

Article

Graph Neural Network and BERT Model for Antimalarial Drug Predictions Using Plasmodium Potential Targets

Medard Edmund Mswahili , Goodwill Erasmo Ndomba, Kyuri Jo  and Young-Seob Jeong * 

Department of Computer Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; medardedmund25@chungbuk.ac.kr (M.E.M.); goodwillndomba22@chungbuk.ac.kr (G.E.N.); kyurijo@chungbuk.ac.kr (K.J.)

* Correspondence: ysjay@chungbuk.ac.kr



Citation: Mswahili, M.E.; Ndomba, G.E.; Jo, K.; Jeong, Y.-S. Graph Neural Network and BERT Model for Antimalarial Drug Predictions Using Plasmodium Potential Targets. *Appl. Sci.* **2024**, *14*, 1472. <https://doi.org/10.3390/app14041472>

Academic Editor: Luis Javier Garcia Villalba

Received: 19 January 2024

Revised: 5 February 2024

Accepted: 7 February 2024

Published: 11 February 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Malaria continues to pose a significant global health burden despite concerted efforts to combat it. In 2020, nearly half of the world's population faced the risk of malaria, underscoring the urgency of innovative strategies to tackle this pervasive threat. One of the major challenges lies in the emergence of the resistance of parasites to existing antimalarial drugs. This challenge necessitates the discovery of new, effective treatments capable of combating the Plasmodium parasite at various stages of its life cycle. Advanced computational approaches have been utilized to accelerate drug development, playing a crucial role in every stage of the drug discovery and development process. We have witnessed impressive and groundbreaking achievements, with GNNs applied to graph data and BERT from transformers across diverse NLP text analysis tasks. In this study, to facilitate a more efficient and effective approach, we proposed the integration of an NLP based model for SMILES (i.e., BERT) and a GNN model (i.e., RGCN) to predict the effect of antimalarial drugs against Plasmodium. The GNN model was trained using designed antimalarial drug and potential target (i.e., PfAcAS, F/GGPPS, and PfMAGL) graph-structured data with nodes representing antimalarial drugs and potential targets, and edges representing relationships between them. The performance of BERT-RGCN was further compared with that of Mordred-RGCN to evaluate its effectiveness. The BERT-RGCN and Mordred-RGCN models performed consistently well across different feature combinations, showcasing high accuracy, sensitivity, specificity, MCC, AUROC, and AUPRC values. These results suggest the effectiveness of the models in predicting antimalarial drugs against Plasmodium falciparum in various scenarios based on different sets of features of drugs and potential antimalarial targets.

Keywords: malaria; graph neural network; BERT; tokenizer; Plasmodium falciparum; machine learning; deep learning; natural language processing; drug discovery and development

1. Introduction

Malaria is a life-threatening disease transmitted through the bites of infected female Anopheles mosquitoes. There are five parasite species that cause malaria in humans, and two of these species, namely Plasmodium falciparum and Plasmodium vivax, pose the greatest threat. Notwithstanding the fact that the substantial efforts and advancements across the globe to combat malaria have improved significantly during the last decade, in 2022, nearly half of the world's population was at risk of malaria. According to the WHO, there were an estimated 249 million infections of malaria in 2022, and the estimated number of malaria deaths stood at 608,000 [1].

Malaria is one of the most devastating and widespread tropical parasitic diseases of those most prevalent in developing countries. The WHO regions of South-east Asia, the Eastern Mediterranean, the Western Pacific, and the Americas reported a significant number of cases and deaths. Among them, the WHO African region carries a disproportionately high share of the global burden of malaria. In 2022, the region was home to 95% and 96%

cases of malaria and deaths, respectively [1–3]. The incidence of malaria cases, as observed in the WHO report, has increased consistently for 22 years from 2000 to 2022. One of the reasons seems to be parasite resistance to existing antimalarial drugs.

Resistance has emerged to most approved antimalarial drugs. This spread and establishment can undermine the gains seen over the last decade. Therefore, the need for new drugs that work through differentiated modes of action is urgent [4]. Antimalarial drug resistance is the ability of a parasite strain to survive and/or to multiply despite the administration and absorption of medicine given in doses equal to or higher than those usually recommended [5]. Resistance is a key challenge for anti-infectives in general, but particularly for antimalarials. There are a number of different mechanisms by which resistance can emerge in pathogens. These include the following: mutations in the target enzyme, amplification of the target enzyme, etc. [6]. However, several factors have contributed to the emergence of resistance to current antimalarial drugs and medications. These include the rate of parasite mutation, the potency of the selected drug, inadequate pharmacokinetic properties, substandard quality of antimalarial drugs, among other elements that can contribute and exacerbate the resistance issue [5].

With the rising resistance to frontline drugs (artemisinin-based combinations), there is a need to accelerate the discovery and development of novel antimalarial drugs [7]. As drugs are urgently needed, the concerted effort of many in the field has resulted in the screening of millions of small molecules in phenotypic assays to assess parasite death [8–10]. These efforts have identified thousands of new chemical scaffolds. A major goal of new drug development for malaria is the discovery of compounds that kill parasites in multiple stages of the life cycle and, thus, could be used both in disease prevention and treatment. In the realm of drug discovery, chemical biology and computational drug design methodologies are harnessed for the effective identification and refinement of lead compounds. Chemical biology focuses primarily on unraveling and elucidating the biological function of a target and understanding the mechanism of action of a chemical modulator. Conversely, the *in silico*-based method or computer-aided drug design leverages structural insights from the target (structure-based) or known bioactive ligands (ligand-based) to facilitate and accelerate the identification of promising candidate drugs [11].

There has been a shift in the discovery of antimalarial drugs from phenotypic screening to target-based approaches, as more potential drug targets have been validated in *Plasmodium* species [4]. Given the high attrition rate, high demand for new drugs, and enormous cost and time-consuming nature of drug discovery [12–16], it is essential to select the targets that are the most likely to deliver progressable drug candidates. Target-based drug discovery is the dominant paradigm of drug discovery [17]. In this study, we utilize the following three potential and high priority targets found in *Plasmodium* that are responsible and ready to enter the antimalarial drug discovery process.

- Acetyl CoA Synthetase (PfAcAS; high priority target [4]) is responsible for the biosynthesis of acetyl coenzyme A from coenzyme A and acetate. Researchers have recently reported on the validation of this enzyme [18,19].
- Bifunctional Farnesyl/Geranylgeranyl Pyrophosphate Synthase (F/GGPPS; high priority target [4]) is a key enzyme in isoprenoid biosynthesis that synthesizes C15 and C20 prenyl chains. Prenyl chains are the substrates of several prenyltransferases that result in isoprenoid products essential for the survival of the parasite [20].
- Monoacylglycerol Lipase (PfMAGL; new and emerging [4]) [21]. Human monoacylglycerol lipase (MAGL) catalyzes the hydrolysis of a variety of monoglycerides into fatty acids and glycerol. In *Plasmodium falciparum*, this enzyme has been reported to play a role in the processing of these monoglycerides, including palmitoyl and oleoyl glycerols [21].

The criteria considered important for selecting the aforementioned targets for antimalarial drug discovery are discussed by Forte et al. [4]. They also describe the analysis of several drug targets within the Malaria Drug Accelerator (MalDA) pipeline, allowing them to prioritize targets that are ready to enter the drug discovery phase.

