Topic: Interpretable Transformer Based Graph Neural Network (GNN) Model Prediction of Antimalarial Drug Efficacy Using SHAP and the Captum Framework

ABSTRACT

Malaria has posed a global health challenge to humans since time immemorial, with much research on drug discovery trying to combat this parasite.  One of the challenges in antimalaria drug discovery is getting an exact, efficient drug that is active in combating the malaria parasite. Many advanced computational approaches have been developed to predict antimalarial drugs, but many of these approaches fail to explain how this model makes its decision in predicting antimalarial drugs. This current study will provide explainability for transformer-based GNN antimalaria drug prediction from its SMILES data. The data use was obtained from ChEMBL. This data has been used by different researchers for similar studies but using different approaches. The result of the study showed that transformer-based GNN architecture has the best performance when compared to GNN and transformer-based GAT models with AUC values of 0.947, 0.947, and 0.939, respectively. Other metrics use to access the performance of this model were accuracy, precision, recall, f1-score, and support were the transformer base GNN shows a remarkably better result. Using SHAP and the Captum framework to perform explainability showed the influence of different molecular descriptors and the SMILES atomic features the model considered in making its prediction. The combination of SHAP and CAPTUM in the explainability of the model performance in predicting antimalaria drug efficacy creates a novel area in the antimalaria drug research.

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

Introduction

Malaria continues to pose a major global public health threat, particularly in tropical and subtropical regions. According to the World Health Organization (WHO), an estimated 263 million cases and 597,000 deaths were reported across 83 countries in 2023. The WHO African Region bears the greatest burden, accounting for 94% of all cases (246 million) and 95% of deaths (569,000) [1]. Malaria, primarily caused by *Plasmodium falciparum*, is both preventable and curable, yet remains life-threatening in areas with limited access to early diagnosis and treatment.

In high-transmission regions, especially across sub-Saharan Africa, frequent reinfections make it difficult to confirm whether a treatment has truly cleared the infection. This challenge is compounded by the steady rise in antimalarial drug resistance over the past three decades, which has undermined the effectiveness of many cost-efficient first-line treatments and threatens the progress made in malaria control [2]. This urgent situation highlights the need for advanced, reliable approaches to assess drug efficacy and identify novel therapeutic candidates.

Drug discovery methods have evolved from traditional high-throughput screening (HTS) to advanced computational techniques. Among these, Quantitative Structure–Activity Relationship (QSAR)-based virtual screening has become a valuable tool for identifying novel compounds active against *P. falciparum* [3]. QSAR models enable medicinal chemists to analyse how the physicochemical properties of compounds influence biological activity, guiding the rational design and optimization of new drug candidates [4]. Compared to HTS, which often yields hit rates below 0.1%; validated QSAR models offer a more targeted, cost-efficient approach, with typical hit rates ranging from 1% to 40% [3,5].

Over the past 50 years, QSAR modelling has progressed from analysing small, structurally related compound sets with simple regression techniques to handling large, diverse datasets using advanced machine learning algorithms [6]. Numerous studies have applied these methods to antimalarial drug discovery. Noviandy et al. (2024) developed an interpretable LightGBM-based QSAR model enhanced with SHAP analysis, achieving 86% predictive accuracy for antimalarial activity and identifying key molecular descriptors to improve both performance and interpretability. Mswahili et al. (2024) integrated BERT for SMILES tokenization with a Relational Graph Convolutional Network (RGCN), achieving 99.7% accuracy in predicting antimalarial drugs against *P. falciparum* by modelling interactions with three prioritized targets (PfAcAS, F/GGPPS, and PfMAGL), outperforming traditional approaches. Danishuddin et al. (2019) used ChEMBL data with SVM and XGBoost models, achieving ~85% accuracy and reliable probability predictions for identifying active compounds. Bharti et al. (2017) applied molecular descriptors from apicoplast-specific bioassays, finding that decision trees, SVM, and Random Forest achieved the highest predictive performance, with test set accuracies between 93% and 95%, making them highly effective for screening large chemical libraries.

In this work, we propose an interpretable transformer-based Graph Neural Network (GNN) model integrated within the Captum framework to predict the efficacy of antimalarial compounds. GNNs excel at capturing molecular graph structures for property prediction [8], while transformer architectures enhance performance by learning long-range dependencies. However, most GNN and transformer-based models operate as “black boxes,” limiting interpretability; a critical factor in pharmacology, where transparency supports validation and regulatory acceptance. To address this, we employ SHAP (SHapley Additive exPlanations) to attribute model predictions to molecular features, and leverage the Captum framework to ensure modular, reproducible, and interpretable graph learning. Our approach aims to achieve competitive predictive performance while providing mechanistic insights into the structural features driving antimalarial activity.

METHODOLOGY

Dataset

The smile antimalaria drug data were collected from the ChEMBL database [67] (https://www.ebi.ac.uk/chembl/, accessed by May 2024) and the PubChem database [66] (https://pubchem.ncbi.nlm.nih.gov/, accessed by May 2024). Specifically, the term “Plasmodium falciparum” was used as a keyword to obtain all related whole-organism bioactivities in PubChem. This data has been used by Mswahili et al. [8], Mswahili et al. [14], Danishuddin et al. [9], and Noviandy et al. [7] in a similar work on antimalaria drug discovery. The classification of the antimalarial drug molecule into active and inactive was done according to the anti-plasmodial activities of the compounds in Mswahili et al. [8], Mswahili et al. [14], and Danishuddin et al. [9] studies, where antimalaria drug molecule is classified active if the compound . The inactive candidates’ inactive instances are experimentally verified as unsuccessful candidates.

Data Preprocessing

The raw SMILES data make of molecular structure of the data was subjected to data wrangling and cleaning. This process involves removing empty values, duplicated entries, and to ensure consistency and reproductivity, molecules with structural errors were excluded from the data.

The data label was encoded into numerical variables were class labels “active” encoded to 0 and label labelled “inactive” encode 1. The class distribution count was made up of 12685 and 111060 inactive. Because the distribution of the class is imbalance, the imblearn [11, 12, 13] library was used to balance the data using the principle of under sampling of the majority inactive class.

Feature Engineering

The features engineering used in this study involve atomic level descriptors and molecular level descriptors.

Atom Features Extraction

The structural properties of the antimalaria compounds, atomic level features of the smiles structure were extracted using the RDKit. During this process, entries that failed SMILES parsing were not used in the atomic feature engineering to maintain data quality. For each valid antimalaria compound, atomic descriptors were comprehensively curated. The atomic features include, element type, formal charge, hybridization state, aromaticity, sum of attached hydrogen atoms, formal charge, degree of bond, atomic number, ring size (3-8), number of radical electrons, implicit valence, and chirality. In antimalarial drug research, this feature has play an important role in molecular reactivity, bioavailability factors and reactivity of the molecules [8, 14, 15]. Across the dataset, the features extracted were aggregated to measure the diversity of their distribution and values that occur frequently.

Molecular Descriptors

In the representation of structural and physiochemical properties of compound of drug and small molecules quantitatively, molecular descriptors have played a big role in achieving this [16, 17, 18].

