays in the training collection sets [80]. GNN are effectively used in the conversion of small molecules into vectorial representation [81]. The outcomes of the process shows stronger performance when compared with other methodology such as CNN and random forest process. It has the potential to transfer information among related but the unique learning task [82].

From the analysis of the above discussion, it gave the suggestion of the new framework in the field of bioinformatics. The proposed framework involves the integration of DT with the incorporation of LSTM related DL methods in the drug designing process. The model make use of attention mechanism with the LSTM model. It has the advantageous properties of determining the long distance correlation with the provided input sequences. DT helps in isolating the similarities of the compounds in the tool and with the support of DL will help in constructing the novel drug. The proposed model of DT-Att-LSTM-DL has been integrated with LSTM for more effective than the conventional approach and has better prediction accuracy. Moreover, the empirical evidence examinations has been included to compare the overall performance of the DT-Att-LSTM-DL model with the conventional methodologies. The examination has been effectively provide the required information mainly to validate and enhance the models efficiency. The LSTM will model the text accurately hence it pave the way to determine the novel drugs in the bioinformatics tool.

4. DISCUSSION

The above review deliberates the incorporation of bioinformatics which considerably transforms the drug design procedure by permitting more effective and operative identification of drug targets, optimization of lead compounds and prediction of drug interactions. Bioinformatics tools support to encounter the issue in conventional drug discovery by using the computational methods for examining biological data, protein structures and ligand binding sites utilizing the techniques namely molecular docking, QSAR modeling and virtual screening. The above review highlights the role of advanced computational frameworks, comprising of DL and GNN, to manage and understand complex biological information and predict drug

target attractions, drug disease relations and drug-drug connections. The combination of DT technology further complements bioinformatics, providing virtual models that simulate drug behavior in real-time, aiding tailored medicine and exactness in the drug development. Despite these developments, the discussion recognizes prevailing issues namely data quality, biological system difficulty and model interpretability, which bioinformatics tools must address to optimize drug discovery outcomes. The proposed integration of DT with attention-based LSTM DL techniques replicates a promising approach to develop accuracy and effectiveness in novel drug design, possibly streamlining pharmaceutical research with the reduction in the time and cost.

5. CONCLUSION

The drug designing process is time-consuming and affluent in the pharmacological industry. Scientists are in examination of novel drug compounds for disease treatment. The estimation of diseases, operational drug treatment and identification of genetic codes gives an effective approach for the generation of novel drugs. The above review focuses on bioinformatics tools and expands on the current trends in the DL process. The digital twin also advances the significance in the drug discovery process as it diminishes the issues faced by the pharmacological industry. The proposed model establishes major efficiency by streamlining drug discovery, enhancing predictive accuracy, using digital twins, enabling tailored medicine, elevating resources, incorporating multi-omics information, permitting real-time information handling, aiding collaboration among researchers, guaranteeing regulatory compliance and stimulating sustainability in research practices. Future the proposed model is essential for the application of applicable technologies in the bioinformatics field to save time and resources in the drug discovery process.

DECLARATION

Conflict of interest: The author reports that there is no conflict of interest
Funding: This research received no external funding.
Acknowledgement: None
Data availability: Data sharing not applicable to this article as no datasets were generated.
Ethical statement for human participant: Not applicable for this research

REFERENCES

1. S. Menon. Bioinformatics approaches to understand gene looping in human genome. *EPRA International Journal of Research & Development* (IJRD), **6**, 170–173 (2021). Doi: https://doi.org/10.36713/epra7666
2. Y.L. Orlov, A.V. Baranova & T.V. Tatarinova. Bioinformatics methods in medical genetics and genomics. *Int. J. Mol. Sci.*, **21**(17), 6224 (2020). Doi: https://doi.org/10.3390/ijms21176224
3. J. K. Eng & E.W. Deutsch. Extending Comet for global amino acid variant and post-translational modification analysis using the PSI extended FASTA format. *Proteomics*, **20**(21-22), 1900362 (2020). Doi: https://doi.org/10.1002/pmic.201900362
4. M. Afzal, S.R. Islam, M. Hussain & S. Lee. Precision medicine informatics: principles, prospects, and challenges. *IEEE Access*, **8**, 13593–13612 (2020). Doi: https://doi.org/2020. 10.1109/ACCESS.2020.2965955