In recent years, machine learning (ML) techniques such as graph neural networks (GNN) and natural language processing (NLP), exemplified by methods such as ChemBERTa, SMILES-BERT, etc. [22–24], have shown remarkable performance in various domains such as cheminformatics and bioinformatics [13,25,26]. Researchers have actively proposed data-driven approaches and have adopted computational biology perspectives based on deep learning to explore drug–target interaction or relations [14–16,27–33]. Accurate prediction and identification of drug–target interactions (DTIs) are pivotal elements in drug discovery [14–16,27,32,33]. The emphasis on computational methods for the prediction of DTIs has grown due to the significant costs and time associated with extensive *in vivo* and *in vitro* experiments with a wide range of potential drug chemical compounds [13–16]. On the other hand, advanced DL frameworks, utilizing a variety of GNNs such as graph convolutional networks (GCNs) [34], graph attention networks (GATs) [35], gated graph neural networks (GGNNs) [36], and residual graph convolutional networks (RGCNs) [37], have achieved groundbreaking performance in various research domains, including social and natural sciences, and knowledge graphs [38,39].

These frameworks have shown their effectiveness in various applications, such as molecular properties prediction [40], widespread application in predicting DTIs [12], and the prediction of protein function [41] within the realm of biochemical problems where the interactions can be represented as graph-like structured data. Specifically, GCNs have found application in addressing pharmacological similarities by considering both sequential and structural properties [12]. The graph representations of biochemical entities have proven capable of capturing structural features akin to Euclidean ones, eliminating the need for extensive feature engineering [42,43]. GNNs have been demonstrated to be relatively powerful in modeling their formally input graph-structured data to the DL pipeline [44–46]. The graph-structured data are basically presented as $G = (V, E, u)$, where G is a graph, V indicates vertices or nodes, E means edges comprising adjacency (A) with weights (W), and u indicates feature vectors of the nodes. All GNN layers are implemented via a message passing network (MPN) interface or framework [47]. Quantitative Structure–Activity Relationship (QSAR) modeling investigates the relationship between the chemical structure of molecules or substances and their biological activities towards a specific target (i.e., a drug target) [48]. By employing statistical techniques, QSAR and ML models have the capacity to predict drugs for drug repurposing or repositioning and pharmacological activity of novel compounds based on their structural attributes [49]. This capability enables chemists to strategically modify molecules, improving the potency of a drug or alleviating side effects, resulting in a more cost-effective and time-efficient drug development process [50].

Herein, this study introduces a pioneering approach that harnesses the potential of advanced ML methodologies to predict and identify antimalarial drugs. The essence of this approach is the integration of state-of-the-art NLP and GNN techniques. The study proposes a fusion of the SMILES-based bidirectional encoder representations from transformers (BERT) model, an NLP model, with the robust capabilities of the GNN model (i.e., relational graph convolutional network (RGCN)). We adopted the deepchem SmilesTokenizer module for embedding features generation [51] that were later treated by RGCN as node features. This tokenizer heavily inherits from the BertTokenizer class found in Huggingface’s transformers library. It runs a WordPiece tokenization algorithm over SMILES strings using the tokenization SMILES regex developed by Schwaller et al. [52]. Furthermore, the RGCN model was trained using multi-relational antimalarial drugs and Plasmodium potential targets graph data, where nodes represent antimalarial drugs and potential targets, and edges represent drug–target relations. The contributions of this study can be summarized as follows:

- Apparently, this is the first study to introduce the NLP-based applications to chemistry and biology (i.e., BERT-based SmilesTokenizer) and the GNN model (i.e., RGCN) for the antimalarial drug prediction against Plasmodium falciparum.
- Secondly, the graph-based network was designed between antimalarial drugs and three potential new targets for antimalarial drugs found in Plasmodium as shown in Figure 1.

- We further developed three independent models, BERT-RGCN, Mordred-RGCN, and BERT-Mordred-RGCN, and compared their performance to understand the contributions of the features.
- Lastly, various experiments were performed using features of individual antimalarial drugs, followed by combination of features of antimalarial drugs and three potential targets for malaria drugs found in Plasmodium.

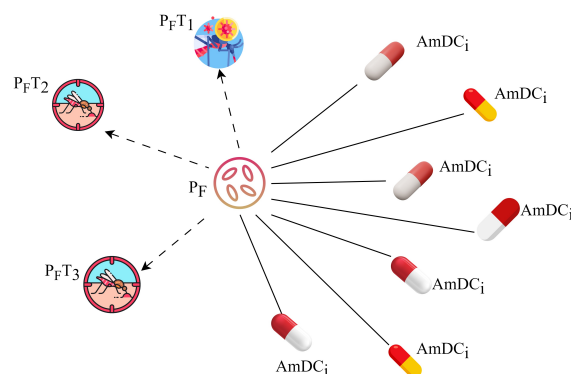


Figure 1. Antimalarial drugs and Plasmodium potential targets relation network. P_F indicates Plasmodium falciparum, P_{FT_1} indicates Acetyl coenzyme A target, P_{FT_2} indicates Geranylgeranyl diphosphate target, P_{FT_3} indicates Monoacylglycerol Lipase Inhibitor 21 target, $P_{FT_{k-th}}$ indicates Plasmodium falciparum targets where $k \in \{1, 2, 3\}$, $AmDC_i$ indicates antimalarial drug compound where $i \in \{1, 2, 3, \dots\}$.

Literature Review

In modern drug discovery, the integration of cheminformatics and QSAR modeling has emerged as a formidable alliance, allowing researchers to harness and leverage the extensive potential of ML techniques for predictive molecular design and analysis [50]. Previous studies have conducted thorough validation of predictive models for antimalarial drugs using ML methods and molecular descriptors. Mswahili et al. [2] proposed various ML models, with the best models achieving accuracies exceeding 82%. In the same approach, Liu et al. [53] proposed traditional ML classification models to predict antimalarial activity against Plasmodium falciparum and achieved accuracies of 87.3% and 88.9% for a support vector machine (SVM) and general regression neural network (GRNN), respectively. Additionally, Danishuddin et al. [54] conducted a comprehensive validation of antimalarial predictive models using ML approaches. Among these, SVM and XGBoost exhibited high performance, achieving an accuracy of approximately 85% on the independent test set. While the results are encouraging and promising, the majority of the in silico-based methods proposed still focus primarily on the features of antimalarial drugs. The common limitation, clearly, is that these previous methods did not encompass the distinctive features of potential targets found in Plasmodium, which are paramount in the discovery of novel antimalarial drugs.

2. Materials and Methods

The cornerstone of this methodology lies in the use of multi-relational data that include pivotal antimalarial drug targets, particularly high priority targets (i.e., PfAcAS and F/GGPPS) and new and emerging target (i.e., PfMAGL). These data form the foundation for training the GNN, where the nodes represent antiplasmodial drugs and potential targets, while the edges delineate the intricate relationships and interactions between them. Furthermore, the GNN is trained on structured data that capture the complex interaction between antimalarial drugs and their potential targets. This comprehensive approach aims to, in the future, identify promising drug candidates capable of disrupting the malaria parasite's life cycle through diverse mechanisms of action.

This paper delves into the methodology, data representation, and potential implications of this interdisciplinary approach in expediting the discovery of novel antimalarial drugs. The amalgamation of NLP model with GNNs is a key component in this research that aims to offer a transformative path to combating malaria, addressing the pressing challenge of drug resistance, and advancing the frontier of drug discovery research.

2.1. Dataset

The dataset used in this study is publicly available and was acquired from the previous study of Mswahili et al. [2] specializing in antimalarial drug predictions and developments. The dataset is a $|D| \times 5$ matrix, where $|D|$ is the number of total instances. The dataset comprises a total of 4794 instances after the removal of the duplicated instances, where it consists of 2070 and 2724 instances for active and inactive classes, respectively. Active instances are experimentally verified as active antimalarial drug candidates, whereas inactive instances are experimentally verified as unsuccessful candidates. The classification of active and inactive was carried out according to the antiplasmodial activities of the drug compounds, as discussed by Mswahili et al. [2].