To enable analysis of the antimalaria drug discovery features, a comprehensive set of molecular descriptors were computed using the RDKit. RDKit was used because of numerous supports for calculating diverse collection of molecular descriptors important for drug discovery and its flexibilities. With RDKit extraction of physiochemical and topological features directly from SMILEs string and then combine it with machine learning workflow that is python base. For this work, the pipeline for the molecular descriptor is designed to capture 91 molecular descriptors which is specifically selected for antimalaria drug analysis research. To compute the molecular descriptors, firstly, we prepared a canonical SMILEs string for each of the antimalaria drug compounds. Secondly, the curated set of molecular descriptors were computed into 1D and 2D molecular features for each of the compound. Each antimalaria drug compound molecular descriptor was converted to real value vector features. To make sure that there is comparability with the different range of values in the molecular descriptors, we applied standardisation were each of the molecular descriptor’s value is arranged between 0 and 1. This step help to improve the performance of the model by aligning the contribution of the each selected molecular descriptors features to the model [18,19].

Topology Graph Construction

In the graphical topological representation of the smile molecular of the antimalaria drug compound into graph involves encoding of the molecular compounds into graph. In this graphical representation, the atoms form the vertices while the edges are consisting of the covalent bonds. This theoretical graph representation conserves both global and local chemical environment and topological features respectively, which are very important for activity and structural representation modelling in research involving drug discovery [21, 22, 25].

In the graph presentation of each drug molecules, the graph

is a unidirectional graph where V depicts the vertex representing the atoms, E depict the edge representing the present chemical bond, denotes the matrix of the node features with a and denotes the matrix of the edge attribute with .

We included molecular descriptors in our methodology which consist of the atomic, bond and molecular scales that can help give more information about representation for the drug molecular structure useful for downstream graph neural network learning.

Our methodology integrates hierarchical molecular descriptors spanning atomic, bond, and molecular scales to provide comprehensive structural representation for downstream graph neural network learning.

Node Feature Engineering

To be able to describe each atom in the drug discovery we designed a 27-dimenstonal features vector, this vector was able to capture chemical and structural relationship in the molecule. The representation of the feature vector atom is

where each features vector captures the properties of the atom that includes atomic number of the identified element, degree of the number of bond present. The total degree consists of the hybridization state of the compounds ( ), hydrogen atoms and the valence electron of the atom. These critical properties provide important information about the pattern of bond and atomic behaviour that happens within the compound [26]. The aromaticity indicator, hydrogen count (which can be total, implicit, and explicit), and the properties of the electron like the formal charge, all makes up the chemical properties of the atoms. This property helps the model to understand the chemical reactivity and the potential of the bonding that exist that is important for the molecular interaction prediction [23, 24, 25].

The ring related features is used to explain the structural properties which is very important for the antimalaria drug molecules. These features consist of the presence of ring in the atom, the number of rings, which is from 3 to 8 membered ring, and specifically the indicators for the ring. The ring features has a very big role to play in stability, rigidity and biological activity of the drug molecules [27, 28, 29]. Other important features included in our study include radical electrons, periodic table position, halogen, metal, heteroatom indicators that makes up the chemical classification, chirality, and summation of the bond order. There properties are important for the characterisation of the atoms which is important for accurately predicting the property of the atoms.

Edge Feature Engineering

Each chemical bond which in our study is a 5-dimensional feature vector described bond properties relevant for investigating the electronic structure and connectivity of the drug molecules. The bond that makes up the edge features noted as

where each feature captures a particular bond property. The present bond properties are single, double, triple, and aromatic bond and this provide information that is fundamental to the strength and electronic configuration. The importance of the aromaticity of the bond is that it tells if the bond is involved in the aromatic that is important in identifying the molecular reactivity and stability [27], while to determine if the bond present is part of a cyclic structure the ring membership that influence the molecular flexibility and constraint conformation is used [24].

To represent the resonance of structures and partial bond character, the numerical order of the bond, a continuous measure of bond strength is used to achieve this. All the bond features play a major role in helping the model determine the strength, and the chemical bond nature that directly impact the reactivity, biological activity, and stability of the drug molecules.

Molecular-Level Descriptors

To create a comprehensive multi scale graph representation that capture global molecular properties, a 36 molecular descriptor that complimented the atomic level features was incorporated into the graph topological structure. The molecular descriptor vector is noted as:

where each component represents specific molecular properties. The relevant properties we used for this study were selected after feature engineering to avoid redundancy in the molecular descriptors features.

The key important features used for this study are lipophilicity (LogP), count of ring, molecular weight, and count of the hydrogen bond donor. This pharmaceutical drug properties are very important for determining the drug absorption, distribution, metabolism and excretion [32, 33, 34]. The electronic and surface properties included in our study is the Van der Waals surface area descriptors and the indices of the electronic state, which makes up the molecular shape and electronic distributions. These properties determine the interactions of molecules with the target biological site, which affect selectivity and binding affinity [34, 35, 36, 37]. Moreso, the features that defines the complexity of the drug molecule structure include rings type counts (saturated, aromatic, and heterocycles) and indicators for their functional group. These features help to determine the motifs of the drug structure in relation to their specific biological activities and their patterns in the molecules that are important pharmacologically [30, 31].

This set of molecular descriptors helps capture the global molecular properties alongside the local atomic feature that aid the model to gather detained information about the structural and molecular features of the drug molecules.

To make sure that molecule descriptor maintains feature quality and consistency across the features, feature normalisation was applied using the z-score standardisation, which make sure that every feature present contribute equally to the model learning process. This transformation normalisation process is denoted as

where μ represent mean value and

I implemented systematic data preprocessing to ensure feature quality and consistency across the dataset. Feature normalization was applied to the molecular descriptors using z-score standardization to ensure all features contribute equally to the learning process. The normalization transformation is expressed as:

z = (x - μ) / σ

where μ is the mean and σ is the standard deviation across the molecular dataset. This standardization process helps prevent molecular features with larger values from having a dominating effect in the model learning process, thereby ensuring neural network training stability.

Model Architecture

In our study we developed three graph neural network architecture and evaluated their learning performance. The models used for this study are the Graph Convolutional Network (GCN), to capture long range dependencies of the model during learning we used the graph transformer layer model, which makes use of attention mechanism, and finally GAT-Transformer model that interchange between capturing of local and global dependency during model learning process by applying the attention mechanism. These dependencies are the cheminformatics descriptors which consist of the local chemical environment and global features of the molecules that is important for predicting biological activities [37, 16].

Graph Convolution Neural Network (GCN)

Our model DrugActivityGNN) architecture is built on graph convolutional network, using a undirected graphs where denotes atom, denotes the edges that is a covalent bond. Individual atom in in the molecule is made up features of a 27-dimensional vector that is encoding atomic number, formal charge, aromaticity, hybridization, hydrogen count, and topological patterns that are local. The bonds are described by a vector of 5-dimension denoted as , this captures the order of the bond, stereochemistry, conjugation, and aromaticity.

Using the message-passing scheme, features propagation is governed with the updates for the nodes at layer denoted by:

Where represent the neighbours of the immediate atoms , and represents the weight of the trainable matrix.

To capture the features from attributes of the atom to substructures that are complex, the model hierarchically uses convolutional layers that has a hidden dimension.

The representation of the final molecule is obtained by mean and max pooling operations concatenation:

The mean pooling is the summation of the overall composition of the molecule while the max pooling emphasizes the pharmacophoric distinctive features.

Graph Transformer

To capture long range molecular dependencies within the graph topology, a multi-head attention mechanism is introduced into the model. This architecture is relevant for large molecules were biological activity through biological conjugation or allosteric effect effects distal atoms.