5. J. Gauthier, A.T. Vincent, S.J. Charette & N. Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, **20**(6), 1981–1996 (2019). Doi: https://doi.org/10.1093/bib/bby063
6. R.Z. Sangeda, A.D. Mwakilili, U. Masamu, S. Nkya, L.A. Mwita, D.P. Massawe, *et al.* Baseline evaluation of bioinformatics capacity in Tanzania. *Frontiers in Education*, **6**, 665313 (2021). Doi: https://doi.org/10.3389/feduc.2021.665313
7. National Human Genome Research Institute (NHGRI). DNA Sequencing Costs: Data. 2018. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
8. M. Liu, J. Liu, Y. Chen, M. Wang, H. Chen & Q. Zheng. AHNG: representation learning on attributed heterogeneous network. *Information Fusion*, **50**, 221–230 (2019). Doi: https://doi.org/10.1016/j.inffus.2019.01.005
9. S. Vascon, M. Frasca, R. Tripodi, G. Valentini & M. Pelillo. Protein function prediction as a graph-transduction game. *Pattern Recognition Letters*, **134**, 96–105 (2020). Doi: https://doi.org/10.1016/j.patrec.2018.04.002
10. Z.-C. Zhang, X.-F. Zhang, M. Wu, L. Ou-Yang, X.-M. Zhao & X.-L. Li. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics*, **36**(11), 3474–3481 (2020). Doi: https://doi.org/10.1093/bioinformatics/btaa157
11. L. Wang, Z.-H. You, Y.-M. Li, K. Zheng & Y.-A. Huang. GCNCDA: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm. *PLoS Computational Biology*, **16**(5), e1007568 (2020). Doi: https://doi.org/10.1371/journal.pcbi.1007568
12. N. Zong, R.S.N. Wong, Y. Yu, A. Wen, M. Huang & N. Li. Drug–target prediction utilizing heterogeneous bio-linked network embeddings. *Briefings in Bioinformatics*, **22**(1), 568–580 (2021). Doi: https://doi.org/10.1093/bib/bbz147
13. H.-C. Yi, Z.-H. You, D.-S. Huang, Z.-H. Guo, K.C. Chan & Y. Li. Learning representations to predict intermolecular interactions on large-scale heterogeneous molecular association network. *iScience*, **23**(7), 101261 (2020). Doi: https://doi.org/10.1016/j.isci.2020.101261
14. Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou & J. Tang. Representation learning for attributed multiplex heterogeneous network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019; pp. 1358–1368. Doi: https://doi.org/10.1145/3292500.3330964
15. N. Schaduangrat, S. Lampa, S. Simeon, M.P. Gleeson, O. Spjuth & C. Nantasenamat. Towards reproducible computational drug discovery. *Journal of Cheminformatics*, **12**, 9 (2020). Doi: https://doi.org/10.1186/s13321-020-0408-x
16. M.R. Javed. CADD and molecular dynamic simulations: potential impacts to conventional medicines. *Combinatorial Chemistry & High Throughput Screening*, **25**(4), 658–659 (2022). Doi: https://doi.org/10.2174/1386207325666220104231827
17. L. Pinzi & G. Rastelli. Molecular docking: shifting paradigms in drug discovery. *International Journal of Molecular Sciences*, **20**(18), 4331 (2019). Doi: https://doi.org/10.3390/ijms20184331
18. M. Rudrapal & D. Chetia. Virtual screening, molecular docking and QSAR studies in drug discovery and development programme. *Journal of Drug Delivery and Therapeutics*, **10**(4), 225–233 (2020). Doi: https://doi.org/10.22270/jddt.v10i4.4218
19. L. Zhao, H.L. Ciallella, L.M. Aleksunes & H. Zhu. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discovery Today*, **25**(9), 1624–1638 (2020). Doi: https://doi.org/10.1016/j.drudis.2020.07.005
20. S. Namasudra. Data access control in the cloud computing environment for bioinformatics. *International Journal of Applied Research in Bioinformatics* (IJARB), **11**(1), 40–50 (2021). Doi: https://doi.org/10.4018/IJARB.2021010105
21. S. Puertas-Martín, A.J. Banegas-Luna, M. Paredes-Ramos, J.L. Redondo, P.M. Ortigosa, O.'h.O. Brovarets', *et al.* Is high performance computing a requirement for novel drug discovery and how will this impact academic efforts?. *Expert Opinion on Drug Discovery*, **15**(9), 981–985 (2020). Doi: https://doi.org/10.1080/17460441.2020.1758664

