INTERNATIONAL JOURNAL OF
INTEGRATIVE AND
MODERN MEDICINE

A Systematic Review of Computational Innovations for Drug Discovery Against Diphtheria in Nigeria

Paul Alemoh Owhenagbo
University of East London

Dr. Ugochukwu Osigwe
African Field Epidemiology Network, Abuja, Nigeria

Annotation: This systematic review explores the impact of computational innovations in drug discovery aimed at combating diphtheria in Nigeria. Diphtheria remains a significant health concern in less developed regions, where traditional drug discovery methods often face challenges of inefficiency and high costs. This review highlights the potential of advanced computational techniques, such as Computer-Aided Drug Design (CADD) and Ligand-Based Drug Design (LBDD), as transformative, cost-effective solutions. By systematically analyzing data from Chembl and PubChem databases, this study applies machine learning algorithms to predict bioactivity of compounds targeting the diphtheria toxin. The findings suggest that computational tools, including machine learning models, can significantly enhance the identification and development of effective treatments for diphtheria, thereby alleviating some of the burden on local healthcare systems. Additionally, this review identifies existing gaps in current research and suggests future directions, emphasizing the need for specialized training and standardized protocols in computational drug design to optimize the efficacy of these innovative approaches.

Keywords: CADD, LBDD, Diphtheria, Drug, Machine Learning Algorithms.

1.0. Introduction

Diphtheria, a life-threatening bacterial infection, primarily affects the respiratory tract and has historically been a significant cause of child mortality. While the introduction of immunization programs in the mid-20th century led to a drastic reduction in cases globally, the disease continues to persist in areas with limited vaccine coverage and healthcare resources. In Nigeria, the resurgence of diphtheria cases in recent years, coupled with an increase in mortality, emphasizes the need for innovative approaches to combat this infectious disease. The current outbreak reported by the Nigeria Centre for Disease Control (NCDC) highlights an urgent health crisis, as diphtheria cases in multiple states reveal a high incidence of unvaccinated children and substantial case fatality rates. The inadequacy of traditional drug discovery processes and the rising need for rapid and cost-effective therapeutic interventions underscore the importance of alternative methods, particularly computational approaches.

Prior to the widespread accessibility of immunizations, diphtheria was a primary factor contributing to child mortality (Zakikhany & Efstratiou, 2012). Diphtheria, a notably contagious illness, is characterized by severe morbidity and mortality (Sharma et al., 2019). The etiological agent of diphtheria is primarily Corynebacterium diphtheriae, a nonmotile, Gram-positive bacterium, belonging to the Corynebacterium species. Transmission of this pathogen typically occurs through respiratory droplets. Upon infection, the bacterium proliferates within the respiratory mucosa. A key aspect of its pathogenicity is the release of exotoxins, which are responsible for inducing both localized and systemic tissue damage (Organization, 2018). Moreover, the primary mechanism by which Corynebacterium diphtheriae, the micro-organism responsible for causing diphtheria, exerts its harmful effects is by the secretion of a toxin. This toxin could hinder the process of cellular protein synthesis, resulting in harm to the nearby tissue and the creation of a unique pseudo-membrane at the infection site. The bacteria generally exhibit an incubation period that spans from 2 to 5 days, with the potential to stretch up to 10 days. It has the potential to impact different mucous membranes throughout the human body (Acosta et al., 2021).

Furthermore, Diphtheria manifests in two main variants: respiratory and non-respiratory, with the respiratory variant being linked to a greater risk of death. The most common form is respiratory diphtheria, which is classified clinically according to the specific anatomical area involved, such as pharyngeal, tonsillar, laryngeal, and nasal diphtheria (Acosta et al., 2021). The occurrence of this kind is usually preceded by preliminary signs, including a mild temperature (often below 38.3°C), runny nose, sore throat, inflammation of the conjunctiva, cough, overall discomfort, and loss of appetite. In severe instances, the development of a pseudo-membrane, mainly on the tonsils, might spread to nearby regions, potentially resulting in blockage of the airway. Manifestations such as voice changes, a harsh cough, and notable swelling and inflammation of the glands under the jaw can result in the distinctive "bull neck" appearance, which signifies the necessity for careful patient observation.

The non-respiratory manifestation of diphtheria includes cutaneous diphtheria, which is characterized by a scaly rash and ulcers with well-defined borders. It is often found in conjunction with long-lasting skin lesions. There are also fewer common forms of diphtheria that affect mucous membranes in areas such as the conjunctiva, auditory canal, and vulvovaginal region. These types typically arise from non-toxigenic strains of Corynebacterium (Acosta et al., 2021).

Typical problems that can occur because of diphtheria infection include inflammation of the heart muscle (myocarditis) and damage to many nerves (polyneuropathies). Other potential consequences may involve nephritis, corneal scarring (which might worsen due to a lack of vitamin A), encephalitis, diarrhea, pneumonia, and subacute sclerosing panencephalitis (Besa et al., 2014)

The implementation and subsequent extensive acceptance of the diphtheria-tetanus-pertussis (DTP) vaccine, especially after World War II, resulted in a swift and substantial decrease in the occurrence of this illness in developed countries. A simultaneous decline in the occurrence of diphtheria was noted in underdeveloped nations with the implementation of the World Health Organization (WHO) Expanded Programme on Immunization in 1974. This initiative promoted the implementation of a three-dose regimen of the DTP immunization for all infants during the first six months of their lives. This recommendation was essential in significantly decreasing the worldwide impact of diphtheria. Hence, the international incidence of diphtheria has been substantially diminished in developed nations, and significant strides have been made in its control over recent decades in low- and middle-income countries, including Nigeria (Clarke, 2017). The integration of the diphtheria vaccine into immunization programs has been instrumental in advancing global initiatives aimed at its eradication. However, Nigeria is presently witnessing a concerning upsurge in the incidence of diphtheria cases nationwide, a development that necessitates urgent attention and intervention (NCDC, 2023). According to the guidelines established by the Nigeria Centre for Disease Control (NCDC), a suspected case of diphtheria is identified based on clinical presentation, specifically characterized by an upper respiratory tract illness. This illness is typified by symptoms such as pharyngitis, nasopharyngitis, tonsillitis, or laryngitis, coupled with the presence of an adherent pseudo-membrane in the pharynx, tonsils, larynx, and/or nasal area. In contrast, a laboratory-confirmed case of diphtheria is defined as an individual from whom Corynebacterium spp. has been isolated via culture methods and has been found positive for toxin production, as determined by the modified Elek test. This confirmation is deemed valid regardless of the presence or absence of clinical symptoms (NCDC, 2023).