The foundation of this study relied on a comprehensive dataset that includes multi-relational information crucial in the discovery of antimalarial drugs and medications. The dataset incorporates various pieces of information on antimalarial drugs, potential drug targets, and their intricate relationships, as shown in Figure 1. The dataset includes information on PfAcAS and F/GGPPS (as high priority), and PfMAGL (new and emerging) targets in the fight against malaria, as shown in Figure 1. Therefore, each antimalarial drug compound was associated with each of the features of three individual targets (i.e., PfAcAS, F/GGPPS, and PfMAGL) as shown in Table 1.

Table 1. Drugs and targets data overview.

Descriptions		# of Instances
Compounds	Antimalarial drugs	4794
Malaria potential targets	PfAcAS (high priority)	4794
	F/GGPPS (high priority)	4794
	PfMAGL (new and emerging)	4794
Label	Active (1)	2070
	Inactive (0)	2724

PfAcAS_F—Acetyl CoA Synthetase features. *GGPP_F*—Geranylgeranyl Pyrophosphate Synthase features. *PfMAGL_F*—Monoacylglycerol Lipase features. #—number.

2.2. Antimalarial Drugs' and Targets' Features' Extraction

All antimalarial drug compounds and the mentioned potential targets were encoded as SMILES strings for the subsequent extraction of features reminiscent of topological structure and semantic information drugs and targets, illustrated in Figure 2.

Using the DeepChem adopted method (i.e., Bert SmilesTokenizer module) [51], the embedding features of both antimalarial drug compounds and potential targets were initially calculated, as described in Table 2. It is important to note that the length of the SMILES strings representing antimalarial drug compounds varied, leading to diverse F_{ALL} dimensional real-numbered feature vectors for each drug compound. The maximum length observed in the values of the real-numbered dimensional feature vector F_{ALL} was 165. To address this variability, a padding function was applied, setting it to the maximum length of 165. In simpler terms, this involved padding the features of SMILES strings with a shorter length using zeros. Consequently, this process (i.e., padding) standardized and resulted in a similar F_{ALL} dimensional real-numbered feature vector for all drug compounds.

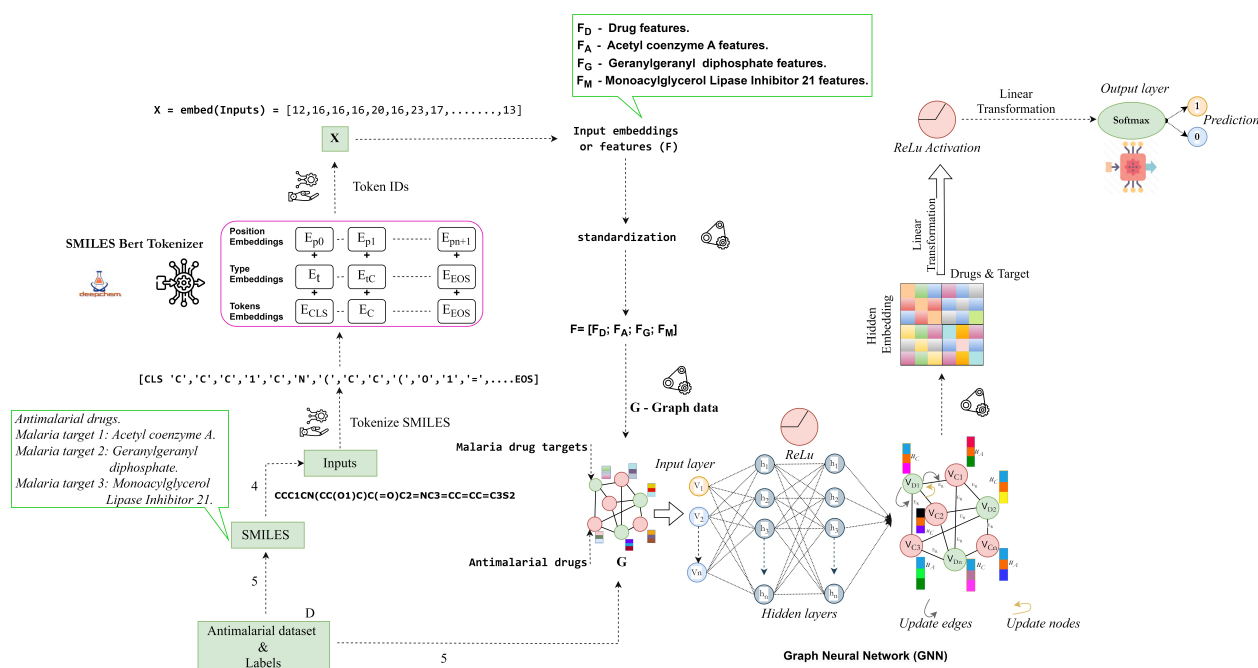


Figure 2. Overall GNNs and SMILES—based BERT prediction models' architecture and implementation.

Table 2. Drugs' and targets' features' analysis.

Features	SMILES Strings	# of F_{ALL} Values
Bert embeddings	Drug compounds	165
	PfAcAS (high priority)	101
	F/GGPPS (high priority)	53
	PfMAGL (new and emerging)	58
Mordred descriptors	Drug compounds	1613
	PfAcAS (high priority)	1613
	F/GGPPS (high priority)	1613
	PfMAGL (new and emerging)	1613

F_{ALL} —All features. $PfAcAS_F$ —Acetyl CoA Synthetase features. $GGPP_F$ —Geranylgeranyl Pyrophosphate Synthase features. $PfMAGL_F$ —Monoacylglycerol Lipase features. #—number.

Molecular descriptors, such as 2D fingerprints and topological indices, play a pivotal role in conjunction with structure–activity relationships (SARs), serving as key elements in paving and unlocking the way for small-molecule drug discovery. Therefore, a well-known descriptor calculator, namely Mordred [55], was implemented to calculate the molecular descriptors values as the model features for each antimalarial drug compound and three potential targets using their respective SMILES strings. The calculated feature values were 2D Mordred descriptor values that were generated using the molecular featurizer provided by the DeepChem library [56]. The total number of feature values calculated by Mordred was $|D| \times 1613$ for each antimalarial drug compound and each of the three potential targets as described in Table 2. In particular, $|D|$ is the same number of instances as discussed in Section 2.1.

Data preprocessing was critical to ensuring uniformity, quality, and compatibility between the diverse datasets. Features in this study were further standardized before being fed to the models using the standard scaler [57]. The standardization of data formats and the preprocessing of inconsistencies facilitated seamless integration into a cohesive framework for subsequent model training and performance analysis and evaluation.

2.3. Topology Graph Construction

Drug–target links information has been widely used to study various drug-related problems, such as drug and target interactions' prediction [12,13,27,33,58,59], drug anatom-

ical therapeutic chemical (ATC) classifiers of drugs [60,61], and adverse reactions' prediction [62,63]. Here, the link information between the drugs was identified according to the combination of antimalarial drugs and the new targets introduced in this study. The future and current trend in DTI is further discussed by Abbasi et al. [15].

The drug–target links graph was constructed based on the dataset with the drug and three potential target links as shown in Table 3. The nodes on the graph indicate antimalarial drugs and the edges indicate interaction relationships between the drugs and three targets found in Plasmodium as described in Figure 1. The node features' values in the graph are the features discussed in Section 2.2 and are treated as feature vectors. The interaction graph consists of 4794 nodes and 9588 edges. The interaction relationship samples are shown in Table 3 and depicted in Figure 1. The built drug–target interaction graph-structured data were also fed into the models for training and evaluation tasks.

Table 3. Data summary used for graph construction.

