JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Predicting Antimalarial Activity in Natural Products Using Pretrained Bidirectional Encoder Representations from Transformers

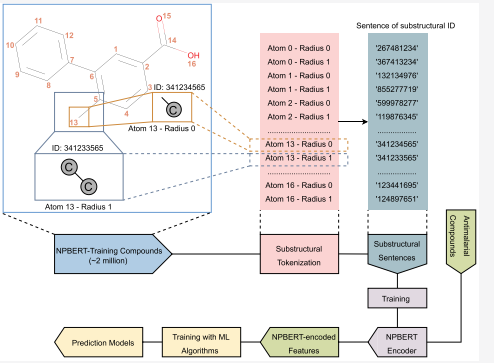
Thanh-Hoang Nguyen-Vo, Quang H. Trinh, Loc Nguyen, Trang T. T. Do, Matthew Chin Heng Chua,* and Binh P. Nguyen*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 5050−5058

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Malaria is a threatening disease that has claimed many lives and has a high prevalence rate annually. Through the past decade, there have been many studies to uncover effective antimalarial compounds to combat this disease. Alongside chemically synthesized chemicals, a number of natural compounds have also been proven to be as effective in their antimalarial properties. Besides experimental approaches to investigate antimalarial activities in natural products, computational methods have been developed with satisfactory outcomes obtained. In this study, we propose a novel molecular encoding scheme based on Bidirectional Encoder Representations from Transformers and used our pretrained encoding model called NPBERT with four machine learning algorithms, including *k*-Nearest Neighbors (*k*-NN), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), and Random Forest (RF), to develop various prediction models to identify antimalarial natural products. The results show that SVM models are the best-performing classifiers, followed by the XGB, *k*-NN, and RF models. Additionally, comparative analysis between our proposed molecular encoding scheme and existing state-of-the-art methods indicates that NPBERT is more effective compared to the others. Moreover, the deployment of transformers in constructing molecular encoders is not limited to this study but can be utilized for other biomedical applications.



1. INTRODUCTION

Malaria is a deadly contagious disease caused by the genus *Plasmodium*'s parasites.[1] For several centuries, malaria was considered one of the leading causes of death worldwide.[2] Annually, malaria causes up to 3 million deaths worldwide, accounting for 0.3−2.2% of total deaths,[1] and more than 200 million infected cases.[3] In tropical regions and several Asian countries, this rate ranges from 11.0% to 30.0%.[1] For decades, continuous efforts and global medical activities have been carried out to reduce mortality and morbidity as well as advance preventive systems, diagnostic processes, and medication for malaria. Several investigations, however, revealed that the prevalence of malaria parasite infection had continuously grown since 2015.[4,5] Due to the severity of this disease, the discovery of potent antimalarial agents has been promoted for decades. Despite numerous studies on developing novel and effective antimalarial compounds, the number of approved drugs is limited.[6] Among approved antimalarial drugs, quinine and its derivatives and antifolate combination drugs are the most prevalently used for clinical treatment.[6] In recent years, antimalarial drug resistance has emerged to be one of the most concerning issues in pharmaceutical science.[6,7] Besides drug resistance, managing the toxicity and side effects of antimalarial drugs is challenging.[8,9] Therefore, exploration of alternative ones derived from natural sources has been carried out to seek better-adapted and more effective candidates.[10] Along with current *in vitro* and *in vivo* platforms, *in silico* approaches employing powerful computing resources and computational advancements contribute considerably to antimalarial drug exploration.[11−15] In the past two decades, wide-spectral virtual screening (VS) using machine learning (ML) and deep learning (DL) has accelerated the search for potent compounds with desired properties.

Recently, numerous studies on the antiplasmodial (or antimalarial) activity of natural products (NPs) showed promising results.[16−20] Additionally, many reliable data sources on experimentally verified bioactive compounds, especially antimalarial compounds and their derivatives, have been made freely accessible.[21] Information on these databases, distinct and diverse structures, and the pharmaceutical importance of NPs