The resurgence of diphtheria in Nigeria presents a pressing public health challenge that underscores the limitations of current treatment and prevention measures. As a bacterial infection caused by *Corynebacterium diphtheriae*, diphtheria is highly contagious and can lead to severe respiratory illness, systemic toxicity, and potentially fatal outcomes, especially in young children and those who are unvaccinated. The recent outbreak, as documented by the Nigeria Centre for Disease Control (NCDC), reveals alarming trends, including high rates of infection in regions with limited access to healthcare and vaccination services. According to NCDC data, the outbreak has spread across multiple states, with significant mortality rates among confirmed cases, primarily affecting unvaccinated children aged 1-14 years. These statistics highlight the need for comprehensive intervention strategies that go beyond vaccination to include novel therapeutic approaches.

Traditional drug discovery methods are often time-consuming, labor-intensive, and costly, posing substantial barriers for countries with limited healthcare resources. The limitations of these methods are particularly evident in the context of diphtheria, where the rapid development of effective therapeutics could save lives and alleviate pressure on healthcare systems. Consequently, there is an urgent need for innovative drug discovery approaches that can provide rapid, efficient, and affordable solutions to emerging health crises such as diphtheria outbreaks. Computational approaches to drug discovery have emerged as promising alternatives to conventional methods. By leveraging computational models, machine learning algorithms, and extensive bioinformatics databases, researchers can now identify and optimize potential therapeutic compounds more efficiently and at a fraction of the traditional cost.

In this context, Computer-Aided Drug Design (CADD) and ligand-based drug design (LBDD) represent critical tools in modern pharmacology, offering valuable insights into the molecular mechanisms of disease and facilitating the identification of compounds with high therapeutic potential. These computational techniques enable the analysis of vast datasets from chemical libraries, such as Chembl and PubChem, which contain detailed information on bioactive compounds. Machine learning algorithms, particularly models like RandomForest, have proven effective in predicting the bioactivity of these compounds, thus enabling researchers to focus on the most promising candidates for further investigation.

The aim of this study is to explore computational innovations in drug discovery tailored to combat diphtheria in Nigeria. By leveraging machine learning and bioinformatics, we assess the potential of computational techniques, such as Computer-Aided Drug Design (CADD) and ligand-based drug design (LBDD), in identifying new and effective compounds against diphtheria toxin. This approach not only addresses the immediate need for novel therapeutic options but also offers a cost-efficient solution that aligns with Nigeria's healthcare context. Through the use of machine learning algorithms to predict compound bioactivity against diphtheria toxin, this study demonstrates the applicability of computational tools in advancing treatment options for neglected diseases in resource-constrained regions.

Research Questions

The following research questions guided the study:

1. What are the most effective strategies for optimizing ligand selection and enhancement in ligand-based drug design (LBDD) to maximize the efficiency of the drug discovery process?
2. What are the different types of machine learning that has contributed to the advancements in computational drug discovery methodologies?
3. What are the various types of machine learning algorithms, and how do they specifically enhance the effectiveness and efficiency of computational drug discovery processes?
4. How do computational strategies enhance the identification of therapeutic targets and optimize drug discovery processes for diphtheria?

2.0. Methodology

This systematic review explores computational innovations in drug discovery aimed at combatting diphtheria in Nigeria, covering studies published from 2011 to 2023. The review's primary goal was to gather and synthesize research on computational drug discovery techniques, focusing on their potential to address the challenges of diphtheria treatment. A broad literature search was performed across various academic and chemical databases, including PubMed, Scopus, Web of Science, Chembl, and PubChem, to capture a diverse selection of studies within the specified timeframe. The search strategy incorporated targeted keywords and phrases, such as "computational drug discovery," "machine learning," "bioinformatics," "diphtheria," "ligand-based drug design," and "Computer-Aided Drug Design (CADD)," using Boolean operators (AND, OR) to refine and optimize search results. Inclusion criteria focused on peer-reviewed research published within the last decade, emphasizing computational methods addressing bacterial infections with an application to diphtheria. Exclusion criteria eliminated conference abstracts, grey literature, and non-English publications to maintain data reliability. Eligible studies underwent data extraction using a standardized template to ensure consistency in key study details: author(s), publication year, journal name, and study design. Methodological details were also documented, focusing on the computational techniques used, specific algorithms (e.g., Random Forest, Support Vector Machine), and databases accessed (e.g., Chembl, PubChem). Key findings, including predictive accuracy and bioactivity outcomes relevant to diphtheria treatments, were organized in a matrix format to facilitate comparative analysis across studies. The findings were analyzed to reveal existing gaps in the literature, such as limited access to high-quality training data and a lack of standardized protocols in computational drug design. The review concludes with recommendations for future research, emphasizing the need for interdisciplinary collaboration and specialized training to support advancements in computational drug discovery methods for infectious diseases like diphtheria.

3.0. Data Analysis and Result

3.1. Demographic Data

In Nigeria, a significant diphtheria outbreak has been reported, encompassing a total of 4,160 suspected cases across 27 states and 139 Local Government Areas (LGAs) as shown in the figure below. Many of these cases were concentrated in specific states, with Kano (3,233 cases), Yobe (477 cases), Katsina (132 cases), Kaduna (101 cases), Bauchi (54 cases), FCT (41 cases), and Lagos (30 cases) accounting for 97.8% of the suspected cases. Of these, 1,534 cases (36.9%) were confirmed, either through laboratory tests (87 cases), epidemiological links (158 cases), or clinical compatibility (1,289 cases). A significant number of cases, 1,700 (40.9%), were discarded, while 639 (15.4%) are pending classification, and 287 (6.9%) remain unknown. The confirmed cases spanned across 56 LGAs in 11 states. Notably, most of the confirmed cases, about 1,018 (66.4%), were among children aged 1 – 14 years. The outbreak has also been marked by a concerning number of fatalities, with 137 deaths recorded among the confirmed cases, translating to a case fatality rate (CFR) of 8.9%. Alarmingly, a significant proportion of the confirmed cases, 1,257 (81.9%), occurred in individuals who were not fully vaccinated against diphtheria. This data underscores the critical need for enhanced vaccination efforts and public health measures in Nigeria, particularly targeting the most affected states and age groups to control and prevent the spread of this serious bacterial infection(NCDC, 2023).