Drugs' ID	Drugs' SMILES	Targets' Strings
CHEMBL233867	<chem>C1CS(=O)(=O)C(=O)N1CC2=CC=CC=C2</chem>	¹ Acetyl CoA Synthetase ² Geranylgeranyl Pyrophosphate Synthase ³ Monoacylglycerol Lipase
CHEMBL295594	<chem>CC(C)C1=NC=C(N1)CCN</chem>	Acetyl CoA Synthetase Geranylgeranyl Pyrophosphate Synthase Monoacylglycerol Lipase
CHEMBL1224171	<chem>CCCCCSC(=O)CCCCC</chem>	Acetyl CoA Synthetase Geranylgeranyl Pyrophosphate Synthase Monoacylglycerol Lipase
CHEMBL1214554	<chem>C1=CC(=C(C=C1C(=NO)N)O)O</chem>	Acetyl CoA Synthetase Geranylgeranyl Pyrophosphate Synthase Monoacylglycerol Lipase
CHEMBL236690	<chem>CC(=C)[CH]1CCCC(=CC1)COCC#C</chem>	Acetyl CoA Synthetase Geranylgeranyl Pyrophosphate Synthase Monoacylglycerol Lipase

¹ *PfAcAS_F*—Acetyl CoA Synthetase. CC(=O)SCCNC(=O)CCNC(=O)C(C(C)(C)COP(=O)(O)OP(=O)(O)OCC1C(C(C(O1)N2C=NC3=C(N=CN=C32)N)O)OP(=O)(O)O)O. ² *GGPP_F*—Geranylgeranyl Pyrophosphate Synthase. CC(=CCCC(=CCCC(=CCCC(=CCOP(=O)(O)OP(=O)(O)O)C)C)C. ³ *PfMAGL_F*—Monoacylglycerol Lipase. C1OC2=C(O1)C=C(C=C2)COC(=O)CCCCC3=CC=C(C=C3)C4=CC=CC=C4.

2.4. R-GCN Model

GCNs have gained substantial attention recently [64], emerging as the preferred methodology for learning graph representations to enhance virtual screening for DTI in the pharmaceutical industry [65]. The RGCN is essentially an extension of the GCN architecture, specifically developed to handle the highly multi-relational data found in genuine knowledge bases [66,67]. In the GCN model, the feature vectors of neighboring nodes undergo a transformation using a shared weight matrix. Conversely, the RGCN introduces the notion of relation types, where the vector transformation is specific to each relation, which means that each type of relation is associated with a unique weight matrix. Consequently, the RGCN model demonstrates the ability to handle relational and heterogeneous graphs effectively [68].

Herein this study, we introduced antimalarial drugs and three potential Plasmodium antimalarial targets that graphically represent structured network data as shown in Figure 1. These graph data, as discussed in Section 2.3, were used as input to the proposed model that was built based on the RGCN. The RGCN input data are as follows: $G = (V, \varepsilon, R)$, whereas G is defined as the graph data, $v_i \in V$ as nodes (entities), and $v_i, r, v_j \in \varepsilon$ as labeled edges (relations), where $r \in R$ is defined as a relation type [66]. In this context, it is defined as different types of drug–target relationships (E_R) as shown in Figure 1. The RGCN model was implemented using PyTorch Geometric [69] with hyperparameters as shown in Table 4. The RGCN learns the local and global features of antimalarial drugs and targets, and the

combined feature module (i.e., BERT- and Mordred-based features) as shown in Table 5 to predict the antimalarial drugs and potential targets' relations. Drug and target nodes' features were treated as integral components of model parameters, subject to updates during the model training process.

Table 4. Hyperparameters and training strategy in this study.

Model	Hyperparameters	Descriptions
R-GCN	# of layers	One RGCNConv
	Training technique	10-fold cross-validation
	Learning rate	0.00195
	Optimizer	Adam
	Dropout	0.645
	Forward module A_F	ReLU
	Predictive module A_F	Softmax

Conv—convolution. A_F —activation function. #—number.

Table 5. Experimental performance comparison and effect of different # of layers on RGCN.

RGCN ^a Input Features	# of Features	Evaluation	Single RGCNconv. Layer	Double RGCNconv. Layer
BERT and Mordred combined	6829	Accuracy	0.9958	0.9931
		Sensitivity	0.9968	0.9904
		Specificity	0.9951	0.9951
		MCC ^b	0.9915	0.9858
		AUROC ^c	0.9947	0.9965
		AUPRC ^d	0.9958	0.9950

[^a] Relational graph convolution network. [^b] Matthew's correlation coefficient. [^c] Area under the receiver operating characteristics. [^d] Area under the precision–recall curve. The RGCN's high scores, as denoted by bold values, are indicative of the layer's architecture. #—number.

2.5. BERT Model for SMILES

Recently in the field of NLP, advanced text analysis has become increasingly important both for academia and industry. This is in fact enabled due to the emergence of powerful language models (LMs) such as transformer-based architecture, namely BERT introduced by Devlin et al. [70], and generative pretraining (GPT) introduced by Radford et al. [71].

Ever since its rise and discovery by Google researchers in 2018, BERT has been considered a relatively powerful and widely used LM for various NLP tasks. The performance of BERT towards a specific task involves two essential mechanisms, which are pretraining and fine-tuning [70]. During pretraining, the model is trained on unlabeled data over different pretraining algorithms. In fine-tuning, the BERT model is first initialized with the pretrained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks.

BERT improves upon standard transformers by removing the unidirectionality constraint by using the masked language modeling (MLM) pretraining objective. In our study, the MLM randomly masks certain tokens within the input sequence, specifically the SMILES strings representing antimalarial drugs and potential targets, and the objective is to predict the original id of the vocabulary of the masked token only based on its context from the SMILES strings. Unlike left-to-right LM pretraining, the MLM objective enables the representation to fuse the left and the right contexts, which allows NLP researchers to pretrain a deep bidirectional transformer. In addition to the MLM, BERT uses a next sentence prediction (NSP) task that jointly pretrains text-pair representations [70]. However, in the pharmaceutical research field, NSP is ignored since, for example, our study involves

only single sequences of antimalarial drugs and targets encoded as SMILES strings during model pretraining.

In this study, we adopted the SmilesTokenizer module from DeepChem [51] inherited from the BertTokenizer class in transformers to generate embedding features as shown in Figure 2 that were later used by our proposed GNN model (i.e., RGCN) as node features after standardize and preprocessing. This runs a WordPiece tokenization algorithm over antimalarial drugs and targets SMILES strings using the tokenization SMILES regex developed by Schwaller et al. [52].

3. Experiment and Evaluation Criteria

The proposed framework integrates two different advanced ML methodologies: NLP (i.e., SMILES-based BERT model) and GNN (i.e., RGCN), as shown in Figure 2. The SMILES-based BERT model, renowned for its proficiency in processing molecular structures, was used to encode molecular information from SMILES strings (i.e., input embeddings or feature values) for antimalarial drugs and three targets (i.e., PfAcAS, F/GGPPS, and PfMAGL). This encoding of SMILES strings facilitated the representation of drugs and three potential targets in a manner conducive to subsequent graph-based analysis as shown in Figure 2. BERT was first introduced in this study specifically for the prediction of antimalarial drugs as a tool for extracting and generating features (i.e., Bert-based features) from drugs and targets SMILES strings as shown in Figure 2. To further validate the contribution and effectiveness of Bert-based features, we compared them with features generated by Mordred (i.e., Mordred-based features) [55], the most widely used tool for feature extraction in this area of antimalarial drugs prediction [2,54].

Thereafter, the GNN model, namely RGCN, was deployed to handle graph-structured data, encapsulating the relationships between antimalarial drugs and potential targets found in Plasmodium, as shown in Figure 1. Input nodes' features used during implementing and training RGCN were of course ones computed by using the DeepChem adopted Bert SmilesTokenizer module and further compared to the ones calculated by a well-known descriptor calculator known as Mordred. Training in RGCN involved iterative processes to discern and capture intricate drug–target interactions embedded within the multi-relational data. An overview of the architecture and training strategy of the proposed RGCN model is summarized in Table 4.

To evaluate the predictive capabilities of the integrated models, rigorous evaluation methodologies were used. The proportion of the dataset used in this study for model training and evaluation is summarized in Table 6.

Table 6. Data separation summary.