Special Issue: Computational Chemistry in Asia

Received: May 24, 2021
Published: August 16, 2021

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

was recently discussed.[22,23] The availability of data sets on antimalarial NPs and modern computational advances has motivated scientists to develop *in silico* frameworks to virtually identify potential NPs possessing strong antimalarial activity. The past two decades have seen an explosion of applications of ML and DL in various fields, especially central science[24−26] and life science.[27−29] The emergence of interdisciplinary fields, such as cheminformatics, bioinformatics, and health informatics, confirms the indispensable role of *in silico* advancement in supporting wet-lab research. In 2018, Egieyeh et al. first introduced a computational model to identify NPs having antimalarial activities.[11] Their data were collected from various confirmatory tests. Their models were constructed using four learning algorithms, including Support Vector Machines (SVM) tuned by using Sequential Minimization Optimization, Naive Bayesian (NB), Random Forest (RF), and Voted Perceptron (VP) using the WEKA 3.6 platform. Their outcomes showed that the RF model had obtained an area under the receiver operating characteristic curve (AUC) of 0.91, followed by the SVM, NB, and VP models with AUC values of 0.86, 0.74, and 0.72, respectively. Generally, besides a satisfactory predictive power, Egieyeh et al.'s prediction frameworks have limitations which need to be addressed. In the same year, Mosaddeque et al. conducted a Quantitative Structure−Activity Relationship (QSAR) study to predict antimalarial compounds having antihemozoin-formation activity.[12] They performed *in vitro* preliminary screening on antihemozoin-formation using high-throughput screening over 9600 compounds and finally obtained 224 hemozoin inhibitors. These hemozoin inhibitors with antimalarial action were then used for QSAR modeling with an accuracy of 91.23%. At the end of the same year, Mason et al. proposed using machine learning approaches to construct a prediction model, called CoSynE, to suggest synergistic combinations between known antimalarial compounds and novel ones.[13] Their model suggested 20 possible synergistic combinations with nine combinations that were then experimentally validated with promising outcomes. In 2019, Danishuddin et al. employed four learning algorithms, including *k*-NN, SVM, RF, and XGB, to develop their prediction models with multiple antimalarial data sets.[14] Their results showed that the SVM and XGB models had been achieved the best performance compared to the others with accuracies of up to 85%.

In this study, we propose a novel molecular encoding scheme and use it to develop several prediction models for identifying antimalarial NPs using four machine learning algorithms: *k*-Nearest Neighbors (*k*-NN), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), and Random Forest (RF). The molecular encoding scheme was developed using **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**),[30] a recent deep learning architecture that is widely used in natural language processing. The BERT architecture has been proven to be a robust and powerful machine translation method to create domain-specific embedding vectors. In 2019, SMILES-BERT[31] and SMILES Transformer,[32] two modified versions of BERT first designed for molecular encoding, were introduced with satisfactory outcomes obtained. In 2020, Chithrananda et al. proposed ChemBERTa for molecular property prediction.[33] To create effective and structurally distinct embedding vectors for antimalarial NPs, we propose a novel molecular encoding scheme called NPBERT, a BERT-based encoder specifically designed for NPs. To create the NPBERT encoder, we used about 2 million compounds, collected from the ChEMBL[34] and ZINC[35] databases, as training data. The antimalarial NPs for model development and evaluation were collected from Egieyeh et al. study[11] and the NPASS database.[36] For a fair assessment, trained classifiers using the NPBERT encoding scheme were compared with those using state-of-the-art methods, including 196-dimensional RDKit molecular descriptors,[37] extended-connectivity fingerprints,[38] and the Mol2Vec[39] encoding scheme.

2. MATERIALS AND METHODS

2.1. Overview of the Method. Figure 1 summarizes key steps in our study. Initially, about 1.9 million compounds and

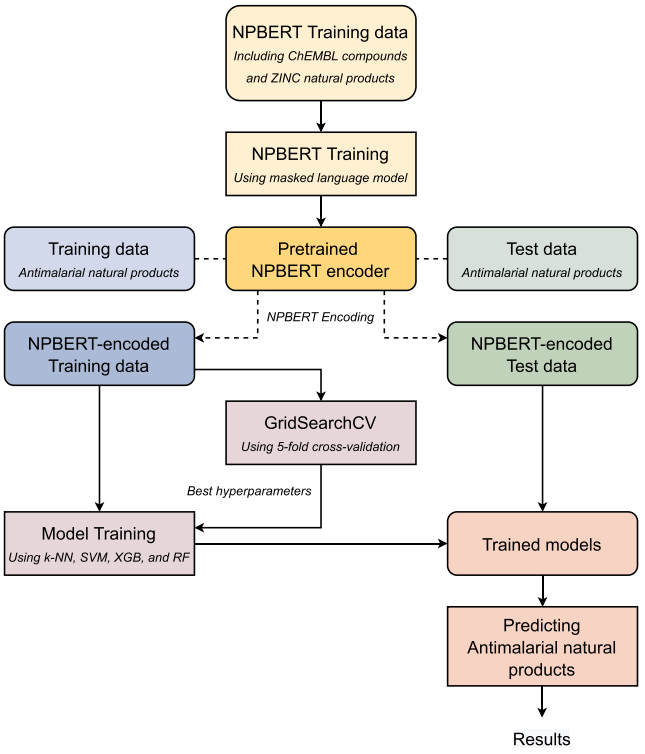


Figure 1. Processing steps in the proposed framework.