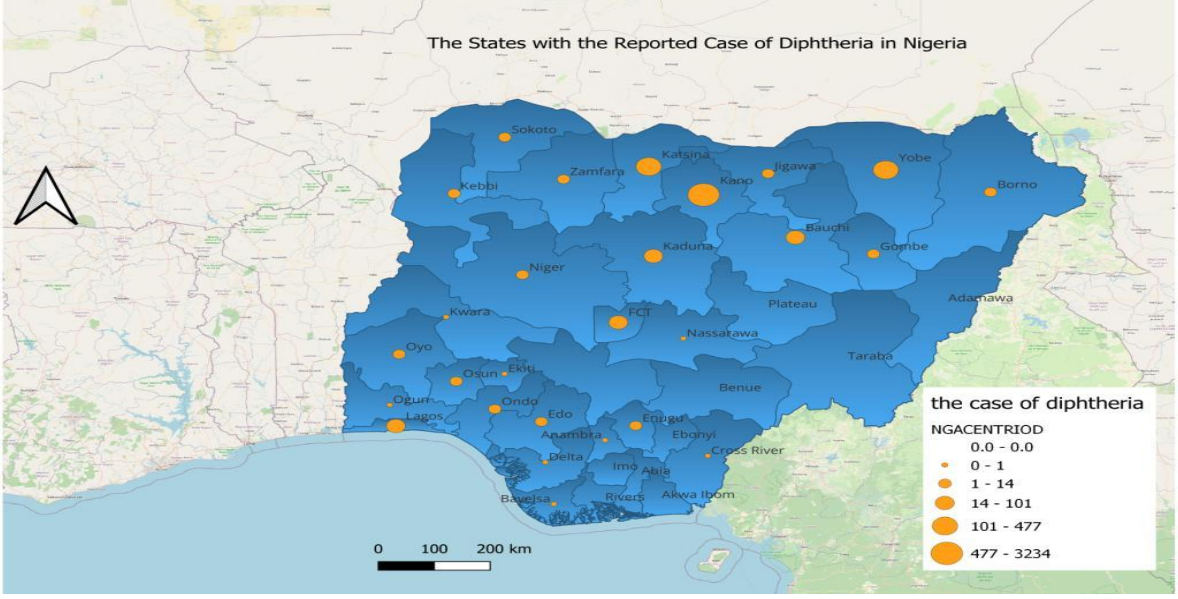

Figure 2-The Geographical distribution of Diphtheria cases in Nigeria

Table 1 - Description of Diphtheria Cases by States, May 2022 – July 2023 (1/2)

States	Suspected Cases	Confirmed Cases (%)	Discarded Cases (%)	Pending Cases (%)	Unknown Cases (%)	Deaths (CFR) (%)
Kano	3,234	1,207 (37.3%)	1,480 (45.8%)	477 (14.7%)	70 (2.2%)	100 (8.3%)
Yobe	477	252 (52.8%)	75 (15.7%)	38 (8.0%)	112 (23.5%)	23 (9.1%)
Katsina	132	9 (6.8%)	29 (22.0%)	56 (42.4%)	38 (28.8%)	2 (22.2%)
Kaduna	101	5 (5.0%)	33 (32.7%)	21 (20.8%)	42 (41.5%)	0
Bauchi	54	41 (75.9%)	10 (18.5%)	3 (5.6%)	0	6 (14.6%)
FCT	41	6 (14.6%)	6 (14.6%)	28 (68.3%)	1 (2.4%)	1 (16.7%)
Lagos	30	8 (26.7%)	17 (56.7%)	5 (16.6%)	0	5 (62.5%)
Sokoto	14	0	7 (50.0%)	0	7 (50.0%)	-
Zamfara	13	0	2 (15.4%)	0	11 (84.6%)	-
Niger	11	2 (18.2%)	9 (81.8%)	0	0	0

Table 2 - Description of Diphtheria Cases by States, May 2022 – July 2023 (2/2)

States	Suspected Cases	Confirmed Cases (%)	Discarded Cases (%)	Pending Cases (%)	Unknown Cases (%)	Deaths (CFR) (%)
Jigawa	4	1 (25.0%)	1 (25.0%)	0	1	0
Kebbi	3	0	-	-	-	-
Ondo	2	0	-	-	-	2 (100.0%)
Edo	2	0	-	-	-	2 (100.0%)
Borno	2	0	-	-	-	0
Ogun	1	0	-	-	-	0
Cross River	1	1 (100%)	-	-	-	0
Kwara	1	0	-	-	-	0
Bayelsa	1	0	-	-	-	0
Delta	1	0	-	-	-	0
Nasarawa	1	0	-	-	-	0
Ekiti	1	0	-	-	-	0
Anambra	1	0	-	-	-	0
TOTAL	**4,160**	**1,534 (36.9%)**	**639 (15.4%)**	**287 (6.9%)**	**139**	**137 (8.9%)**

- ✓ ***Lab confirmed (LC): a person with Corynebacterium spp. isolated by culture and positive for toxin production, regardless of symptoms.***
- ✓ ***Epidemiologically linked (EL):*** *a person that meets the definition of a suspected case and is linked epidemiologically to a laboratory-confirmed case.*
- ✓ ***Clin compatible (CC):*** *a person that meets the definition of a suspected case and lacks both a confirmatory laboratory test result and epidemiologic linkage to a laboratory confirmed case.*
- ✓ ***Confirmed case*** *= LC + EL + CC*

Adapted from "Diphtheria health advisory for health care workers amidst outbreak in Nigeria," by Nigeria Centre for Disease Control and Prevention, 2023 (https://ncdc.gov.ng/news/436/diphtheria-health-advisory-for-health-care-workers-amidst-outbreak-in-nigeria).

The data presented in Tables 1 and 2 provides a comprehensive overview of diphtheria cases reported across various Nigerian states from May 2022 to July 2023, highlighting significant variations in the incidence and outcomes of suspected cases. In total, there were 4,160 suspected cases, with 1,534 (36.9%) confirmed as diphtheria, indicating a substantial level of diagnostic activity; however, the high percentage of discarded cases (639 or 15.4%) and pending cases (287 or 6.9%) raises concerns about the efficiency of case management and surveillance systems. Notably, states such as Kano and Yobe reported the highest numbers of suspected cases, along with considerable confirmation rates of 37.3% and 52.8%, respectively, while Katsina and Kaduna exhibited low confirmation rates of 6.8% and 5.0%, underscoring disparities in disease burden and reporting efficacy. Additionally, the case fatality rate (CFR) varied across states, with Lagos exhibiting a high CFR of 62.5% among confirmed cases, contrasting with states like Kaduna, which reported no deaths. Overall, this data not only illustrates the urgency for improved public health interventions in the affected regions but also emphasizes the importance of continued epidemiological surveillance and targeted healthcare strategies to mitigate the impact of diphtheria in Nigeria.