22. Y.-L. Wang, F. Wang, X.-X. Shi, C.-Y. Jia, F.-X. Wu, G.-F. Hao, *et al.* Cloud 3D-QSAR: a web tool for the development of quantitative structure–activity relationship models in drug discovery. *Briefings in Bioinformatics*, **22**(4), bbaa276 (2021). Doi: https://doi.org/10.1093/bib/bbaa276
23. A. Olğaç, A. Türe, S. Olğaç & S. Möller. Cloud-based high throughput virtual screening in novel drug discovery. In: J. Kołodziej & H. González-Vélez (editors). *High-Performance Modelling and Simulation for Big Data Applications.* Lecture Notes in Computer Science, vol. 11400. Springer, Cham, 2019; pp. 250–278. Doi: https://doi.org/10.1007/978-3-030-16272-6_9
24. B. Seth, S. Dalal & R. Kumar. Securing bioinformatics cloud for big data: Budding buzzword or a glance of the future. In: R. Kumar & U. Wiil (editors). *Recent Advances in Computational Intelligence*. Studies in Computational Intelligence, vol. 823. Springer, Cham, 2019; pp. 121–147. Doi: https://doi.org/10.1007/978-3-030-12500-4_8
25. S. Armoogum & X. Li. Big data analytics and deep learning in bioinformatics with Hadoop. In: A.K. Sangaiah (editor). *Deep Learning and Parallel Computing Environment for Bioengineering Systems*. Academic Press, 2019; pp. 17–36. Doi: https://doi.org/10.1016/B978-0-12-816718-2.00009-9
26. L. Frye, S. Bhat, K. Akinsanya & R. Abel. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discovery Today: Technologies*, **39**, 111–117 (2021). Doi: https://doi.org/10.1016/j.ddtec.2021.08.001
27. A.S. Rifaioglu, H. Atas, M.J. Martin, R. Cetin-Atalay, V. Atalay & T. Doğan. Recent applications of deep learning and machine intelligence on *in silico* drug discovery: Methods, tools and databases. *Briefings in Bioinformatics*, **20**(5), 1878–1912 (2019). Doi: https://doi.org/10.1093/bib/bby061
28. H. Khawar, M.M. Zaheer, H. Khan, U. Akhtar, L. Bukhari, R. Zamir, *et al.* Drug design assisted by computers: New tools in bioinformatics. *Journal of Survey in Fisheries Sciences*, **10**(3), 2998–3035 (2023). Doi: https://doi.org/10.53555/sfs.v10i3.2179
29. X.-M. Zhang, L. Liang, L. Liu & M.-J. Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, **12**, 690049 (2021). Doi: https://doi.org/10.3389/fgene.2021.690049
30. L. Gong & Q. Cheng. Exploiting edge features for graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2019; pp. 9211–9219. Doi: https://doi.org/10.1109/CVPR.2019.00943
31. T. Nguyen, H. Le, T.P. Quinn, T. Nguyen, T.D. Le & S. Venkatesh. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, **37**(8), 1140–1147 (2021). Doi: https://doi.org/10.1093/bioinformatics/btaa921
32. N. Sheng, H. Cui, T. Zhang & P. Xuan. Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA–disease association prediction. *Briefings in Bioinformatics*, **22**(3), bbaa067 (2021). Doi: https://doi.org/10.1093/bib/bbaa067
33. H.E. Manoochehri & M. Nourani. Drug-target interaction prediction using semi-bipartite graph model and deep learning. *BMC Bioinformatics*, **21**, 248 (2020). Doi: https://doi.org/10.1186/s12859-020-3518-6
34. T. Zhao, Y. Hu, L. R. Valsdottir, T. Zang & J. Peng. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in Bioinformatics*, **22**(2), 2141–2150 (2021). Doi: https://doi.org/10.1093/bib/bbaa044
35. J. Peng, Y. Wang, J. Guan, J. Li, R. Han, J. Hao, *et al.* An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Briefings in Bioinformatics*, **22**(5), bbaa430 (2021). Doi: https://doi.org/10.1093/bib/bbaa430
36. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang & S.Y. Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**(1), 4–24 (2020). Doi: https://doi.org/10.1109/TNNLS.2020.2978386
37. H. Tan, Q. Sun, G. Li, Q. Xiao, P. Ding, J. Luo, *et al.* Multiview consensus graph learning for lncRNA–disease association prediction. *Frontiers in Genetics*, **11**, 89 (2020). Doi: https://doi.org/10.3389/fgene.2020.00089
38. A.C. Mara, J. Lijffijt, S. Günnemann & T.D. Bie. A systematic evaluation of node embedding robustness. Proceedings of the First Learning on Graphs Conference (LoG 2022), PMLR 198, Virtual