Individual layers make use of:

The computation of each attention head is:

Where helps maintain efficiency and stability. attention heads are used by the model to integrates residual connections and normalisation of the layers:

The feature of the initial atom is projected to the model space using:

To form the final embedding, three pooling strategies that are complementary are use, these strategies are the mean, max, and sum, which are all concatenated as follow:

Our approach ensures that there is robustness across different sizes of the molecule and complexities.

Gat-Transformer Hybrid

To model both global and local molecular dependencies, our study combines graph attention networks and transformer layers.

The graph attention networks compute edge-specific attentions:

With updates that is aggregated as:

The model alternated between the transformer layers (global) and the GAT layers (Local):

By using the learnable attention-based mechanism the final pooling is performed as expressed below:

This process allows the model to adaptively set attention on biological and chemical relevant substructures.

Descriptor Integration and Unified Prediction

To complement the learning process of the graph model, molecular descriptor is integrated into the graph structures. These molecular descriptors capture the physiological characteristics of the drug molecule, such properties are molecular weight, polar surface area, topological indices, electronic properties and lipophilicity.

These descriptors are denoted as:

Where the descriptor input vector is the compressed representation is

The embedding of the molecular descriptors and the graph are concatenated into a vector that is unified:

A three-layer classifier is used to process the representation as expressed below:

Model Training and Evaluation

Our study used binary cross-entropy loss to train the model with a L2 regularization:

To encourage generalization and sparsity of the weight , the additional regularization our study used are for Dropout in all layers, normalization in GNN layers, in the transformer layers we used batch normalization and for all learnable parameter we used weight decay.

The models were trained, validated, and evaluated with the following hyperparameters as presented in table 1.

Table 1: Hyperparameter Tunning for GNN, Transformer-based GNN, and Transformer + GAT

|  |  |  |  |
| --- | --- | --- | --- |
| Hyperparameter | GNN | Transformer-based GNN | Transformer + GAT |
| epochs | 300 | 300 | 300 |
| Initial learning rate | 0.001 | 0.0005 | 0.0005 |
| Weight decay | 1e-4 | 1e-4 | 1e-4 |
| Scheduler factor | 0.5 | 0.5 | 0.5 |
| Scheduler patience | 10 | 10 | 10 |
| Early stop patience | 25 | 25 | 25 |
| Warmup epochs | 5 | 5 | 5 |
| Cosine annealing | False | True | True |
| Dropout schedule | False | False | False |
| Gradient clip value | 1.0 | 1.0 | 1.0 |
| label smoothing | None | 0.1 | 0.1 |

The model was trained and evaluated on an epoch of 300 under the hyperparameter value presented in table 1. The models were evaluated using metrics that are essential for binary classification, these metrics are accuracy, precision, recall, fi-score, and AUC-ROC evaluation metrics.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) serves as the primary model evaluation metrics across the various turned hyperparameter. The ROC curve gives a trade-off between the true positive and the false positive. The selection of AUC-ROC as the main evaluation metrics is because for antimalarial drug discovery process, it helps the model’s ability to differentiate between active and inactive antimalarial drug compounds based on the best hyperparameter settings. The robustness and stability of AUC makes it suitable to be used as an evaluation metrics for this model.

Other metrics used for this study are accuracy [64], precision [64], recall [65], and F1-Score [65]. This metrics also provide assessment of the model predictive ability, specifically across the different model’s performance dimension, this metrics can detect active and inactive antimalarial drug group important to pharmaceutical screening applications.

RESULT

Feature Selection

In our study, we identify and remove any redundant molecular descriptors using a correlation-based feature selection approach. This is important to avoid multicollinearity and reduce dimensionality that could hinder a perfect model performance and interpretability. This feature selection process was significant for optimizing the space of the molecular descriptor and keep important chemical information that is needed for the prediction of antimalarial activity.

A screenshot of a computer screen

AI-generated content may be incorrect. A screenshot of a graph

AI-generated content may be incorrect.

Figure 1: Active Antimalarial Drug Discovery from Correlation Based Feature Selection. Molecular descriptor correlation comparative analysis of highly correlated and non-highly correlated features. Figure 1a shows highly correlated features with r values greater or equal to 0.8, indicating reductant molecular descriptors. Figure 1b shows non-redundant molecular descriptors indicating reduction in multicollinearity while preserving important chemical diversity relevant for this work.

Model Training Accuracy Performance for the prediction of Active Antimalaria Drug

A graph with orange and blue lines

AI-generated content may be incorrect.

a.

A graph with blue and orange lines

AI-generated content may be incorrect.

b.

A graph of a graph

AI-generated content may be incorrect.

c.

Figure 2. Accuracy curves for model training and validation performance in Active antimalaria drug prediction. Training dynamics of the neural network’s architecture used in our study. a. Graph Neural Networks (GNN) base model b. Graph Transformer Layered Model. c. Hybrid of Graph Attention and Transformer Layered Model. The representation of the training and validation accuracy used the blue and orange line respectively. Making use of early stop criteria, the green line point base shows the optimal stop point base.

Model Training Performance using Area Under the Curve Receiver Operating Characteristics (AUC-ROC) metrics for the prediction of Active Antimalaria Drug

A graph with blue and orange lines

AI-generated content may be incorrect.

a.

A graph with blue and orange lines

AI-generated content may be incorrect.

b.

A graph with lines and numbers

AI-generated content may be incorrect.

c.

Figure 3. Area Under the Curve Receiver Operating Characteristics (AUC-ROC) for model training and validation performance in Active antimalaria drug prediction. Training and validation dynamics architecture performance. a. Graph Neural Networks (GNN) base model b. Graph Transformer Layered Model. c. Hybrid of Graph Attention and Transformer Layered Model. The representation of the training and validation AUC-ROC used the blue and orange line respectively. Making use of early stop criteria, the green line point base shows the optimal stop point base.

Learning Rate Optimisation and Scheduling for the prediction of Active Antimalaria Drug

A graph with a line going up

AI-generated content may be incorrect.

a.

A graph with a curved line

AI-generated content may be incorrect.

b.

A graph with a blue line

AI-generated content may be incorrect.

c.

Figure 4. Learning Rate Decay Schedule a. Graph Neural Networks (GNN) base model Learning Rate Decay Schedule b. Graph Transformer Layered Model Learning Rate Decay Schedule. c. Hybrid of Graph Attention and Transformer Layered Model Decay Schedule. Making use of early stop criteria, the green line point base shows the optimal stop point base.

Confusion Matrix Classification Performance for the prediction of Active Antimalaria Drug

A diagram of a blue box

AI-generated content may be incorrect.

a

A diagram of a diagram

AI-generated content may be incorrect.

b.

A diagram of a diagram

AI-generated content may be incorrect.

c

Figure 5. Confusion matrix classification performance for the antimalaria active drug detection on the test set. a. Graph Neural Networks (GNN) base model Learning confusion matrix performance b. Graph Transformer Layered Model Learning confusion matrix performance. c. Hybrid of Graph Attention and Transformer Layered Model confusion matrix performance. Making use of early stop criteria, the green line point base shows the optimal stop point base. The matrices showed True Vs Predicted labels with the value of cells indicating the smiles molecular count. The hight count is marked by dark blue colour.

A graph with a line

AI-generated content may be incorrect.

A graph of a function

AI-generated content may be incorrect.

A graph with orange line

AI-generated content may be incorrect.