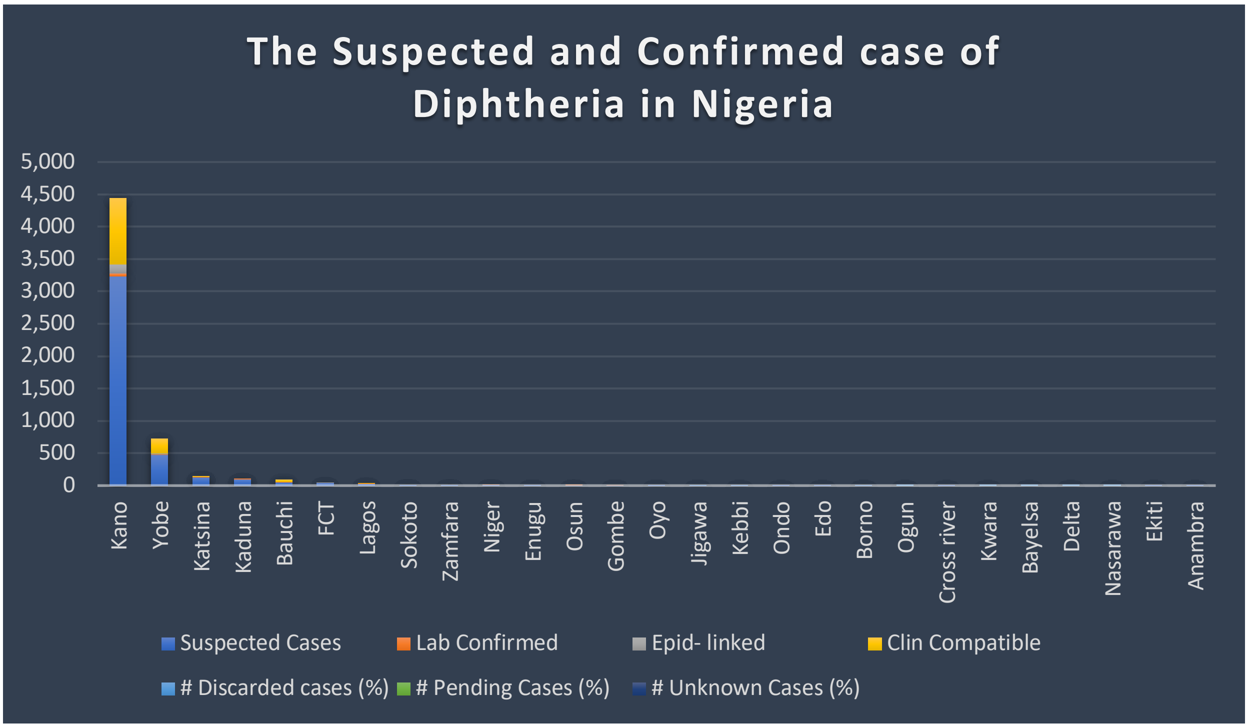

Figure 3-Distribution of Cases

3.2. Data Analysis and Result

Research Question One: What are the most effective strategies for optimizing ligand selection and enhancement in ligand-based drug design (LBDD) to maximize the efficiency of the drug discovery process?

The integration of computational methodologies in the realm of drug discovery has become exceedingly crucial. It involves two main approaches. The structure-based drug design (SBDD) and ligand-based drug design (LBDD). Structure-based drug design (SBDD) includes using knowledge of the three-dimensional structure of the biological target to understand how a possible drug can interact and fit with it (Sliwoski et al., 2014). However, LBDD does not require prior knowledge of the target's structure. Instead, it uses existing drug molecules and their pharmacological features to guide the development of new drug candidates. In this paper, LBDD was been employed. The creation of ligand libraries is an important process that entails choosing and enhancing ligands based on their drug-like qualities and other relevant physicochemical features related to the target of interest. The selection procedure is crucial for maximizing the efficiency of the drug development process. Despite the availability of fast docking methods, the computational binding of millions of molecules requires significant resources. Efficient use of time and resources can be achieved by proactively eliminating compounds that lack drug-like features, unstable, or have negative qualities.

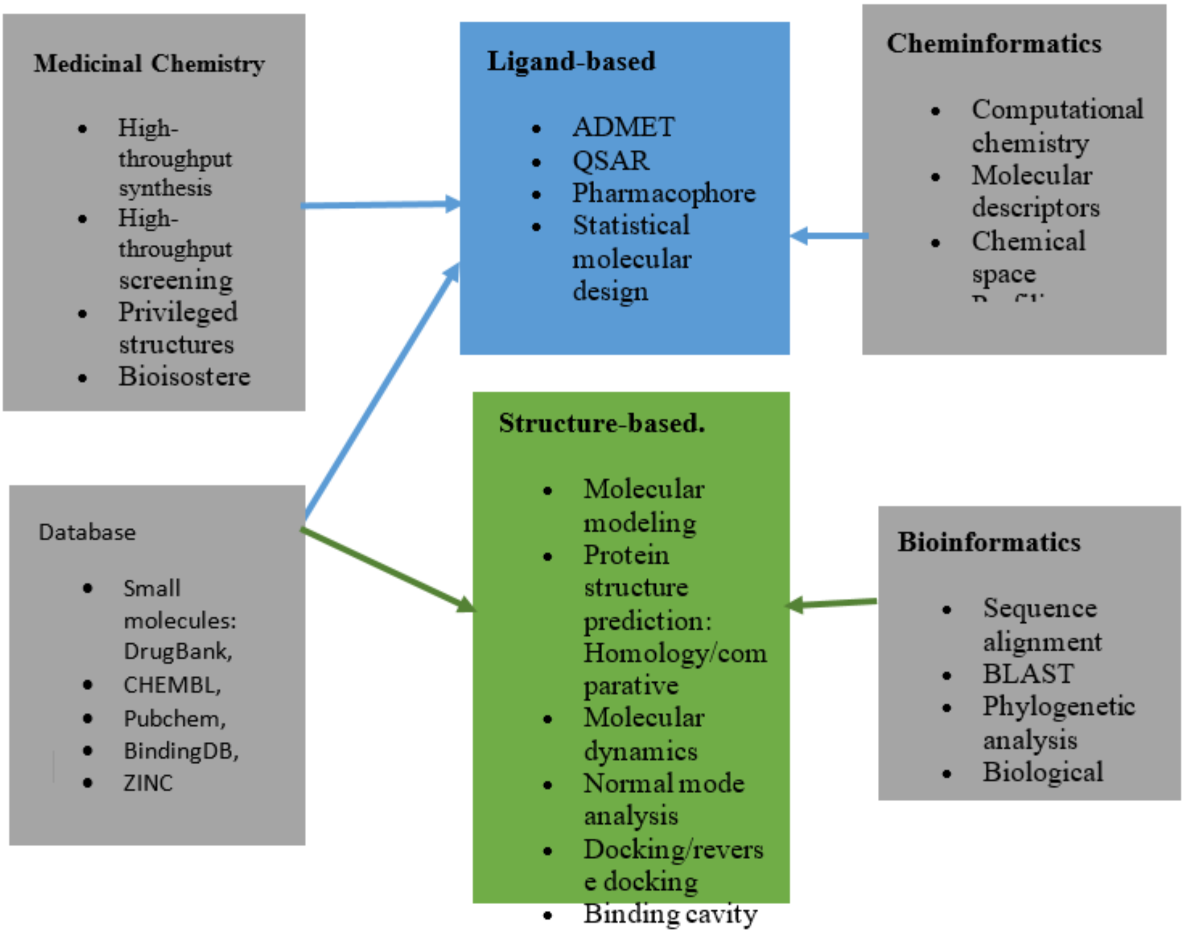

Figure 4- computation of drug discovery with SBDD and LBDD

***Research Question Two:* What are the different types of machine learning that contribute to advancements in computational drug discovery methodologies?**

The utilization of artificial intelligence (AI) is prevalent in both the research industrial and academic sectors. Machine learning (ML), a fundamental element of artificial intelligence (AI), has been incorporated into various domains, including data production and analytics (Patel et al., 2020). This can be further divided into Supervised and Unsupervised.