	Training	Validation	Test
# of instances	3355	719	720

#—number.

The datasets were partitioned into training, validation, and test sets to facilitate model training and subsequent assessment. Cross-validation techniques, such as 10-fold cross-validation, ensured robustness and mitigated biases in the model evaluation process. Evaluation metrics, including accuracy, sensitivity, specificity, Matthew's correlation coefficient (MCC), area under receiver operating characteristics (AUROC), and area under the precision–recall curve (AUPRC), were used to quantify the predictive performance of the model. Comparative analyses against baseline models and existing approaches to antimalarial drug discovery served as benchmarks to gauge the efficacy and superiority of the proposed frameworks.

4. Results

The performance of the RGCN models (i.e., BERT-RGCN, Mordred-RGCN, and Bert and Mordred-RGCN) was measured based on the Bert-based features, the Mordred-

based features, and the combined set of the Bert and Mordred features as shown in Table 5, respectively, resulting in three main expression entities of the independent experiments. However, two proposed models (i.e., BERT-RGCN and Mordred-RGCN) while training, RGCN was initially trained and tested using only antimalarial drug compounds' features followed by a combination of multiple features of the drugs and targets for a further comparison task, analysis, and contribution of the Plasmodium potential targets' features.

Furthermore, to understand the importance of different architectures and hyperparameters in this study, we report the ablation studies on different layers' tasks, as shown in Table 5. We can detect a slight decrease in the model's accuracy when transitioning from a single RGCNconv. layer to stack up double RGCNconv. layers. Additionally, we observe a slight variation in sensitivity following adjustments in the hyperparameters. Specifically, the sensitivity changes from 0.9968 with a single RGCNconv. layer to 0.9904 with two RGCNconv. layers.

In the detailed models' performance evaluation, we divided the Results Section into two parts of Sections 4.1 and 4.2.

4.1. Model Training and Evaluation

The RGCN model was first implemented to demonstrate its role in the prediction of antimalarial drugs during training. The RGCN model was trained using two types of features as discussed in Section 2.2 resulting in two main models (i.e., BERT-RGCN, which was trained using Bert-based features, and Mordred-RGCN, which utilizes Mordred-based features as shown in Table 7). Both models were trained using the 10-fold cross-validation technique and evaluated with prepared graph data. One might debate concerning the training performance of the proposed models; Table 7 illustrates the training results of these models.

Table 7. Performance comparison of the model (i.e., BERT-RGCN) on training dataset using BERT-based features.

Model	Training result	AmD_F	AmD_F and $PfAcA_F$	AmD_F and $GGPP_F$	AmD_F and $PfMAGL_F$	AmD_F and Ts_F
BERT-RGCN	Accuracy	0.9937	0.9943	0.9914	0.9958	0.9925
	Sensitivity	0.9917	0.9903	0.9828	0.9945	0.9855
	Specificity	0.9953	0.9974	0.9979	0.9969	0.9979
	MCC	0.9873	0.9885	0.9826	0.9915	0.9850
	AUROC	0.9992	0.9989	0.9990	0.9993	0.9990
	AUPRC	0.9992	0.9990	0.9992	0.9995	0.9992
Mordred-RGCN	Accuracy	0.9967	0.9952	0.9940	0.9973	0.9934
	Sensitivity	0.9959	0.9952	0.9952	0.9965	0.9876
	Specificity	0.9974	0.9953	0.9932	0.9979	0.9979
	MCC	0.9933	0.9903	0.9880	0.9945	0.9868
	AUROC	0.9999	0.9999	0.9998	0.9998	0.9998
	AUPRC	0.9989	0.9989	0.9988	0.9989	0.9974

AmD_F —Antimalarial drug features. $PfAcA_F$ —Acetyl CoA Synthetase features. $GGPP_F$ —Geranylgeranyl Pyrophosphate Synthase features. $PfMAGL_F$ —Monoacylglycerol Lipase features. Ts_F —All targets' features.

4.1.1. RGCN Model Training Performance Featuring BERT-Based Features

Table 7 shows the performance of RGCN (i.e., BERT-RGCN) when trained using the features based on BERT for SMILES. The RGCN model achieved high and comparable performance in terms of all performance metrics when trained using only drugs' features (i.e., AmD_F) and also when drugs' features were combined with antimalarial potential targets' features, as shown in Table 7. However, it revealed slightly higher performance achievement (i.e., accuracy: 0.9985, sensitivity: 0.9945, AUROC: 0.9993, and AUPRC: 0.9995) when trained using concatenated features of the antimalarial drugs and the monoacylglycerol Lipase target (i.e., AmD_F and $PfMAGL_F$) as shown in Table 7.

4.1.2. RGCN Model Training Performance Featuring Mordred-Based Features

Meanwhile, Table 7 shows the training performance of RGCN (i.e., Mordred-RGCN) when trained using SMILES Mordred-based features. Our model also achieved remarkable and comparable performance in terms of all performance metrics (i.e., accuracy, sensitivity, specificity, MCC, AUROC, and AUPRC) when trained using only Mordred drugs' features (i.e., AmD_F). When Mordred the drugs' features set was combined or concatenated with the Mordred antimalarial potential targets' features set as shown in Table 7, the proposed model still exhibited high performance. It also showed slightly higher performance achievement (i.e., accuracy: 0.9973, sensitivity: 0.9965, specificity: 0.9979, and MCC: 0.9945) when trained using concatenated features of antimalarial drugs and the monoacylglycerol Lipase target (i.e., AmD_F and $PfMAGL_F$) as shown in Table 7. However, Mordred-RGCN exhibited a remarkably similar performance (i.e., AUROC and AUPRC) during training as BERT-RGCN. In summary, Mordred-RGCN and BERT-RGCN, as shown in Table 7, prevailed with high performance regardless of the slight difference.

In general, BERT-RGCN exhibits a high performance similar to Mordred-RGCN in terms of all metrics regardless of the type of feature set, whether it is only antimalarial drugs or when combined with the targets' feature set. However, $PfMAGL_F$ in particular appears to be slightly more promising compared to other targets during training.

4.2. Model Testing and Evaluation

The RGCN models (i.e., BERT-RGCN, which utilized Bert-based features, and Mordred-RGCN, which utilized Mordred-based features) after 10-fold cross-validation training with prepared graph data, were therefore tested and evaluated. The testing results of the models are shown in Table 8.

Table 8. Ablation experimental results and independent test performance on BERT features.

Model	Test result	AmD_F	AmD_F and $PfAcA_F$	AmD_F and $GGPP_F$	AmD_F and $PfMAGL_F$	AmD_F and Ts_F
BERT-RGCN	Accuracy	0.9972	0.9972	0.9917	0.9972	0.9958
	Sensitivity	0.9968	0.9968	0.9839	0.9968	0.9968
	Specificity	0.9976	0.9976	0.9976	0.9976	0.9951
	MCC	0.9943	0.9943	0.9831	0.9943	0.9915
	AUROC	0.9971	0.9974	0.9977	0.9965	0.9942
	AUPRC	0.9960	0.9961	0.9972	0.9923	0.9896
Mordred-RGCN	Accuracy	0.9958	0.9958	0.9972	0.9958	0.9958
	Sensitivity	0.9968	0.9968	0.9968	0.9968	0.9968
	Specificity	0.9951	0.9951	0.9976	0.9951	0.9951
	MCC	0.9915	0.9915	0.9943	0.9915	0.9915
	AUROC	0.9995	0.9991	0.9990	0.9986	0.9971
	AUPRC	0.9996	0.9992	0.9991	0.9986	0.9973

AmD_F —Antimalarial drug features. $PfAcA_F$ —Acetyl CoA Synthetase features. $GGPP_F$ —Geranylgeranyl Pyrophosphate Synthase features. $PfMAGL_F$ —Monoacylglycerol Lipase features. Ts_F —All targets features.