Figure 6. Receiver Operating Characteristic (ROC) Curves for performance evaluation for the antimalaria active drug detection. a. Graph Neural Networks (GNN) base model ROC curve performance evaluation (AUC = 0.947). b. Graph Transformer Layered Model ROC curve performance evaluation (AUC = 0.947). c. Hybrid of Graph Attention and Transformer Layered Model ROC curve performance evaluation (AUC = 0.939). The random classify (AUC = 0.5) is showed by the Gray dashed line, while the model performance is indicated by the orange curve.

Table 2: Evaluation Metrics Performance for GNN, Transformer-based GNN, and Transformer + GAT

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | GNN | Transformer-based GNN | Transformer + GAT |
| Accuracy | 0.8807 | 0.8812 | 0.8744 |
| Precision | 0.8717 | 0.8628 | 0.8685 |
| Recall | 0.8911 | 0.9048 | 0.8805 |
| F1-Score | 0.8813 | 0.8833 | 0.8745 |
| AUC | 0.9471 | 0.9473 | 0.9389 |

A graph with red and blue dots

AI-generated content may be incorrect.

A graph with numbers and text

AI-generated content may be incorrect.

Figure 7. SHAP Explainability a. Node features b. Molecular Descriptors

A graph with numbers and a number of numbers

AI-generated content may be incorrect.

A graph with red bars

AI-generated content may be incorrect.

Figure 8. Captum Feature Ablation Explainability a. Node features b. Molecular Descriptors

A graph with red bars

AI-generated content may be incorrect.

A bar graph with red and white text

AI-generated content may be incorrect.

Figure 9. Captum Saliency Explainability a. Node features b. Molecular Descriptors

A graph of a number of different colored bars

AI-generated content may be incorrect.

A graph of a number of colored bars

AI-generated content may be incorrect.

Figure 10. Captum Integrated Gradients Explainability a. Node features b. Molecular Descriptors

Discussion

This comprehensive evaluation of deep learning architectures for antimalarial activity prediction provides significant insights into the application of modern neural network methodologies to structure-activity relationship modelling. In our study, the transformer architecture demonstrated superior performance characteristics that position it as the optimal solution for computational screening applications. The bedrock of this outstanding result is in the methodological approach applied during the feature selection process, where we implemented the feature selection using the correlation method. During this process, features that were highly correlated were eliminated; this is to avoid redundancy and multicollinearity in the molecular descriptors. The eliminated descriptor clusters that are highly correlated include descriptors for the molecular shape, distribution parameters for the molecule charge, and topological connectivity indices (Figures 1a and 1b). This is to ensure that the model predictive ability operates based on molecular properties that are independent and informative.

The process applied in the optimisation of the features is important to the transformer architecture model, and this aids the distribution of the attention weight effectively between specific properties of the molecules. The selected features capture key chemical space properties that are relevant in antimalaria drug discovery research. These features include LogP and the weight of the molecules. Additional features include the hydrogen bond ring and count of the rings, pharmacophore-relevant fragment, and electronic properties and topological indices (Figure 1b). Across various chemical scaffolds, this diverse essential feature is significant in validating and predicting active antimalarial drugs.

Across the various models used in the prediction of active antimalarial drugs, the transformer-based GNN architecture showed excellent results across all evaluation parameters used in the current study. The transformer architecture showed training and validation accuracies that converge well at 88.70% with the AUC-ROC value of 0.947, while the transformer model supported with a graph attention converges at 85.00% with an AUC-ROC score of 0.939, and the pure GNN architecture model has the same training and validation accuracy convergence value (figure 2) as that of the transformer architecture; however, the training and validation curve shows irregularities during the training and validation process. The transformer model architecture showed a close, perfect alignment in the training and validation curves during the training and validation process (Figure 2b and Figure 3b), showing a dynamic optimisation that is robust and a regularisation ability that is exceptional. This is in contrast with the pure GNN and GAT optimised architectures, respectively, that showed a slight overfitting during training and validation of the models.

Most importantly, in our study on active antimalarial drug discovery, the transformer-based GNN model had the best sensitivity score of 0.9048, where 1711 (Table 2, figure 5b) were correctly predicted as active drug compounds out of the 1,891 active compounds. In pharmaceutical computational screening research, this high value of sensitivity performance has significant importance in proving the model performance in predicting the active compound, where the opportunity cost is the missing potential active antimalarial drug. This missing potential active antimalarial drug, representing the false negative rate of 9.5%, provides a high confidence of the transformer-based GNN model in predicting positive active antimalarial drugs, indicating that the transformer-based GNN model has better identification performance for antimalarial drug discovery. Across all the thresholds for the various model classifications, as shown in the ROC-curve analysis, there are exceptional characteristics in the transformer-based GNN model performance that are revealed, with a rapid increase in the value of the true positive rate of up to 90% with a false positive rate of approximately 12%, showing a specific strength in high-sensitivity applications (Figure 5 and Figure 6).

The better performance of the transformer-based GNN, being the best against the pure GNN and the GAT optimised model, is due to the advantages the transformer layers get from its fundamental architecture, which is well designed to use the property of the antimalarial drug molecule in making its prediction. The model uses the mechanism of multi-head self-attention to simultaneously evaluate the molecular descriptors by learning feature interactions that are complex and relevant for structural activity relationship capturing. While the parallel attention computations aid the molecular properties modelling, ranging from physicochemical properties that are simple to complex pattern pharmacophores.

The transformer-based GNN in our current study has a gradient flow that is stable throughout the training process. This is possible because of the residual connections in the transformer layer and normalisation of the layers that are effectively optimised in the complex parameter space that is needed for accurate model relationships in the molecular structure activities. The learning rate schedule also contributes to this, which transitions from 10⁻⁴ to 10⁻⁵ (figure 4b) over a training epoch of 300, proving optimisation in the attention weight that is optimal, which does not interfere with the delicate balance that exists in the attention weight.

Even though the pure GNN model achieved an AUC-ROC value of 0.947, which is the same as that of the transformer-based model but different from the GAT-based GNN model (figure 2), different factors have placed it as a suboptimal model for the prediction of antimalarial active drug prediction in our study when compared to the transformer-based model. This may be because of possible overfitting as seen in the accuracy gap between the training and validation of 1.3% (Figure 2a); this is in contrast with the transformer-based GNN model that has an accuracy gap of almost zero. Additionally, the learning rate schedule, even though a rapid convergence is seen at an epoch of 85, in later training phases showed limited fine-tuning capabilities. The pure GNN model's total reliance on the graph structural representation that is explicit may have hindered the model's ability to explore the molecular descriptor space that contains the rich information needed for making the antimalarial drug discovery.

The GAT-optimised transformer-based GNN model showed the least performance value across all model evaluation metrics with an accuracy score of 86.5%, an AUC-ROC of 0.939, and a sensitivity score of 88.1%. Despite the theoretical potential that combining two complementary layers may have to improve model performance, the GAT-optimised transformer model performed worse compared to the transformer-based GNN and pure GNN model, which may be attributed to the fact that the GAT layer might make the GAT-optimised transformer-based GNN model very complex. With this model's architectural complexity, the performance of the GAT-optimised transformer-based GNN model is inferior compared to other architectures used in our study. This architectural complexity from the combination of the graph attentions and transformer-based GNN model created a challenge in the model optimisation process that outweighed the potential benefit from his synergistic relationship. Also, the poor performance of the GAT-optimised transformer-based GNN model may be attributed to the model learning rate, though having a regularisation property that is excellent, as seen in its training validation gap of almost 0.3%, may appear to be responsible for limiting its full potential to explore the molecular feature space's discriminative potential.