1. Supervised Learning: This is a machine learning where models are trained with dataset that contains input-output pairs, often known as labeled data. By being exposed to these labeled samples, the algorithm can identify and understand patterns and connections between the input attributes and the related output labels. The primary aim is to acquire a mapping function that can precisely forecast the output labels for instances that have not been encountered before (Rifaioglu et al., 2019). This procedure entails iteratively modifying the parameters of the model to minimize the disparity between its predictions and the actual labels. Supervised learning commonly involves the use of regression algorithms for predicting continuous results and classification methods for categorizing discrete outcomes (Rifaioglu et al., 2019).

2. Unsupervised: Unsupervised learning, on the other hand, works with datasets that do not include explicit labels or established categories. The main objective is to reveal the inherent structures, patterns, or relationships that exist within the data. Unsupervised learning algorithms can independently discover and describe hidden patterns without any prior knowledge of the correct answers. Clustering techniques, such as K-means and hierarchical clustering, are examples of such algorithm that divide the data into separate groups based on similarity criteria, making it easier to identify natural groupings or clusters (Xu & Tian, 2015)

***Research Question Three:* What are the various types of machine learning algorithms, and how do they specifically enhance the effectiveness and efficiency of computational drug discovery processes?**

Table 3: Types of machine learning algorithms

Learning Type	**Method**	**Description**	**Reference**
Supervised Learning	k-Nearest Neighbor	This method classifies an object based on the most frequent label among its 'k' closest neighbors in the dataset, with 'k' being a chosen positive integer.	(Kadir et al., 2020)
	Naive Bayes	It's a probabilistic model that predicts group membership by applying Bayes' theorem with the assumption that the features in the dataset are independent of each other.	(H.-C. Kim et al., 2020)
	Random Forest	This approach uses a collection of decision trees to perform classification tasks, where the majority vote from the trees determines the final output.	(Sun et al., 2024)
	Support Vector Machine	Data points are projected into a higher-dimensional space to identify the optimal hyperplane that maximizes the margin between the closest points of different classes, known as support vectors.	(Maltarollo et al., 2019)
	Independent Component Analysis	This method identifies and separates a multivariate signal into additive, independent components, assuming the statistical independence of the non-Gaussian source signals.	(Monakhova & Rutledge, 2020)
Unsupervised Learning	Hierarchical Clustering	Clusters are formed by either merging smaller clusters into larger ones (agglomerative) or by breaking down a large cluster into smaller clusters (divisive), creating a hierarchy of clusters.	(Reddy & Vinzamuri, 2018)
	k-Means Clustering	This algorithm groups the data into 'k' number of clusters by minimizing the distance between data points and the center of their assigned cluster	(Malhat et al., 2014)
	Principal Component Analysis (PCA)	PCA transforms a set of potentially correlated variables into a set of values of linearly uncorrelated variables, called principal components, through an orthogonal transformation.	(Imaizumi et al., 2020)

1. **Random Forest Algorithm:** Random Forest is an algorithm in machine learning that uses several decision trees to enhance the accuracy of predictions. The algorithm use bootstrap sampling and random feature selection to generate a variety of trees, thereby mitigating the issue of overfitting (Kim et al., 2020). It is proficient in both classification and regression tasks, providing exceptional accuracy, identification of important features, and resilience to missing data. Nevertheless, it is more intricate and requires more computer resources compared to single decision trees. The Random Forest algorithm is extensively utilized across diverse domains, ranging from fraud detection to drug discovery. In the field of drug discovery, Random Forests (RFs) are predominantly employed for feature selection, classification, or regression tasks. Cano and colleagues demonstrated the use of RF techniques to enhance the prediction of ligand-protein affinity. This was achieved through virtual screening, where molecular descriptors were selected based on a training dataset encompassing enzyme ligands, including those for kinases and nuclear hormone receptors. Key advantages of employing RF in drug discovery include speeding up the training process, requiring fewer parameters, the ability to handle missing data, and accommodating nonparametric data (Cano et al., 2017).
2. **Support Vector Algorithm:** Support Vector Machines (SVMs) are crucial in drug discovery since they classify molecules into active and inactive categories by utilizing an ideal hyperplane in a feature-based chemical space (Heikamp & Bajorath, 2014). They demonstrate exceptional skill in establishing decision boundaries by optimizing the margin between distinct classes in an N-dimensional feature space. SVMs are essential for the identification of potential compounds, as they utilize regression models to predict interactions between drugs and ligands. Additionally, SVMs aid in the ranking of compounds based on their probability of exhibiting activity. SVMs employ kernel functions to transform non-linear data into a higher-dimensional space, enabling efficient separation of classes. SVMs in drug-target interaction studies use many data sources, such as ligand and protein information, to enhance prediction accuracy (Heikamp & Bajorath, 2014). The research conducted by Wang et al. illustrates the ability of Support Vector Machines (SVMs) to integrate different types of data, such as drug pharmacology, chemical structures, and genomic information, to predict drug-target interactions and multiple bioactivities. All this are being done through the SVR. Support vector regression (SVR) is a modified version of the SVM technique that can be used to predict numerical property values, such as chemical potency. In Support Vector Regression (SVR), a distinct function is derived from the training data to predict numerical values, rather than constructing a hyperplane for class label prediction (Rodríguez-Pérez et al., 2017). This highlights the usefulness and effectiveness of SVMs in advancing drug discovery efforts.

Research Question Four: How do computational strategies enhance the identification of therapeutic targets and optimize drug discovery processes for diphtheria, and what specific challenges and advancements have been observed in recent studies?