RGCN Model Test Performance Featuring BERT-Based and Mordred-Based Features

Test results for the proposed method (i.e., BERT-RGCN model) are presented in Table 8. The presented table contains test results for the BERT-RGCN model, detailing its performance in various combinations of features. Each row corresponds to a specific performance metric, while each column represents a different scenario with distinct sets of features. Below is a detailed explanation of each metric.

Accuracy represents the overall correctness of the RGCN models' (i.e., BERT-RGCN and Mordred-RGCN) predictions. Accuracy performances were higher score values, indicating better accuracy. The BERT-RGCN and Mordred-RGCN models achieved high accuracy consistently across all combination of features, ranging from 0.9917 to 0.9972 and 0.9958 to 0.9972, respectively, as shown in Table 8. Sensitivity, also known as True Positive Rate or Recall, measures the ability of the model to correctly identify positive instances. This indicates the proportion of actual positive instances correctly classified by RGCN models.

In Table 8, the BERT-RGCN model demonstrated high sensitivity, ranging from 0.9839 to 0.9968. Meanwhile, the Mordred-RGCN model consistently achieved a high sensitivity of 0.9968 across all features. Specificity measures the ability of RGCN-based models to correctly identify negative instances. This indicates the proportion of actual negative instances correctly classified by the proposed methods (i.e., BERT-RGCN and Mordred-RGCN models), and values close to 1 signify promising performance. The BERT-RGCN model consistently showed high specificity, ranging from 0.9951 to 0.9976. Consequently, Mordred-RGCN also had a similar performance of 0.9951 across all features and 0.9976 for the remaining combination of features, as shown in Table 8.

In this study, MCC takes into account true and false positives and negatives, providing a balanced measure even for imbalanced datasets. The range of values achieved by BERT-RGCN (i.e., MCC values ranging from 0.9831 to 0.9943) and Mordred-RGCN (i.e., 0.9943 for AmD_F and $GGPP_F$ and 0.9915 for other features) were high, which signifies perfect predictions of antimalarial drugs of the models as shown in Table 8.

Furthermore, the models' performances in this study were evaluated through the implemented AUROC metric. AUROC evaluates the trade-off between the true positive rate and the false positive rate across different threshold settings. Values close to 1 resemble higher values, which signify better discrimination. As shown in Table 8, the BERT-RGCN and Mordred-RGCN models exhibited high AUROC values, ranging from 0.9942 to 0.9977 and from 0.9971 to 0.9995, respectively.

Lastly, AUPRC measures the trade-off between precision and recall in various threshold settings. The values range from 0 to 1, with higher values indicating better precision and recall. The BERT-RGCN model demonstrated robust performance with AUPRC values ranging from 0.9896 to 0.9972. Meanwhile, Mordred-RGCN prevailed with high AUPRC (i.e., 0.9973 to 0.9996 shown by AmD_F) as shown in Table 8.

One might be curious as to what the performance of the RGCN model using combined Bert and Mordred features will be. To answer the aforementioned concern, we performed another experiment using the concatenated features (i.e., Bert and Mordred features) and the test performance is as shown in Table 5. In summary, the BERT-RGCN and Mordred-RGCN models consistently performed well across different feature combinations, showcasing high accuracy, sensitivity, specificity, MCC, AUROC, and AUPRC values. These results suggest the effectiveness of the models in predicting new antimalarial drugs against *Plasmodium falciparum* in various scenarios based on different sets of features of drugs and potential antimalarial targets.

5. Discussion

The integration of NLP, represented by the SMILES-based BERT model, and GNN, exemplified by RGCN, presents a novel and potent approach to identify novel antimalarial drug candidates. The use of these advanced ML techniques has shown promise in overcoming the complexities associated with drug–target interactions, offering a pathway towards combating the challenges posed by drug resistance in malaria treatment. The successful prediction of potential antimalarial drugs validated the efficacy of our proposed approach. By harnessing multi-relational data that encompass critical antimalarial drug targets such as PfAcAS, F/GGPPS, and PfMAGL, our proposed models such as BERT-RGCN and Mordred-RGCN demonstrated their efficacy in capturing multi-drug–target relations, showcasing the potential of the proposed approaches in uncovering novel drug candidates with varying modes of action against malaria. The prediction accuracy, validated through rigorous testing and cross-validation methodologies, substantiates the reliability and robustness of the proposed frameworks. This capability is perhaps pivotal in uncovering novel compounds capable of disrupting the malaria parasite's life cycle at various stages.

Comprehensive analysis of graph-structured data revealed various drug–target interactions, shedding light on potential mechanisms of action and the polypharmacology essential in addressing the complexity of malaria treatment. The ability of the RGCN models such as BERT-RGCN and Mordred-RGCN to identify and infer within complex

relational data allowed the identification of previously unexplored antimalarial drug–target interactions, presenting their potential and a foundation for further exploration and experimentation for accelerating drug development in the realm of antimalarial drug discovery.

However, it is important to acknowledge certain limitations. The availability of comprehensive and curated datasets remains a challenge in drug discovery research. Despite the success of the models (i.e., BERT-RGCN and Mordred-RGCN) in navigating complex relational data, the quality and depth of available data profoundly influence predictive accuracy. Additionally, the need for experimental validation of predicted drug candidates remains imperative to confirm their efficacy and safety profiles. Moreover, while our approach showcases promise, further refinement and optimization are necessary. Fine-tuning the model architecture and leveraging larger and more diverse datasets could enhance the prediction accuracy and broaden the scope of potential drug candidates identified.

In conclusion, the effectiveness of the integrated BERT model and RGCN in predicting potential antimalarial drugs offers a promising avenue to accelerate drug discovery efforts and address the challenges posed by drug resistance in the fight against malaria. The results also showed that the use of new potential targets such as PfAcAS, F/GGPPS, and PfMAGL for antimalarial drugs regardless of the extraction of features was promising and with the use of graphs to express antimalarial drug molecules and potential targets in *Plasmodium*, RGCN yielded better results. The success in predicting drug candidates against *Plasmodium* highlights the potential impact of advanced ML in addressing global health challenges. Moving forward, continued advances in methodologies and collaborative efforts between computational and experimental research are crucial in translating predictions into tangible therapeutic interventions to combat malaria effectively.

6. Conclusions

The pursuit of novel antimalarial drugs to combat the persistent threat posed by *Plasmodium* demands innovative and interdisciplinary approaches. Our study embarked on a pioneering journey, leveraging the synergy between advanced ML techniques such as NLP (i.e., BERT) and GNN (i.e., RGCN) to predict antimalarial drugs against *Plasmodium falciparum*. Graph-structured data can effectively represent the relationships between antimalarial drugs and *Plasmodium* potential targets, allowing for a more holistic understanding of their interactions. Training the RGCN on multi-relational data related to potential targets like PfAcAS, F/GGPPS, and PfMAGL seems like a strategic move to capture diverse malaria drug–target interactions.

Integration of the SMILES-based BERT model with the robust RGCN showcased immense potential to decipher intricate drug–target interactions and accelerate drug discovery processes, especially in combating diseases like malaria where resistance to existing treatments is a significant challenge. The model's adeptness in navigating multi-relational data, particularly focusing on critical antimalarial drug targets like PfAcAS, F/GGPPS, and PfMAGL, underscores its capacity to uncover promising compounds capable of disrupting the malarial various stages of the parasite's life cycle. Our study's success in predicting antimalarial drugs offers a glimpse into the transformative impact of computational methodologies in drug discovery. However, this progress does not exist in isolation; it is based on collaborative efforts across scientific disciplines, emphasizing the importance of interdisciplinary research in addressing global health challenges.