The learning rate optimisation and scheduling showed the optimisation strength and limitation of each model’s architecture used in the current study. Out of all the models used in the current study, only the transformer-based GNN architecture exceptionally showed a stable training process that is dynamic, as revealed by its smooth convergence with a fluctuation between training and validation performance that is very minimal (figures 4a, b, and c). These not only showed the efficiency of the transformer-based model in predicting the antimalarial active drug but also revealed optimisation that is superior to other models used in the current study. This stability during the transformer-based model aids long training periods without risk of overfitting, thereby giving room for exploring the parameter spaces for structure-activity relationships.

Explainability of the Transformer based GNN model

The transformer based GNN model performance explainability was performed using SHapley Addictive exPlanation (SHAP) and Captum with focus on integrated gradients, feature ablation, and saliency. The node features and molecular descriptors were investigated to know which features have high impact on the model’s decision.

Model Explainability with SHAP

The node features were evaluated to note the feature that had the most influence in the model-making decision. The SHAP analysis for node features as shown in figure 7a reveals that the hydrogen count value in the antimalarial drug molecule has a major role to play in the antimalarial drug prediction. In the SHAP, this hydrogen count contribution value ranges from low negative to highly positive value, indicating that hydrogen atom counts in the antimalaria drug study have a significant role to play in determining the transformer-based GNN model decision in predicting active or inactive antimalarial drugs from SMILES molecules (Figure 7a). The relationship between hydrogen bonding and complex biological activities has been reported by several studies, where the scaffold of molecules and binding targets depends heavily on the optimal profile of donor and acceptor. This is in consonance with the study of Wolf et al. [38] and Srinivasa et al. [39]; they revealed that the formation of hydrogen bonds for the foundation of quinoline-based antimalarial drug mechanisms prevents the detoxification of the parasite, thereby promoting the death of the malaria parasite.

Similarly, the aromaticity feature importance plays a high impact for the transformer-based GNN model to make its prediction of active antimalaria drugs (figure 7a). The aromaticity can affect the model decision negatively or positively, as shown in Figure 7a, and this decision depends on the molecular positioning with the antimalarial drug scaffold. This supports the claim of Tran et al. [40] and Valsecchi et al. [41] that the significant contribution to intricate interaction between molecular solubility and the surrounding chemical environment is the aromatic rings, which the transformer-based GNN model has the ability to effectively capture; thus, improving the accuracy of the model prediction. The positive influence in the transformer-based GNN prediction may be due to the stacking of π-π interaction with the aromatic residues of amino acids in the protein of the antimalarial, while the negative influence may be in the instance where metabolic liabilities are introduced, thus comprising aromaticity selectiveness. For instance, protein complexes that have a π-π interaction stacking have been reported to contribute significantly to stabilising interaction between drug and protein [42, 43].

The implicit valence of the antimalaria drug molecule also serves as an important feature the transformer-based model considered in making the prediction, as shown in Figure 7a. The SHAP explainability framework used showed a positive feature that is concentrated, which reaffirms the importance of the antimalarial drug electronic environment and its activity, where electronic hybridisation states features and lone pair availability directly impacting the molecular recognisable events that take place in the active drug prediction from the antimalarial drug prediction as supported by the study of Noviandy et al. [7], which revealed in machine learning analysis using SHAP that antimalarial drugs involving electronic descriptor features enhance drug activity prediction, design and model interpretability. Additionally, SHAP also performs the explainability of the transformer-based GNN model based on the molecular descriptors of the SMILES molecules. The top molecular descriptor feature that played a major impact in the prediction of the model is the PEOE\_VSA1, as shown in figure 7b. This molecular descriptor captures the Van der Waals surface area that is weighted by the partial atomic charges, revealing that specific electrostatic surface patterns consistently improve the efficiency of antimalarial drugs. This supports the claim by Grassmann et al. [44] that in ligand-receptor pairs, specifically in improving binding specificity and affinity towards malaria targets, electrostatic complementarity plays a major role, which is in conjunction with the shape of the drug molecule that is usually considered during antimalarial drug design. According to Siqueira-Neto et al. [45], initiatives in the chemistry of the antimalarial drug structure have triggered the optimisation of surface properties of the molecules to improve therapeutic efficacy and reduce antimalarial drug resistance.

### Model explainability with Captum Feature Ablation

Figure 8a shows the captum feature ablation presentation for the top atomic features that have much influence in the transformer-based GNN model making its decision. The explicit hydrogen count showed the most influential features (Figure 8a), which is similar to that of SHAP (Figure 7a). Other atomic features that showed important impact in the model’s prediction performance are the halogen and ring membership features. Halogen atom features contributing to the model performance imply that they have a significant role to play in the active antimalaria drug outcome of the model. This is in consonance with studies of Dudek et al. [46] and Paulikat et al. [47], which showed that for the enhancement of membrane permeability and stability of drug metabolism, halogenation of drug molecules is important. Additionally, it is reported that present in the scaffold of antimalarial drugs are halogen substituents, which is due to their modulatory pharmacokinetic properties and biological activity [48, 49, 50]. The ring membership feature contributes to heterocyclic and aromatic motif contributions, which are normally seen in active antimalarial compounds [51]. Previous studies have emphasised that the heterocyclic rings serve as one of the hallmarks for clinically discovered antimalarial drugs, which have provided a strong and excellent interaction with biological antimalarial drug targets [52, 53].

For molecular descriptors, the highest features for the feature ablation analysis are the Electro-topological State Van der Waals Area (EState\_VSA) feature, as seen in Figure 8b. The EState\_VSA is followed by Partial Equalisation of Orbital Electronegativities Van der Waals Surface Area (PEOE\_VSA), as shown in figure 8b, which combines electronic atomic states or partial charges with their Van der Waals surface charge to capture important molecular environments that impart biological activities. These findings are consistent with the previous studies in computational antimalaria drug research where molecular descriptors that involve electronic-level and surface atomic properties played a key role in model predictive ability and performance [8, 14]. This surface charge molecular descriptors feature is supported by the work of Mswahili et al. [14]; they stated that in their work, where they predicted antimalaria drug bioactivities using a machine learning model, their analysis of feature importance revealed that the top features that influenced the predictive ability of the model were features that captured spatial and electronic atom-level properties, thus supporting the observed Estate\_VSA-like features in our study.

**Model explainability with Captum Saliency**

The saliency analysis for the feature importance for node and molecular descriptor features are presented in Figures 9a and 9b, respectively. The most important node feature that the transformer model considered in making his prediction is the atomic number of the antimalarial drug molecule, followed by implicant valence (Figure 9a). This atomic number and implicant valence of the antimalarial drug molecule show that heteroatom identity is an important factor that impacts the predictive ability of the transformer base model used in this study. Therefore, any modification in the heteroatomic features may possibly cause significant changes to the antimalaria drug predictive decision of the transformer-based GNN model, thus emphasising the importance of heteroatom optimisation in the antimalaria drug discovery process.

Other top atomic features that have a high impact on the model’s decision are the implicit valence and total degree (figure 9a). These atomic features show that the electronic environment and connectivity of the antimalarial drug molecules are important features the model depends on for accurately making predictions for the antimalarial drug. These findings provide important insight that shows that antimalarial drug connectivity and its electronic states play an important role in determining the bioactivity of antimalarial drugs [14, 54].