Event, December 9–12, 2022. 2022; 42 p. URL: https://proceedings.mlr.press/v198/mara22a/mara22a.pdf

39. M.A. Thafar, R.S. Olayan, H. Ashoor, S. Albaradei, V.B. Bajic, X. Gao, *et al.* DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, **12**, 44 (2020). Doi: https://doi.org/10.1186/s13321-020-00447-2
40. X. Tang, J. Luo, C. Shen & Z. Lai. Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction. *Briefings in Bioinformatics*, **22**(6), bbab174 (2021). Doi: https://doi.org/10.1093/bib/bbab174
41. W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, *et al.* Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics*, **19**, 233 (2018). Doi: https://doi.org/10.1186/s12859-018-2220-4
42. C. Ruiz, M. Zitnik & J. Leskovec. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications*, **12**, p. 1796 (2021). Doi: https://doi.org/10.1038/s41467-021-21770-8
43. R. Celebi, H. Uyar, E. Yasar, O. Gumus, O. Dikenelli & M. Dumontier. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinformatics*, **20**, 726 (2019). Doi: https://doi.org/10.1186/s12859-019-3284-5
44. M.R. Karim, M. Cochez, J.B. Jares, M. Uddin, O. Beyan & S. Decker. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019; pp. 113–123. Doi: https://doi.org/10.1145/3307339.3342161
45. C. Park, J. Park & S. Park. AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*, **159**, 113538 (2020). Doi: https://doi.org/10.1016/j.eswa.2020.113538
46. S. Pan, R. Hu, S.-f. Fung, G. Long, J. Jiang & C. Zhang. Learning graph embedding with adversarial training methods. *IEEE Transactions on Cybernetics*, **50**, 2475–2487 (2019). Doi: https://doi.org/10.1109/TCYB.2019.2932096
47. M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou & F. Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics*, **21**(3), 919–935 (2020). Doi: https://doi.org/10.1093/bib/bbz042
48. T. Wu, Y. Wang, Y. Wang, E. Zhao & Y. Yuan. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Scientific Reports*, **11**, 5858 (2021). Doi: https://doi.org/10.1038/s41598-021-85255-w
49. L. Wright & S. Davidson. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, **7**, 13 (2020). Doi: https://doi.org/10.1186/s40323-020-00147-4
50. M.N. Kamel-Boulos & P. Zhang. Digital twins: From personalised medicine to precision public health. *Journal of Personalized Medicine*, **11**(8), 745 (2021). Doi: https://doi.org/10.3390/jpm11080745
51. H. Garg. Digital twin technology: Revolutionary to improve personalized healthcare. *Science Progress and Research* (SPR), **1**, 32–34 (2021). Doi: https://doi.org/10.52152/spr/2021.105
52. K. Bruynseels, F. Santoni de Sio & J. Van den Hoven. Digital twins in health care: ethical implications of an emerging engineering paradigm. *Frontiers in Genetics*, **9**, 31 (2018). Doi: https://doi.org/10.3389/fgene.2018.00031
53. R.M. Portela, C. Varsakelis, A. Richelle, N. Giannelos, J. Pence, S. Dessoy, *et al.* When is an *in silico* representation a digital twin? A biopharmaceutical industry approach to the digital twin concept. In: C. Herwig, R. Pörtner & J. Möller (editores). *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Advances in Biochemical Engineering/Biotechnology, vol 176. Springer, Cham, 2021; pp. 35–55. Doi: https://doi.org/10.1007/10_2020_138
54. E.O. Popa, M. van Hilten, E. Oosterkamp & M.-J. Bogaardt. The use of digital twins in healthcare: Socio-ethical benefits and socio-ethical risks. *Life Sciences, Society and Policy*, **17**, 6 (2021). Doi: https://doi.org/10.1186/s40504-021-00113-x