Nevertheless, several challenges persist; these include dependence on data availability and quality, the necessity for rigorous experimental validation, and the ongoing evolution of the malaria parasite. The journey towards translating computational predictions into viable therapeutic interventions demands a concerted effort from both computational and experimental researchers. The convergence of cutting-edge ML methodologies with traditional drug discovery approaches holds promise. Fine-tuning and predictive models, expanding datasets, fostering collaborations, and integrating insights from diverse domains will accelerate the field towards identifying more effective and resilient antimalarial drugs. In conclusion, our study represents a crucial step towards leveraging advanced ML in the

pursuit of addressing the challenges of malaria treatment. The intersection of computational methodologies and drug discovery research presents a promising path forward in the quest to alleviate the global burden of malaria, emphasizing the transformative potential of integrative approaches in shaping the future of medicine against malaria. We believe that our method that exploits the RGCN and BERT is applicable to other domains such as object recognition and autonomous driving, because it is based and relies on distances in the embedding representations which can be applicable in domains where the data has local patterns.

Author Contributions: M.E.M., G.E.N. and Y.-S.J., data curation; M.E.M., G.E.N., Y.-S.J. and K.J., conceptualization; M.E.M., G.E.N. and Y.-S.J. conceived the experiment(s); M.E.M. conducted the experiment(s); M.E.M., K.J. and Y.-S.J. analyzed the results; M.E.M. and Y.-S.J. wrote the original manuscript; M.E.M., G.E.N., Y.-S.J. and K.J. reviewed and edited the manuscript; M.E.M. and Y.-S.J., investigation; M.E.M., Y.-S.J. and K.J., methodology. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R111A3053015). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2023R1A2C1003355).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and/or analysed during the this study are publicly available at <https://sites.google.com/view/medardemswahili/publications-awards> (accessed on 25 September 2023).

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

References

1. World Health Organization. Malaria. Available online: <https://www.who.int/news-room/fact-sheets/detail/malaria> (accessed on 27 October 2023).
2. Mswahili, M.E.; Martin, G.L.; Woo, J.; Choi, G.J.; Jeong, Y.S. Antimalarial drug predictions using molecular descriptors and machine learning against plasmodium falciparum. *Biomolecules* **2021**, *11*, 1750. [CrossRef]
3. World Health Organization. World Malaria Report 2022. Available online: <https://www.who.int/publications/i/item/9789240064898> (accessed on 27 October 2023).
4. Forte, B.; Otilie, S.; Plater, A.; Campo, B.; Dechering, K.J.; Gamo, F.J.; Goldberg, D.E.; Istvan, E.S.; Lee, M.; Lukens, A.K.; et al. Prioritization of molecular targets for antimalarial drug discovery. *ACS Infect. Dis.* **2021**, *7*, 2764–2776. [CrossRef]
5. Shibeshi, M.A.; Kifle, Z.D.; Atnafie, S.A. Antimalarial drug resistance and novel targets for antimalarial drug discovery. *Infect. Drug Resist.* **2020**, *13*, 4047–4060. [CrossRef]
6. Arendse, L.B.; Wyllie, S.; Chibale, K.; Gilbert, I.H. Plasmodium kinases as potential drug targets for malaria: Challenges and opportunities. *ACS Infect. Dis.* **2021**, *7*, 518–534. [CrossRef]
7. Tajbakhsh, E.; Kwent, T.E.; Kheyri, P.; Nezaratizade, S.; Lindsay, D.S.; Khamesipour, F. Antiplasmodial, antimalarial activities and toxicity of African medicinal plants: A systematic review of literature. *Malar. J.* **2021**, *20*, 349. [CrossRef]
8. Gamo, F.J.; Sanz, L.M.; Vidal, J.; De Cozar, C.; Alvarez, E.; Lavandera, J.L.; Vanderwall, D.E.; Green, D.V.; Kumar, V.; Hasan, S.; et al. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305–310. [CrossRef]
9. Guiguemde, W.A.; Shelat, A.A.; Bouck, D.; Duffy, S.; Crowther, G.J.; Davis, P.H.; Smithson, D.C.; Connelly, M.; Clark, J.; Zhu, F.; et al. Chemical genetics of Plasmodium falciparum. *Nature* **2010**, *465*, 311–315. [CrossRef]
10. Plouffe, D.; Brinker, A.; McNamara, C.; Henson, K.; Kato, N.; Kuhen, K.; Nagle, A.; Adrián, F.; Matzen, J.T.; Anderson, P.; et al. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 9059–9064. [CrossRef]
11. Macalino, S.J.Y.; Gosu, V.; Hong, S.; Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharmacol. Res.* **2015**, *38*, 1686–1701. [CrossRef]
12. Yin, Q.; Cao, X.; Fan, R.; Liu, Q.; Jiang, R.; Zeng, W. DeepDrug: A general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction. *Quant. Biol.* **2020**, *11*, 260–274. [CrossRef]
13. Lu, Y.; Liu, J.; Jiang, T.; Cui, Z.; Wu, H. Drug-target Binding Affinity Prediction Based on Three-branched Multiscale Convolutional Neural Networks. *Curr. Bioinform.* **2023**, *18*, 853–862. [CrossRef]

14. Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-learning-based drug–target interaction prediction. *J. Proteome Res.* **2017**, *16*, 1401–1409. [[CrossRef](#)] [[PubMed](#)]
15. Abbasi, K.; Razzaghi, P.; Poso, A.; Ghanbari-Ara, S.; Masoudi-Nejad, A. Deep learning in drug target interaction prediction: Current and future perspectives. *Curr. Med. Chem.* **2021**, *28*, 2100–2113. [[CrossRef](#)] [[PubMed](#)]
16. Yang, Z.; Bai, B.; Long, J.; Wei, P.; Li, J. Multi-scale Feature Fusion Neural Network for Accurate Prediction of Drug-Target Interactions. In *International Conference on Neural Information Processing, Proceedings of the 30th International Conference, ICONIP 2023, Changsha, China, 20–23 November 2023*; Springer: Singapore, 2023; pp. 176–188.
17. Sadri, A. Is Target-Based Drug Discovery Efficient? Discovery and “Off-Target” Mechanisms of All Drugs. *J. Med. Chem.* **2023**, *66*, 12651–12677. [[CrossRef](#)] [[PubMed](#)]
18. Schalkwijk, J.; Allman, E.L.; Jansen, P.A.; De Vries, L.E.; Verhoef, J.M.; Jackowski, S.; Botman, P.N.; Beuckens-Schortinghuis, C.A.; Koolen, K.M.; Bolscher, J.M.; et al. Antimalarial pantotheneamide metabolites target acetyl-coenzyme A biosynthesis in *Plasmodium falciparum*. *Sci. Transl. Med.* **2019**, *11*, eaas9917. [[CrossRef](#)] [[PubMed](#)]
19. Summers, R.L.; Pasaje, C.F.A.; Pisco, J.P.; Striepen, J.; Luth, M.R.; Kumpornsin, K.; Carpenter, E.F.; Munro, J.T.; Lin, D.; Plater, A.; et al. Chemogenomics identifies acetyl-coenzyme A synthetase as a target for malaria treatment and prevention. *Cell Chem. Biol.* **2022**, *29*, 191–201. [[CrossRef](#)] [[PubMed](#)]
20. Gisselberg, J.E.; Herrera, Z.; Orchard, L.M.; Llinás, M.; Yeh, E. Specific inhibition of the bifunctional farnesyl/geranylgeranyl diphosphate synthase in malaria parasites via a new small-molecule binding site. *Cell Chem. Biol.* **2018**, *25*, 185–193. [[CrossRef](#)]
21. Yoo, E.; Schulze, C.J.; Stokes, B.H.; Onguka, O.; Yeo, T.; Mok, S.; Gnädig, N.F.; Zhou, Y.; Kurita, K.; Foe, I.T.; et al. The antimalarial natural product salinipostin A identifies essential α/β serine hydrolases involved in lipid metabolism in *P. falciparum* parasites. *Cell Chem. Biol.* **2020**, *27*, 143–157. [[CrossRef](#)]
22. Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv* **2020**, arXiv:2010.09885
23. Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 7–10 September 2019*; pp. 429–436.
24. Huang, K.; Xiao, C.; Glass, L.M.; Sun, J. MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* **2021**, *37*, 830–836. [[CrossRef](#)]
25. Thafar, M.A.; Alshahrani, M.; Albaradei, S.; Gojobori, T.; Essack, M.; Gao, X. Affinity2Vec: Drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Sci. Rep.* **2022**, *12*, 4751. [[CrossRef](#)] [[PubMed](#)]
26. Lin, X. DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction. *arXiv* **2020**, arXiv:2003.13902.
27. Li, Y.; Qiao, G.; Wang, K.; Wang, G. Drug–target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* **2022**, *23*, bbab346. [[CrossRef](#)]
28. Liu, Q.; Wan, J.; Wang, G. A survey on computational methods in discovering protein inhibitors of SARS-CoV-2. *Brief. Bioinform.* **2022**, *23*, bbab416. [[CrossRef](#)]
29. He, Y.; Shen, Z.; Zhang, Q.; Wang, S.; Huang, D.S. A survey on deep learning in DNA/RNA motif mining. *Brief. Bioinform.* **2021**, *22*, bbaa229. [[CrossRef](#)] [[PubMed](#)]
30. Wang, L.; You, Z.H.; Huang, Y.A.; Huang, D.S.; Chan, K.C. An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network. *Bioinformatics* **2020**, *36*, 4038–4046. [[CrossRef](#)]
31. Wang, L.; You, Z.H.; Huang, D.S.; Zhou, F. Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *17*, 972–980. [[CrossRef](#)] [[PubMed](#)]
32. Wang, S.; Jiang, M.; Zhang, S.; Wang, X.; Yuan, Q.; Wei, Z.; Li, Z. MCN-CPI: Multiscale convolutional network for compound–protein interaction prediction. *Biomolecules* **2021**, *11*, 1119. [[CrossRef](#)]
33. Yang, Z.; Zhong, W.; Zhao, L.; Chen, C.Y.C. MGraphDTA: Deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chem. Sci.* **2022**, *13*, 816–833. [[CrossRef](#)]
34. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
35. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 10–48550.
36. Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated graph sequence neural networks. *arXiv* **2015**, arXiv:1511.05493.
37. Bresson, X.; Laurent, T. Residual gated graph convnets. *arXiv* **2017**, arXiv:1711.07553.
38. Xu, C.; Liu, Q.; Huang, M.; Jiang, T. Reinforced molecular optimization with neighborhood-controlled grammars. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 8366–8377.
39. Ding, K.; Zhou, M.; Wang, Z.; Liu, Q.; Arnold, C.W.; Zhang, S.; Metaxas, D.N. Graph convolutional networks for multi-modality medical imaging: Methods, architectures, and clinical applications. *arXiv* **2022**, arXiv:2202.08916.
40. Duvenaud, D.K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
41. Fout, A.; Byrd, J.; Shariat, B.; Ben-Hur, A. Protein interface prediction using graph convolutional networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

42. Feng, Q.; Dueva, E.; Cherkasov, A.; Ester, M. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv* **2018**, arXiv:1807.09741.
43. Zamora-Resendiz, R.; Crivelli, S. Structural learning of proteins using graph convolutional neural networks. *bioRxiv* **2019**. [CrossRef]
44. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
45. Ju, W.; Fang, Z.; Gu, Y.; Liu, Z.; Long, Q.; Qiao, Z.; Qin, Y.; Shen, J.; Sun, F.; Xiao, Z.; et al. A Comprehensive Survey on Deep Graph Representation Learning. *arXiv* **2023**, arXiv:2304.05055.
46. Hu, Z.; Dong, Y.; Wang, K.; Sun, Y. Heterogeneous graph transformer. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2704–2710.
47. Creating Message Passing Networks. Available online: https://pytorch-geometric.readthedocs.io/en/latest/notes/create_gnn.html (accessed on 15 January 2023).
48. Mswahili, M.E.; Lee, M.J.; Martin, G.L.; Kim, J.; Kim, P.; Choi, G.J.; Jeong, Y.S. Cocystal prediction using machine learning models and descriptors. *Appl. Sci.* **2021**, *11*, 1323. [CrossRef]
49. Niazi, S.K.; Mariam, Z. Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis. *Pharmaceuticals* **2023**, *17*, 22. [CrossRef]
50. Niazi, S.K.; Mariam, Z. Recent advances in machine-learning-based chemoinformatics: A comprehensive review. *Int. J. Mol. Sci.* **2023**, *24*, 11488. [CrossRef]
51. DeepChem Tokenizers. Available online: https://deepchem.readthedocs.io/en/2.4.0/api_reference/tokenizers.html (accessed on 25 September 2023).
52. Schwaller, P.; Probst, D.; Vaucher, A.C.; Nair, V.H.; Kreutter, D.; Laino, T.; Reymond, J.L. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **2021**, *3*, 144–152. [CrossRef]
53. Liu, Q.; Deng, J.; Liu, M. Classification models for predicting the antimalarial activity against Plasmodium falciparum. *SAR QSAR Environ. Res.* **2020**, *31*, 313–324. [CrossRef]
54. Danishuddin; Madhukar, G.; Malik, M.; Subbarao, N. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR QSAR Environ. Res.* **2019**, *30*, 543–560. [CrossRef] [PubMed]
55. Moriwaki, H.; Tian, Y.S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef] [PubMed]
56. Featurizers MordredDescriptors. Available online: https://deepchem.readthedocs.io/en/latest/api_reference/featurizers.html#mordreddescriptors (accessed on 25 September 2023).
57. sklearn.preprocessing.StandardScaler. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> (accessed on 25 September 2023).
58. Cui, C.; Ding, X.; Wang, D.; Chen, L.; Xiao, F.; Xu, T.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. Drug repurposing against breast cancer by integrating drug-exposure expression profiles and drug–drug links based on graph neural network. *Bioinformatics* **2021**, *37*, 2930–2937. [CrossRef] [PubMed]
59. Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* **2019**, *35*, 104–111. [CrossRef] [PubMed]
60. Zhou, J.P.; Chen, L.; Guo, Z.H. iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* **2020**, *36*, 1391–1396. [CrossRef]
61. Zhou, J.P.; Chen, L.; Wang, T.; Liu, M. iATC-FRAKEL: A simple multi-label web server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* **2020**, *36*, 3568–3569. [CrossRef]
62. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J.K.; Ceulemans, H.; Clevert, D.A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441–5451. [CrossRef]
63. Zhao, X.; Chen, L.; Guo, Z.H.; Liu, T. Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* **2019**, *14*, 709–720. [CrossRef]
64. Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; Weinberger, K. Simplifying graph convolutional networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6861–6871.
65. Kim, S.; Bae, S.; Piao, Y.; Jo, K. Graph convolutional network for drug response prediction using gene expression data. *Mathematics* **2021**, *9*, 772. [CrossRef]
66. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In Proceedings of the Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, 3–7 June 2018; Proceedings 15; Springer: Berlin/Heidelberg, Germany, 2018; pp. 593–607.
67. Thanapalasingam, T.; van Berkel, L.; Bloem, P.; Groth, P. Relational graph convolutional networks: A closer look. *PeerJ Comput. Sci.* **2022**, *8*, e1073. [CrossRef] [PubMed]
68. Ding, Y.; Jiang, X.; Kim, Y. Relational graph convolutional networks for predicting blood–brain barrier penetration of drug molecules. *Bioinformatics* **2022**, *38*, 2826–2831. [CrossRef] [PubMed]
69. Fey, M.; Lenssen, J.E. Fast graph representation learning with PyTorch Geometric. *arXiv* **2019**, arXiv:1903.02428.

70. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
71. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *OpenAI* **2018**.

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