The molecular descriptor analysis using saliency to unravel the most important features that impact the transformer-based GNN model decision in predicting the antimalarial drug shows that the aromatic amine fragments (fr\_Ar\_NH) serve as the most important features, thereby serving as the most critical features considered by the model in making its prediction. In the design of antimalaria drugs, many studies have reported that antimalarials often have an aromatic nitrogen and amine functionality that is associated with important bioactivity and drug-like properties. Our findings support the study of Tiwari et al. [55] on 4-aminoquinoline derivative, where they report that amino group modification of antimalarial drugs can significantly improve the antimalarial drug effect against malaria-resistant strains. Therefore, this careful optimisation of aromatic amine in the antimalaria drug study is strategic in the development of an active and effective antimalaria drug.

Model explainability with Captum Integrated Gradients

The integrated gradients analysis of feature importance revealed that the most important feature that influences the decision of the transformer-based model GNN is the total hydrogen count as presented in figure 10a, which is aligned with the analysis from SHAP and feature ablation (figure 7a and figure 8a). This finding further confirms that one of the determinants of the bioactivity of antimalarial drugs is the overall hydrogen content of the antimalarial drug, which supports the principle of hydrogen bonding as stated by Jeffrey and Saenger [56] and Herschlag and Pinney [57]. The transformer-based GNN model also considered the number of implicit hydrogen counts, aromaticity, and implicit valence in making its predictive decision. Their features are important for key relevant electronic events that happen between the antimalaria drug molecules.

The molecular descriptors that contribute to the transformer-based GNN model’s decision are the thiazole and the thiophene descriptors (Figure 10b). In antimalaria drug design, the thiazole systems are well-stabilised structures that provide a vast number of interaction modalities, which include π-π stacking, bonds that involve hydrogen, and essential heme-target and parasite enzyme meta-coordination [58, 59]. This aligns with the hydrogen count feature's importance discovered by the SHAP, Captum feature ablation, and Saliency. The successful contribution to the antimalaria scaffold is the rise of thiazole fragments, which are present in numerous validated antimalaria drugs represented in pharmacore-validated recognisable drug-target events [59, 60, 61]. The thiophene fragment features (fr\_thiophene) show the importance of aromatic heterocycles in sulphur-containing antimalaria drugs in the antimalaria drug prediction. The thiophene features in the antimalaria target provide unique geometric constraints and electronic properties that optimise antimalaria targets' molecular recognition [62, 63].

Conclusion

The study demonstrated a successful application of a GNN-based model to investigate the prediction of active and inactive antimalarial drugs from their SMILES data. Out of the various GNN models used, the transformer-based GNN model achieves a better predictive ability than the other models used. This model showed better performance metrics in all evaluation metrics, with a training and validation accuracy of 88.70% with an AUC-ROC value of 0.947, with a more stable and convergent training and validation curve compared to the pure GNN and GAT optimised transformer-based GNN model, which showed overfitting and an unstable training and validation curve. The current study also revealed the features that contributed more to the transformer-based model to make its predictive decision. This was achieved using the SHAP and Captum frameworks that include feature ablation, saliency, and integrated gradients, which identified the node features in the GNN structure and the molecular descriptors that played a major role in the model transformer-based GNN model architecture predictive ability process for the active antimalarial drug. The excellent performance of the transformer-based model over the other architecture can be attributed to the self-attention mechanism that captures long-range molecular dependencies and the interaction of complex features. The layer normalisation and residual connections help facilitate the stable gradient flow, which is important for structure relationship modelling that is accurate.

The interpretability for the model using SHAP and Captum showed that for the node feature important hydrogen count served as the most important atomic feature the transformer-based model considered before making its prediction for the antimalarial drug, which reflectsthe role hydrogen bonding plays in the interaction between drug and target. Aromaticity features also played a major role in the model’s prediction ability of the antimalarial drug, and these are followed by electronic descriptors like the implicit valence and atomic number that are significant in determining biological activity through its electronic environment. The top molecular descriptor that has the highest impact on the transformer-based GNN models’ decision in predicting the antimalaria drug is partial atomic charges (PEOE\_VSA1) and electro-topological state features (EState\_VSA). These features are surface descriptors that play a major role in antimalarial drug design efficacy through their interaction with molecular surface properties and electrostatic complementarity.

In antimalarial drug design, our current study addresses the need to improve and design a drug discovery computational tool that is reliable for an in silico antimalarial drug design process, particularly in this area of global rise in malarial drug-resistant parasites. The interpretable approach we have implemented into this study helps to provide insight to antimalarial drug design researchers or chemists on atomic and molecular features that determine if an antimalarial drug will be active and the reason they are active based on their physicochemical and structural properties.

Limitations

This study has several limitations. First, due to the imbalance between active and inactive molecules, under-sampling of the majority class was applied, which may have led to the loss of potentially useful chemical information. Second, the dataset may be relatively small or lack sufficient chemical diversity, limiting the model’s ability to generalize across broader chemical space. Third, the absence of wet-lab validation means the computational predictions cannot yet be confirmed in real-world antimalarial activity. Finally, while SHAP and the Captum framework were employed for interpretability, the dataset limitations and lack of experimental confirmation constrain the depth and reliability of the explanations.

Future Work

Future studies will address these limitations by (i) using larger, more chemically diverse, and higher-quality datasets to minimize information loss and improve model generalization, (ii) incorporating multimodal data sources such as omics datasets, structural biology data, and phenotypic screening to provide a more holistic representation of drug activity, (iii) exploring additional interpretability techniques such as LIME, counterfactual explanations, and attention visualization, and (iv) validating computational predictions with wet-lab experiments to confirm their real-world efficacy in malaria drug discovery.