Table 4: Summary of Findings on Computational Drug Discovery for Diphtheria

Author(s)	Year	Topic	Findings and Recommendations
Jamal et al.	2017	Genome Analysis of Corynebacterium diphtheria	Employed computational methods to analyze the complete genome, identifying potential therapeutic targets beyond proteins, which is essential for comprehensive drug development strategies.
Schneider	2018	Overview of Computational Drug Discovery	Highlighted improvements in duration, costs, and error rates in drug discovery processes through computational methods, emphasizing their significance in modern drug

			development.
Khalid et al.	2018	Target Identification in Corynebacterium diphtheriae	Focused on identifying proteins as potential drug targets and the use of sophisticated computational techniques to analyze protein structures, yielding insights for drug development.
Balasubramanian	2018	Lead Optimization Using Computational Methods	Discussed the use of computer models to predict pharmacokinetic and pharmacodynamic properties, aiding in enhancing drug efficacy and safety profiles during the optimization phase.
Sinha & Vohora	2018	Importance of ADMET Properties	Underlined the significance of pharmacokinetic properties (ADMET) in assessing the safety and efficacy of bioactive compounds, recommending early evaluation during drug optimization.
Neves et al.	2018	QSAR Techniques in Drug Discovery	Explored how Quantitative Structure-Activity Relationship (QSAR) models relate chemical structures to biological activity, emphasizing the importance of quality descriptors in model reliability.
Brogi	2019	Utilizing Computer Techniques for Candidate Discovery	Demonstrated rapid analysis of large compound libraries to discover promising candidates, advocating for computational methods over traditional chemical screening.
Mohamed et al.	2019	Knowledge Graphs in Drug Target Prediction	Highlighted the use of knowledge graphs to enhance drug target predictions, identifying intricate correlations that traditional models might overlook, while acknowledging challenges like class imbalance.
Ferreira & Andricopulo	2019	Financial Impact of ADMET Evaluations	Noted the historical financial losses associated with pursuing drugs with inadequate ADMET profiles, advocating for early assessments to optimize resource allocation in drug development.
Redkar et al.	2020	Challenges in Knowledge Graph Embeddings	Discussed issues such as class imbalance and high dimensionality in datasets, suggesting advanced techniques for data preprocessing and innovative training methodologies to improve model accuracy.
Qureshi et al.	2023	Effectiveness of In Silico Techniques	Emphasized the role of in silico techniques in simulating drug-target interactions, reducing the need for extensive in vitro or in vivo testing, and expediting the drug discovery process.

Table 4 highlights the contributions of various authors to the field of computational drug discovery for diphtheria, summarizing their key findings and recommendations. Jamal et al. (2017) initiated the discussion by using computational methods to analyze the genome of Corynebacterium diphtheriae, uncovering potential therapeutic targets. Schneider (2018) provided an overview of improvements in drug discovery processes, emphasizing reductions in duration, costs, and error rates due to computational techniques. Khalid et al. (2018) focused on specific protein identification as drug targets through advanced computational analysis. Balasubramanian (2018) discussed the role of computer models in predicting pharmacokinetic and pharmacodynamic properties during lead optimization, while Sinha and Vohora (2018) stressed the importance of early ADMET assessments for bioactive compounds. Neves et al. (2018) introduced Quantitative Structure-Activity Relationship (QSAR) techniques to link chemical structures with biological activity. Brogi (2019) advocated for computational methods in large compound library analysis, and Mohamed et al. (2019) highlighted knowledge graphs for improved drug target predictions. Redkar et al. (2020) addressed challenges in data processing to enhance model accuracy, and Qureshi et al. (2023) emphasized the effectiveness of in silico techniques in simulating drug-target interactions, expediting the drug discovery process. Overall, these findings underscore the transformative impact of computational strategies in identifying therapeutic targets and optimizing drug development for diphtheria.

Drug likeness assessment is commonly performed using the criteria established by Lipinski's Rule of Five, which provides essential standards for evaluating the potential of drug candidates for oral administration. According to this regulation, a drug should not violate more than one of the specified criteria: a maximum of five hydrogen bond donors, ten oxygen and nitrogen atoms, a molecular mass of less than 500 Da, and an octanol-water partition coefficient (LogP) of five or lower. These criteria are crucial in preparing ligand libraries for Computer-Aided Drug Design (CADD), highlighting the importance of evaluating drug-like properties early in the drug development process (X. Chen et al., 2020).

The advantages of computational drug discovery over traditional approaches are well documented. Computational drug discovery utilizes algorithms, data analysis, and in silico modeling to enhance the identification and optimization of new therapeutic drugs (Macarron et al., 2011). Unlike conventional methods, which often require extensive specialized computing expertise, computational techniques improve accessibility for researchers lacking computer backgrounds. Traditional drug discovery has historically relied on experimental approaches characterized by extensive trial-and-error, beginning with identifying biological targets, such as proteins, followed by screening hundreds to millions of compounds. This time-consuming and labor-intensive process has been criticized for its inefficiency (Sinha & Vohora, 2018). In contrast, computational drug discovery employs computer-based models to understand the complex interactions between drugs and their targets. This methodology encompasses molecular docking, quantitative structure-activity relationships (QSAR), and machine learning, which facilitate predictions of efficacy, safety, and metabolism.

Moreover, traditional approaches are often associated with prohibitive time and cost, typically requiring over a decade and incurring billions of dollars to bring a single drug to market. This high cost is partly due to the significant failure rates observed at various stages of drug development (Wouters et al., 2020). Computational approaches have been shown to substantially reduce both the time and expenses associated with drug discovery. By facilitating rapid screening and optimization of compounds, researchers can refine the candidate pool that necessitates synthesis and experimental testing, resulting in lower overall duration and financial investment in development (DiMasi et al., 2016).

While traditional methods are widely recognized, they face challenges due to their high-throughput characteristics, which can result in low success rates and often overlook the complexities of biological systems (Sinha & Vohora, 2018). In contrast, computational methods provide greater accuracy in forecasting drug-target interactions, especially when combined with machine learning techniques. However, the efficacy of these models is contingent upon the quality of the data on which they are trained, indicating that these approaches may not fully capture the intricacies inherent in biological systems (Schneider, 2018).

Additionally, conventional drug development typically aims to provide broadly effective treatments, with less focus on individual differences among patients. In this context, computational approaches are pivotal in advancing personalized medicine by integrating patient-specific data that correspond to distinct genetic profiles (H. Zhang et al., 2020). Furthermore, traditional drug research tends to exhibit incremental innovation due to its reliance on established frameworks and molecules. Conversely, computational tools play a critical role in fostering innovation by identifying new pharmacological targets and mechanisms, as well as in the process of drug repurposing, where existing drugs are rapidly screened for novel therapeutic applications.