55. S. Nargund, K. Guenther & K. Mauch. The move toward biopharma 4.0: insilico biotechnology develops "smart" processes that benefit biomanufacturing through digital twins. *Genetic Engineering & Biotechnology News*, **39**(6), 53–55 (2019). Doi: https://doi.org/10.1089/gen.39.06.18
56. Z. Huang, Y. Shen, J. Li, M. Fey & C. Brecher. A survey on AI-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics. *Sensors*, **21**(19), 6340 (2021). Doi: https://doi.org/10.3390/s21196340
57. B.R. Barricelli, E. Casiraghi & D. Fogli. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, **7**, 167653–167671 (2019). Doi: https://doi.org/10.1109/ACCESS.2019.2953499
58. R. Kaul, C. Ossai, A.R.M. Forkan, P.P. Jayaraman, J. Zelcer, S. Vaughan, *et al.* The role of AI for developing digital twins in healthcare: The case of cancer care. *WIREs: Data Mining and Knowledge Discovery*, **13**(1), e1480 (2023). Doi: https://doi.org/10.1002/widm.1480
59. T. Sun, X. He & Z. Li. Digital twin in healthcare: Recent updates and challenges. *Digital Health*, **9**, 20552076221149651 (2023). Doi: https://doi.org/10.1177/20552076221149651
60. H. Li, S. Tian, Y. Li, Q. Fang, R. Tan, Y. Pan, *et al.* Modern deep learning in bioinformatics. *Journal of Molecular Cell Biology*, **12**(11), 823–827 (2020). Doi: https://doi.org/10.1093/jmcb/mjaa030
61. V.D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A.G. Papadiamantis, V. Aidinis, et al. Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences*, **22**(4), 1676 (2021). Doi: https://doi.org/10.3390/ijms22041676
62. W. Wang & X. Gao. Deep learning in bioinformatics and biomedicine. *Methods*, **209**, 39–39 (2023). Doi: https://doi.org/10.1016/j.ymeth.2022.12.001
63. R. Vijayan, J. Kihlberg, J.B. Cross & V. Poongavanam. Enhancing preclinical drug discovery with artificial intelligence. *Drug Discovery Today*, **27**(4), 967–984 (2022). Doi: https://doi.org/10.1016/j.drudis.2021.11.023
64. Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan & X. Gao. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, **166**, 4–21 (2019). Doi: https://doi.org/10.1016/j.ymeth.2019.04.008
65. S. Quazi & Z. Fatima. Role of artificial intelligence and machine learning in bioinformatics: Drug discovery and drug repurposing. In: Information Resources Management Association. *Research Anthology on Bioinformatics, Genomics, and Computational Biology*. IGI Global, 2024; 1394–1405. Doi: https://doi.org/10.4018/979-8-3693-3026-5.ch062
66. N. Auslander, A.B. Gussow & E.V. Koonin. Incorporating machine learning into established bioinformatics frameworks. *International Journal of Molecular Sciences*, **22**(6), 2903 (2021). Doi: https://doi.org/10.3390/ijms22062903
67. B. Tang, Z. Pan, K. Yin & A. Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics*, **10**, 214 (2019). Doi: https://doi.org/10.3389/fgene.2019.00214
68. S.M. Kandathil, J.G. Greener & D.T. Jones. Recent developments in deep learning applied to protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, **87**(12), 1179–1189 (2019). Doi: https://doi.org/10.1002/prot.25824
69. Y. Cao, T.A. Geddes, J.Y.H. Yang & P. Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, **2**, 500–508 (2020). Doi: https://doi.org/10.1038/s42256-020-0217-y
70. L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, **36**(16), 4406–4414 (2020). Doi: https://doi.org/10.1093/bioinformatics/btaa524
71. J. Wang, X. Liu, S. Shen, L. Deng & H. Liu. DeepDDS: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings in Bioinformatics*, **23**(1), bbab390 (2022). Doi: https://doi.org/10.1093/bib/bbab390
72. S. Liu, X. Xu, Z. Yang, X. Zhao, S. Liu & W. Zhang. EPIHC: Improving enhancer-promoter interaction prediction by using hybrid features and communicative learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **19**(6), 3435–3443 (2021). Doi: https://doi.org/10.1109/TCBB.2021.3109488