REFERENCES

1. <https://www.who.int/news-room/fact-sheets/detail/malaria>
2. White, N.J., 2002. The assessment of antimalarial drug efficacy. *Trends in parasitology*, *18*(10), pp.458-464.
3. Neves, B.J., Braga, R.C., Melo-Filho, C.C., Moreira-Filho, J.T., Muratov, E.N. and Andrade, C.H., 2018. QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology*, *9*, p.1275.
4. Tse, E.G., Aithani, L., Anderson, M., Cardoso-Silva, J., Cincilla, G., Conduit, G.J., Galushka, M., Guan, D., Hallyburton, I., Irwin, B.W. and Kirk, K., 2021. An open drug discovery competition: experimental validation of predictive models in a series of novel antimalarials. *Journal of Medicinal Chemistry*, *64*(22), pp.16450-16463.
5. Thorne, N., Auld, D.S. and Inglese, J., 2010. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Current opinion in chemical biology*, *14*(3), pp.315-324.
6. Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R. and Consonni, V., 2014. QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, *57*(12), pp.4977-5010.
7. Noviandy, T.R., Idroes, G.M. and Hardi, I., 2024. An interpretable machine learning strategy for antimalarial drug discovery with lightgbm and shap. *Journal of Future Artificial Intelligence and Technologies*, *1*(2), pp.84-95.
8. Mswahili, M.E., Ndomba, G.E., Jo, K. and Jeong, Y.S., 2024. Graph neural network and bert model for antimalarial drug predictions using plasmodium potential targets. *Applied Sciences*, *14*(4), p.1472.
9. Danishuddin, Madhukar, G., Malik, M.Z. and Subbarao, N., 2019. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR and QSAR in Environmental Research*, *30*(8), pp.543-560.
10. Bharti, D.R. and Lynn, A.M., 2017. QSAR based predictive modeling for anti-malarial molecules. *Bioinformation*, *13*(5), p.154.
11. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5. Link: <https://www.jmlr.org/papers/volume18/16-365/16-365.pdf>
12. Ali, M., 2025. Classification of imbalanced travel mode choice dataset with SMOTE and prediction using interpretable machine learning. Sustainable Futures, 10, p.101119.
13. Ileberi, E., Sun, Y. and Wang, Z., 2021. Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. IEEE access, 9, pp.165286-165294.
14. Mswahili, M.E., Martin, G.L., Woo, J., Choi, G.J. and Jeong, Y.S., 2021. Antimalarial drug predictions using molecular descriptors and machine learning against Plasmodium falciparum. Biomolecules, 11(12), p.1750.
15. Van Heerden, A., Turon, G., Duran-Frigola, M., Pillay, N. and Birkholtz, L.M., 2023. Machine learning approaches identify chemical features for stage-specific antimalarial compounds. ACS omega, 8(46), pp.43813-43826.
16. Alam, S.M., Nigar, N., Rasheed, M.W. and Amin, L., 2024. Analyzing the role of reducible molecular descriptors and thermodynamic aspects of anti-tuberculosis drugs via QSPR study. Frontiers in Physics, 12, p.1383216.
17. McGibbon, M., Shave, S., Dong, J., Gao, Y., Houston, D.R., Xie, J., Yang, Y., Schwaller, P. and Blay, V., 2024. From intuition to AI: evolution of small molecule representations in drug discovery. Briefings in bioinformatics, 25(1), p.bbad422.
18. Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F.J., Carballal, A., Maojo, V., Pazos, A. and Fernandez-Lozano, C., 2021. A review on machine learning approaches and trends in drug discovery. Computational and structural biotechnology journal, 19, pp.4538-4558.
19. Orosz, Á., Héberger, K. and Rácz, A., 2022. Comparison of descriptor-and fingerprint sets in machine learning models for ADME-Tox targets. Frontiers in Chemistry, 10, p.852893.
20. Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F.J., Carballal, A., Maojo, V., Pazos, A. and Fernandez-Lozano, C., 2021. A review on machine learning approaches and trends in drug discovery. Computational and structural biotechnology journal, 19, pp.4538-4558.
21. Zhao, B., Xu, W., Guan, J. and Zhou, S., 2024. Molecular property prediction based on graph structure learning. Bioinformatics, 40(5), p.btae304.
22. Han, S., Fu, H., Wu, Y., Zhao, G., Song, Z., Huang, F., Zhang, Z., Liu, S. and Zhang, W., 2023. HimGNN: a novel hierarchical molecular graph representation learning framework for property prediction. Briefings in Bioinformatics, 24(5).
23. Gallegos, M., Barrena-Espés, D., Guevara-Vela, J.M., Rocha-Rinza, T. and Pendás, Á.M., 2022. A QCT view of the interplay between hydrogen bonds and aromaticity in small CHON derivatives. Molecules, 27(18), p.6039.
24. Polêto, M.D., Rusu, V.H., Grisci, B.I., Dorn, M., Lins, R.D. and Verli, H., 2018. Aromatic rings commonly used in medicinal chemistry: force fields comparison and interactions with water toward the design of new chemical entities. Frontiers in pharmacology, 9, p.395.
25. Hong, H., Wu, X., Sun, H., Xie, C., Wang, Q. and Li, Y., 2025. Learning Hierarchical Interaction for Accurate Molecular Property Prediction. arXiv preprint arXiv:2504.20127.
26. Glendening, E.D. and Weinhold, F., 2021. Pauling’s conceptions of hybridization and resonance in modern quantum chemistry. Molecules, 26(14), p.4110.
27. Jackson, S.N., Barbacci, D.C., Bonci, A. and Woods, A.S., 2019. An in vitro study of aromatic stacking of drug molecules. Journal of The American Society for Mass Spectrometry, 30(7), pp.1199-1203.
28. Shearer, J., Castro, J.L., Lawson, A.D., MacCoss, M. and Taylor, R.D., 2022. Rings in clinical trials and drugs: present and future. Journal of medicinal chemistry, 65(13), pp.8699-8712.
29. Uthumange, S.S., Liew, A.J.H., Chee, X.W. and Yeong, K.Y., 2024. Ringing medicinal chemistry: The importance of 3-membered rings in drug discovery. Bioorganic & Medicinal Chemistry, 116, p.117980.
30. Shearer, J., Castro, J.L., Lawson, A.D., MacCoss, M. and Taylor, R.D., 2022. Rings in clinical trials and drugs: present and future. Journal of medicinal chemistry, 65(13), pp.8699-8712.
31. Mao, F., Ni, W., Xu, X., Wang, H., Wang, J., Ji, M. and Li, J., 2016. Chemical structure-related drug-like criteria of global approved drugs. Molecules, 21(1), p.75.
32. Menestrina, L., Parrondo-Pizarro, R., Gómez, I., Garcia-Serna, R., Boyer, S. and Mestres, J., 2025. Refined ADME Profiles for ATC Drug Classes. Pharmaceutics, 17(3), p.308.
33. Leeson, P.D., Bento, A.P., Gaulton, A., Hersey, A., Manners, E.J., Radoux, C.J. and Leach, A.R., 2021. Target-based evaluation of “drug-like” properties and ligand efficiencies. Journal of medicinal chemistry, 64(11), pp.7210-7230.
34. Daina, A., Michielin, O. and Zoete, V., 2017. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Scientific reports, 7(1), p.42717.
35. Rana, M.M. and Nguyen, D.D., 2022. EISA-Score: Element Interactive Surface Area score for protein–ligand binding affinity prediction. Journal of Chemical Information and Modeling, 62(18), pp.4329-4341.
36. Xie, D., Tian, Y., Cao, L., Guo, P., Cai, Z. and Zhou, J., 2025. 3D-QSAR, design, molecular docking and dynamics simulation studies of novel 6-hydroxybenzothiazole-2-carboxamides as potentially potent and selective monoamine oxidase B inhibitors. Frontiers in Pharmacology, 16, p.1545791.