Conventional drug discovery methods have long been integral to medical advancements, offering valuable insights and direct validation through hands-on experimentation. Despite the increasing use of computational techniques, traditional methods retain unique advantages that contribute to a thorough understanding of drug efficacy and biological complexity. The following are the advantages of Conventional drug discovery:

i. **Experimental Validation:** Conventional approaches offer a means of directly validating the effectiveness and safety of drugs through experimentation, a crucial aspect in obtaining regulatory approval. Experimental validation is generally necessary to validate CADD forecasts, despite their great value (del Carmen Quintal Bojórquez & Campos, 2023).

ii. **Understanding of Biological Complexity**: The comprehension of biological complexity can be enhanced by traditional research methods, as they provide a more profound understanding of biological pathways and the intricate nature of diseases. This is mostly due to the utilization of hands-on experiments and observations, which may not be adequately captured by computational models.

iii. **Discovery of Serendipitous Findings**: The utilization of traditional methods might result in surprising findings, wherein unforeseen outcomes from tests can create novel opportunities for research and medication development(Foletti & Fais, 2019). This advantage is less probable in the more focused CADD approach.

While computational approaches in drug discovery offer speed and precision, they also face certain limitations, particularly in workforce expertise and data quality. These challenges can impact the accuracy and reliability of predictive models, which are fundamental to Computer-Aided Drug Design (CADD).

i. **Scarcity of Skilled Professionals:** There is a notable shortage of professionals proficient in the advanced computational and machine learning techniques required for effective Computer-Aided Drug Design. This gap presents a challenge to fully leveraging CADD's capabilities. To address this, training programs and recruitment efforts focused on computational drug discovery skills are essential. These efforts aim to bridge the skills gap and integrate the latest tools, ultimately accelerating drug discovery, enhancing model accuracy, and producing more effective therapeutics (Akhtar et al., 2020).

ii. **Dependence on Data Quality and Quantity**: The effectiveness of CADD heavily depends on the quality and quantity of the data used to train predictive models. Insufficient or low-quality data can reduce the accuracy of these models, leading to unreliable predictions. Key steps like removing outliers, standardizing data formats, and implementing consistent experimental protocols—such as uniform assay settings and endpoint measurements—are critical for ensuring the quality of molecular interaction datasets and the robustness of computational models (Lee et al., 2022; Zhu, 2020).

iii. **Computational Limitations**: Despite advancements in computing power, there remain computational constraints in processing large-scale, complex biological data. High-performance computing resources are essential to analyze intricate interactions, but these resources are costly and require continuous updates to keep pace with advancements, potentially limiting access for smaller research facilities.

iv. **Limited Biological Context:** Computational models often struggle to capture the full complexity of biological systems. Many interactions and pathways are highly nuanced and context-dependent, making it difficult for CADD to fully replicate the physiological environment where drugs ultimately function. This limitation can result in predictions that may not accurately translate to in vivo results, underscoring the importance of complementary experimental validation.

v. **Risk of Overfitting:** When computational models are trained on limited or biased datasets, there is a risk of overfitting, where the model performs well on training data but poorly on new, unseen data. This can lead to misleadingly optimistic results and suggests a need for careful data management and regular model evaluation to ensure generalizability.

Collectively, these findings underscore the transformative impact of computational strategies in drug discovery, particularly in enhancing the identification of therapeutic targets, optimizing drug development, and addressing the inherent challenges faced in the field. The integration of computational methods not only supports innovation but also addresses key limitations of traditional drug discovery processes, thus representing a significant advancement in pharmaceutical research and development.

4.0. Conclusion, Recommendations, and Future Research Gaps to Be Filled

Computational methods, especially Computer-Aided Drug Design (CADD), have advanced drug discovery by improving prediction accuracy, optimizing candidate drugs, and lowering development costs. However, limitations such as the need for highly skilled personnel, dependence on quality data, computational constraints, and challenges in fully replicating biological complexity remain significant. Addressing these limitations is essential to maximize the potential of CADD in creating effective treatments, particularly for diseases like diphtheria. The following recommendations are proposed to advance computational drug discovery:

i. **Investment in Specialized Training Programs:** To mitigate the skills gap, institutions should establish programs that integrate computational techniques with drug discovery expertise. Collaborations between academia and industry can foster a workforce adept in both areas, enhancing the application of CADD methods.

ii. **Promoting Interdisciplinary Collaboration:** Bridging expertise across computational biology, pharmacology, and data science is crucial. By encouraging interdisciplinary research, institutions can improve the effectiveness of CADD models and create more comprehensive solutions for drug discovery.

iii. **Supporting Open-Source Data Repositories:** Building shared databases and open-access platforms can improve data quality, availability, and reliability. Policymakers and funding agencies should support and incentivize data sharing, enabling researchers to access robust datasets and facilitate collaborative advancements in computational drug design.

Future research should focus on refining algorithms to better mimic biological complexities, addressing computational resource demands, and developing adaptable CADD models suited for infectious disease research, especially in under-resourced settings.

Reference

1. Acosta, A. M., Moro, P. L., Hariri, S., Tiwari, T. S. P., Diphtheria, Hamborsky, J., Kroger, A., & Wolfe, S. (2021). Epidemiology and prevention of vaccine-preventable diseases. *The Pink Book: Diphtheria. Centre of Disease Control and Prevention: CDC Pink Book.*
2. Akhtar, S., Khan, M. K. A., & Osama, K. (2020). Machine learning approaches to rational drug design. *Computer-Aided Drug Design*, 279–306.

3. Balasubramanian, K. (2018). Mathematical and computational techniques for drug discovery: promises and developments. *Current Topics in Medicinal Chemistry*, *18*(32), 2774–2799.

4. Besa, N. C., Coldiron, M. E., Bakri, A., Raji, A., Nsuami, M. J., Rousseau, C., Hurtado, N., & Porten, K. (2014). Diphtheria outbreak with high mortality in northeastern Nigeria. *Epidemiology & Infection*, *142*(4), 797–802.

5. Brogi, S. (2019). Computational approaches for drug discovery. In *Molecules* (Vol. 24, Issue 17, p. 3061). MDPI.

6. Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., & Barr, A. (2017). Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications*, *72*, 151–159.

7. Chen, R., Liu, X., Jin, S., Lin, J., & Liu, J. (2018). Machine learning for drug-target interaction prediction. *Molecules*, *23*(9), 2208.

8. Chen, X., Li, H., Tian, L., Li, Q., Luo, J., & Zhang, Y. (2020). Analysis of the physicochemical properties of acaricides based on Lipinski's rule of five. *Journal of Computational Biology*, *27*(9), 1397–1406.

9. Clarke, K. E. N. (2017). Review of the epidemiology of diphtheria 2000-2016. *World Health Organization*, 2005–2021.

10. del Carmen Quintal Bojórquez, N., & Campos, M. R. S. (2023). Traditional and Novel Computer-Aided Drug Design (CADD) Approaches in the Anticancer Drug Discovery Process. *Current Cancer Drug Targets*, *23*(5), 333–345.

11. DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33.

12. Ferreira, L. L. G., & Andricopulo, A. D. (2019). ADMET modeling approaches in drug discovery. *Drug Discovery Today*, *24*(5), 1157–1165.

13. Foletti, A., & Fais, S. (2019). Unexpected discoveries should be reconsidered in science—a look to the past? In *International Journal of Molecular Sciences* (Vol. 20, Issue 16, p. 3973). MDPI.

14. Heikamp, K., & Bajorath, J. (2014). Support vector machines for drug discovery. *Expert Opinion on Drug Discovery*, *9*(1), 93–104.

15. Henriques, R., & Madeira, S. C. (2015). Towards robust performance guarantees for models learned from high-dimensional data. *Big Data in Complex Systems: Challenges and Opportunities*, 71–104.

16. Imaizumi, T., Nakayama, A., & Yokoyama, S. (2020). *Advanced Studies in Behaviormetrics and Data Science*. Springer.

17. Jamal, S. B., Hassan, S. S., Tiwari, S., Viana, M. V, Benevides, L. de J., Ullah, A., Turjanski, A. G., Barh, D., Ghosh, P., & Costa, D. A. (2017). An integrative in-silico approach for therapeutic target identification in the human pathogen Corynebacterium diphtheriae. *PLoS One*, *12*(10), e0186401.

18. Kadir, M. E., Akash, P. S., Sharmin, S., Ali, A. A., & Shoyaib, M. (2020). A proximity weighted evidential k nearest neighbor classifier for imbalanced data. *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*, 71–83.

19. Kaushik, A. C., Kumar, A., Bharadwaj, S., Chaudhary, R., Sahi, S., Kaushik, A. C., Kumar, A., Bharadwaj, S., Chaudhary, R., & Sahi, S. (2018). Ligand-based approach for in-silico drug designing. *Bioinformatics Techniques for Drug Discovery: Applications for Complex Diseases*, 11–19.

20. Khalid, Z., Ahmad, S., Raza, S., & Azam, S. S. (2018). Subtractive proteomics revealed plausible drug candidates in the proteome of multi-drug resistant Corynebacterium diphtheriae. *Meta Gene*, *17*, 34–42.

21. Kim, H.-C., Park, J.-H., Kim, D.-W., & Lee, J. (2020). Multilabel naïve Bayes classification considering label dependence. *Pattern Recognition Letters*, *136*, 279–285. https://doi.org/https://doi.org/10.1016/j.patrec.2020.06.021

22. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., & Yu, B. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, *47*(D1), D1102–D1109.

23. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., & Shoemaker, B. A. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213.

24. Lee, J. W., Maria-Solano, M. A., Vu, T. N. L., Yoon, S., & Choi, S. (2022). Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD). *Biochemical Society Transactions*, *50*(1), 241–252.

25. Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., Green, D. V. S., Hertzberg, R. P., Janzen, W. P., & Paslay, J. W. (2011). Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, *10*(3), 188–195.

26. Malhat, M. G., Mousa, H. M., & El-Sisi, A. B. (2014). Clustering of chemical data sets for drug discovery. *2014 9th International Conference on Informatics and Systems*, DEKM-11.

27. Maltarollo, V. G., Kronenberger, T., Espinoza, G. Z., Oliveira, P. R., & Honorio, K. M. (2019). Advances with support vector machines for novel drug discovery. *Expert Opinion on Drug Discovery*, *14*(1), 23–33.

28. Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., & Nowotka, M. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940.

29. Mohamed, S. K., Nováek, V., & Nounu, A. (n.d.). *Discovering protein drug targets using knowledge graph*.

30. Monakhova, Y. B., & Rutledge, D. N. (2020). Independent components analysis (ICA) at the "cocktail-party" in analytical chemistry. *Talanta*, *208*, 120451.

31. NCDC. (2023). Diphtheria Health Advisory for Health Care Workers amidst Outbreak in Nigeria. https://ncdc.gov.ng/news/436/diphtheria-health-advisory-for-health-care-workers-amidst-outbreak-in-nigeria

32. Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., & Andrade, C. H. (2018). QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in Pharmacology*, *9*, 1275.

33. Organization, W. H. (2018). Diphtheria vaccine: WHO position paper, August 2017–Recommendations. *Vaccine*, *36*(2), 199–201.

34. Padole, S. S., Asnani, A. J., Chaple, D. R., & Katre, S. G. (2022). A review of approaches in computer-aided drug design in drug discovery. *GSC Biological and Pharmaceutical Sciences*, *19*(2), 75–83.

35. Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. *Molecules*, *25*(22), 5277.

36. Qureshi, R., Irfan, M., Gondal, T. M., Khan, S., Wu, J., Hadi, M. U., Heymach, J., Le, X., Yan, H., & Alam, T. (2023). AI in drug discovery and its clinical relevance. *Heliyon*.

37. Reddy, C. K., & Vinzamuri, B. (2018). A survey of partitional and hierarchical clustering algorithms. In *Data clustering* (pp. 87–110). Chapman and Hall/CRC.

38. Redkar, S., Mondal, S., Joseph, A., & Hareesha, K. S. (2020). A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing. *Molecular Informatics*, *39*(5), 1900062.

39. Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, *20*(5), 1878–1912.

40. Rodríguez-Pérez, R., Vogt, M., & Bajorath, J. (2017). Support vector machine classification and regression prioritize different structural features for binary compound activity and potency value prediction. *ACS Omega*, *2*(10), 6371–6379.

41. Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, *17*(2), 97–113.

42. Sharma, N. C., Efstratiou, A., Mokrousov, I., Mutreja, A., Das, B., & Ramamurthy, T. (2019). Diphtheria. *Nature Reviews Disease Primers*, *5*(1), 1–18.

43. Sinha, S., & Vohora, D. (2018). *Chapter 2 - Drug Discovery and Development: An Overview* (D. Vohora & G. B. T.-P. M. and T. C. R. Singh (eds.); pp. 19–32). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-802103-3.00002-X

44. Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews*, *66*(1), 334–395.

45. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *237*, 121549.

46. Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., & Zhang, J. (2017). Pubchem bioassay: 2017 update. *Nucleic Acids Research*, *45*(D1), D955–D963.

47. Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, *323*(9), 844–853.

48. Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165–193.

49. Zakikhany, K., & Efstratiou, A. (2012). Diphtheria in Europe: current problems and new challenges. *Future Microbiology*, *7*(5), 595–607.

50. Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician*, *71*(4), 310–316.

51. Zhang, H., Wang, S., Zhang, K., Tang, Z., Jiang, Y., Xiao, Y., Yan, W., & Yang, W.-Y. (2020). Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2407–2416.

52. Zhu, H. (2020). Big data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology*, *60*, 573–589.