73. M.M. Fareed. Impact of deep learning-DL in the modern computational biology. *European Journal of Volunteering and Community-Based Projects*, **1**(4), 35–48 (2021). URL: https://journal.odvcasarco-baleno.it/index.php/ejvcbp/article/view/56
74. N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason & J. Bostrom. Deep reinforcement learning for multiparameter optimization in *de novo* drug design. *Journal of Chemical Information and Modeling*, **59**(7), 3166–3176 (2019). Doi: https://doi.org/10.1021/acs.jcim.9b00325
75. L. Patel, T. Shukla, X. Huang, D.W. Ussery & S. Wang. Machine learning methods in drug discovery. *Molecules*, **25**(22), 5277 (2020). Doi: https://doi.org/10.3390/molecules25225277
76. Y. Deng, Y. Qiu, X. Xu, S. Liu, Z. Zhang, S. Zhu, *et al.* META-DDIE: predicting drug–drug interaction events with few-shot learning. *Brief. Bioinform.*, **23**(1), bbab514 (2022). Doi: https://doi.org/10.1093/bib/bbab514
77. S. Gull & F. Minhas. AMP 0: Species-specific prediction of anti-microbial peptides using zero and few shot learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **19**(1), 275–283 (2020). Doi: https://doi.org/10.1109/TCBB.2020.2999399
78. S. Sen, R. Das, S. Dasgupta & U. Maulik. Application of deep architecture in bioinformatics. In: S. Dash, B. Acharya, M. Mittal, A. Abraham & A. Kelemen (editors). *Deep Learning Techniques for Biomedical and Health Informatics*. Studies in Big Data, vol 68. Springer, Cham, 2020; pp. 167-186. Doi: https://doi.org/10.1007/978-3-030-33966-1_9
79. L. Rundo, A. Tangherloni & C. Militello. Artificial intelligence applied to medical imaging and computational biology. *Appl. Sci.*, **12**(18), 9052 (2022). Doi: https://doi.org/10.3390/app12189052
80. H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, *et al.* ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Molecular Therapy: Nucleic Acids*, **17**, 1–9, 2019. Doi: https://doi.org/10.1016/j.omtn.2019.04.025
81. L.A. Bugnon, C. Yones, D.H. Milone & G. Stegmayer. Deep neural architectures for highly imbalanced data in bioinformatics. *IEEE Transactions on Neural Networks and Learning Systems*, **31**(8), 2857–2867 (2019). Doi: https://doi.org/10.1109/TNNLS.2019.2914471
82. B. Pfeifer, A. Saranti & A. Holzinger. GNN-SubNet: disease subnetwork detection with explainable graph neural networks. *Bioinformatics*, **38**(Suppl. 2), ii120–ii126 (2022). Doi: https://doi.org/10.1093/bioinformatics/btac478

HOW TO CITE THIS ARTICLE

A.I. Oraibi, A.M. Zheoat, H.N. Sameer, M.A. Alboreadi & H.M. Abdulhamza. Drug design and discovery with bioinformatics tools. *Rev. Colomb. Cienc. Quim. Farm.*, **54**(3), 621–642 (2025). Doi: https://doi.org/10.15446/rcciquifa.v54n3.118428