37. Nguyen, H.T., Yoshinouchi, Y., Hirano, M., Nomiyama, K., Nakata, H., Kim, E.Y. and Iwata, H., 2023. In silico simulations and molecular descriptors to predict in vitro transactivation potencies of Baikal seal estrogen receptors by environmental contaminants. Ecotoxicology and Environmental Safety, 265, p.115495.
38. Wolf, S., Domes, R., Domes, C. and Frosch, T., 2024. Spectrally resolved and highly parallelized Raman difference spectroscopy for the analysis of drug–target interactions between the antimalarial drug chloroquine and hematin. Analytical Chemistry, 96(8), pp.3345-3353.
39. Srinivasa, S.B., Ullal, S.N. and Kalal, B.S., 2025. Quinoline conjugates for enhanced antimalarial activity: a review on synthesis by molecular hybridization and structure-activity relationship (SAR) investigation. American Journal of Translational Research, 17(2), p.1335..
40. Tran, T. and Ekenna, C., 2023. Molecular descriptors property prediction using transformer-based approach. International Journal of Molecular Sciences, 24(15), p.11948.
41. Valsecchi, C., Arjona-Medina, J.A., Dyubankova, N. and Nugmanov, R., 2025. Benchmarking molecular conformer augmentation with context-enriched training: graph-based transformer versus GNN models. Journal of Cheminformatics, 17(1), p.80.
42. Shao, J., Kuiper, B.P., Thunnissen, A.M.W., Cool, R.H., Zhou, L., Huang, C., Dijkstra, B.W. and Broos, J., 2022. The role of tryptophan in π interactions in proteins: An experimental approach. Journal of the American Chemical Society, 144(30), pp.13815-13822.
43. Sakata, Y., Yabunaka, K., Kobayashi, Y., Omiya, H., Umezawa, N., Kim, H.S., Wataya, Y., Tomita, Y., Hisamatsu, Y., Kato, N. and Yagi, H., 2018. Potent antimalarial activity of two arenes linked with triamine designed to have multiple interactions with heme. ACS Medicinal Chemistry Letters, 9(10), pp.980-985.
44. Grassmann, G., Di Rienzo, L., Gosti, G., Leonetti, M., Ruocco, G., Miotto, M. and Milanetti, E., 2023. Electrostatic complementarity at the interface drives transient protein-protein interactions. Scientific Reports, 13(1), p.10207.
45. Siqueira-Neto, J.L., Wicht, K.J., Chibale, K., Burrows, J.N., Fidock, D.A. and Winzeler, E.A., 2023. Antimalarial drug discovery: progress and approaches. Nature Reviews Drug Discovery, 22(10), pp.807-826.
46. Dudek, A., Szulc, N., Pawlak, A., Strugała-Danak, P., Krawczyk-Łebek, A., Perz, M., Kostrzewa-Susłow, E. and Pruchnik, H., 2024. Structural investigation of interactions between halogenated flavonoids and the lipid membrane along with their role as cytotoxic agents. Scientific Reports, 14(1), p.10561.
47. Paulikat, M., Piccini, G., Ippoliti, E., Rossetti, G., Arnesano, F. and Carloni, P., 2023. Physical chemistry of chloroquine permeation through the cell membrane with atomistic detail. Journal of Chemical Information and Modeling, 63(22), pp.7124-7132.
48. Charoensutthivarakul, S., Lohawittayanan, D., Kanjanasirirat, P., Jearawuttanakul, K., Seemakhan, S., Chabang, N., Schlaeppi, P., Tantivess, V., Limboonreung, T. and Phanchana, M., 2023. Rational Design and Lead Optimisation of Potent Antimalarial Quinazolinediones and Their Cytotoxicity against MCF-7. Molecules, 28(7), p.2999.
49. Smith, H.S., Giuliani, B., Wijesekera, K., Lum, K.Y., Duffy, S., Lock, A., White, J.M., Avery, V.M. and Davis, R.A., 2025. Investigations of amination reactions on an antimalarial 1, 2, 4-triazolo [4, 3-a] pyrazine scaffold. Beilstein Journal of Organic Chemistry, 21(1), pp.1126-1134.
50. Charoensutthivarakul, S., Lohawittayanan, D., Kanjanasirirat, P., Jearawuttanakul, K., Seemakhan, S., Chabang, N., Schlaeppi, P., Tantivess, V., Limboonreung, T. and Phanchana, M., 2023. Rational Design and Lead Optimisation of Potent Antimalarial Quinazolinediones and Their Cytotoxicity against MCF-7. Molecules, 28(7), p.2999.
51. Parhizgar, A.R. and Tahghighi, A., 2017. Introducing new antimalarial analogues of chloroquine and amodiaquine: a narrative review. Iranian journal of medical sciences, 42(2), p.115.
52. Orozco, M.I., Moreno, P., Guevara, M., Abonia, R., Quiroga, J., Insuasty, B., Barreto, M., Burbano, M.E. and Crespo-Ortiz, M.D.P., 2024. In silico prediction and in vitro assessment of novel heterocyclics with antimalarial activity. Parasitology Research, 123(1), p.75.
53. Kalaria, P.N., Karad, S.C. and Raval, D.K., 2018. A review on diverse heterocyclic compounds as the privileged scaffolds in antimalarial drug discovery. European journal of medicinal chemistry, 158, pp.917-936.
54. Bhanot, A. and Sundriyal, S., 2021. Physicochemical profiling and comparison of research antiplasmodials and advanced stage antimalarials with oral drugs. ACS omega, 6(9), pp.6424-6437.
55. Tiwari, V.S., Joshi, P., Yadav, K., Sharma, A., Chowdhury, S., Manhas, A., Kumar, N., Tripathi, R. and Haq, W., 2021. Synthesis and antimalarial activity of 4-methylaminoquinoline compounds against drug-resistant parasite. ACS omega, 6(20), pp.12984-12994.
56. Jeffrey, G.A. and Saenger, W., 2012. Hydrogen bonding in biological structures. Springer Science & Business Media.
57. Herschlag, D. and Pinney, M.M., 2018. Hydrogen bonds: simple after all?. Biochemistry, 57(24), pp.3338-3352.
58. Liu, X., Chu, A., Nekouei, M., Lan, C.B., Auclair, K. and Saliba, K.J., 2025. Thiazole substitution of a labile amide bond-a new option towards stable pantothenamide-mimics. bioRxiv, pp.2025-02.
59. Bueno, J.M., Carda, M., Crespo, B., Cuñat, A.C., De Cozar, C., León, M.L., Marco, J.A., Roda, N. and Sanz-Cervera, J.F., 2016. Design, synthesis and antimalarial evaluation of novel thiazole derivatives. Bioorganic & Medicinal Chemistry Letters, 26(16), pp.3938-3944.
60. Petrou, A., Fesatidou, M. and Geronikaki, A., 2021. Thiazole ring - A biologically active scaffold. Molecules, 26(11), p.3166.
61. Kalita, T., Choudhury, A., Shakya, A., Ghosh, S.K., Singh, U.P. and Bhat, H.R., 2024. A review on synthetic thiazole derivatives as an antimalarial agent. Current Drug Discovery Technologies, 21(5), pp.10-42.
62. Thakur, S., Kumar, D., Goel, K.K., Rawat, P., Srivastava, V., Dhiman, S., Jadhav, H.R. and Dwivedi, A.R., 2024. Medicinal chemistry-based perspective on thiophene and its derivatives: exploring the structural insights to discover plausible druggable leads. RSC Medicinal Chemistry.
63. Morales-Tenorio, M., Lasala, F., Garcia-Rubia, A., Aledavood, E., Heung, M., Olal, C., Escudero-Pérez, B., Alonso, C., Martínez, A., Muñoz-Fontela, C. and Delgado, R., 2024. Discovery of thiophene derivatives as potent, orally bioavailable, and Blood–Brain Barrier-Permeable Ebola virus entry inhibitors. Journal of Medicinal Chemistry, 67(18), pp.16381-16402.
64. Dowd, P.A., 2023. Accuracy and precision. In Encyclopedia of Mathematical Geosciences (pp. 1-4). Cham: Springer International Publishing.
65. Fränti, P. and Mariescu-Istodor, R., 2023. Soft precision and recall. Pattern Recognition Letters, 167, pp.115-121
66. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B. and Zaslavsky, L., 2019. PubChem 2019 update: improved access to chemical data. Nucleic acids research, 47(D1), pp.D1102-D1109
67. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research, 40(D1), pp.D1100-D1107.