250 000 natural products (NPs) were downloaded from the ChEMBL database[34] and the ZINC database[35] to be used as training data for the NPBERT pretrained model (or shortly, NPBERT encoder). The original data set was then checked to remove duplicates and invalid molecular structures. The processed NPBERT encoder's training set contains nearly 2 million structural, verified compounds with high diversity in molecular scaffolds. The NPBERT encoder was constructed using the masked language modeling method. After obtaining the NPBERT encoder, we passed all antimalarial NPs (including training set and independent test set) through it to transform them into their corresponding NPBERT-encoded vectors. Those NPBERT-embedded vectors were then used for developing prediction models using four ML algorithms, including *k*-NN, SVM, XGB, and RF. To find the best hyperparameters for each model, we performed an exhaustive search over a determined grid of parameter values for these classifiers using 5-fold cross-validation. For each algorithm, the parameter set whose corresponding trained classifier showed the best performance was defined as the best hyperparameters. Those hyperparameters were then used to retrain the models

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

using the whole training set. Finally, the trained models were tested with the independent test set.

2.2. Data Set. *2.2.1. Data Set for NPBERT Encoder Training.* To build the NPBERT encoder, we used two sources of compounds: the ChEMBL[34] and ZINC databases.[35] The ChEMBL compound set has about 1.9 million molecules while the ZINC natural compound set contains about 250 000 molecules. Two sources of compounds were then merged and checked to remove any duplicates. Although the number of ZINC natural compounds is far smaller than that of ChEMBL compounds, the presence of these compounds can enrich substructural diversity, especially naturally distinct substructures. The processed NPBERT encoder's training data have about 2.0 million molecules. For training the NPBERT encoder, 85% of the total compounds were used as training data, while the rest of the compounds were used as validation data.

2.2.2. Data Set for Prediction Model Development. To construct models for predicting antimalarial activities of natural products (NPs), we collected training data from Egieyeh et al.[11] and the NPASS database[36] with 1155 and 1175 samples, respectively. The NPASS database is a reliable source of information on NPs and their inhibitory concentrations toward various targets (e.g., cell lines, proteins). In Egieyeh et al.'s[11] data set, 70% of the samples were labeled as "active," and the rest of the samples were labeled "inactive." Those compounds were collected from experimentally verified sources (literature, thesis, and public chemical databases). Those compounds were divided into two groups: active NPs (positive samples) and inactive NPs (negative samples), based on their 50% inhibitory concentration (IC_{50}) values. Compounds having IC_{50} values of <10 μM were considered active antimalarial compounds, while ones with IC_{50} values of ≥10 μM were considered inactive antimalarial compounds. Although Egieyeh et al.'s[11] data set is a valuable manually curated source of antimalarial NPs, there are several limitations, including duplicated and conflicting data (Table S6, Supporting Information). The duplicated data refer to compounds having similar molecular structures, while conflicting data are compounds with similar molecular structures but different labels assigned. For duplicated data, only one sample was kept. For conflicting data, all samples were removed. After cleaning the original Egieyeh et al. data set, we obtained the refined one. To enrich data for model development, we collected more antimalarial NPs from the NPASS database using a target search using the keyword "*Plasmodium falciparum*." The active NPs and inactive antimalarial NP were selected based on the same cutoff values of IC_{50} used in Egieyeh et al.[11] The two sources of data, including the refined Egieyeh et al. data set and the collected NPASS data set, were then merged and checked for duplicates, conflicting data, and invalid structures. The data refinement for the newly merged data set was processed in the same manner as being executed in the original Egieyeh et al. data set. The SMILES of NPs were cross-referenced with the PubChem database[21] to verify the structural validity. Finally, we obtained 1829 structurally verified NPs, including 1101 inactive antimalarial NPs (negative samples) and 728 active antimalarial NPs (positive samples). The independent test set was designed with 100 positive samples and 100 negative samples using random sampling. The rest of the compounds were used as training data with 628 positive samples and 1001 negative samples (Table 1). There are no overlapping samples between the training set and independent test set. The independent test set was an unseen data set that was not engaged in any steps related to the training process.

Table 1. Data for Prediction Model Development and Evaluation

data set	no. of samples		total
	active	inactive	
training set	628 (38.5%)	1001 (61.5%)	1629
independent test set	100 (50%)	100 (50%)	200
total	728	1101	1829

2.2.3. Class Rebalancing. Since the data set used has an unequal distribution of positive and negative samples, the Synthetic Minority Oversampling Technique (SMOTE)[40] was used to rebalance the classes. SMOTE was applied in two stages: model tuning and model development. To avoid overfitting, SMOTE was strictly controlled to apply to involved parts after splitting the original data set. In model tuning, we performed 5-fold cross-validations to find optimal hyperparameters. During 5-fold cross-validation, SMOTE was applied to the training folds only. Similarly, in model development, SMOTE was applied to the training set only.

2.3. Training NPBERT Encoder. Representation learning has been successfully applied in the domain of Natural Language Processing (NLP) to create word embeddings. Being inspired by this idea, various versions of representation learning for chemical[26,39,41] and biological[42] data have been proposed. These novel approaches, such as Mol2Vec, represent the molecule as a sequence of molecular substructures and learn vector embeddings for each present substructure. They have been demonstrated to work more effectively compared to classical molecular fingerprints in the form of binary vectors. These molecular encoders, however, are insufficiently strong to deal with long-distance bidirectional dependencies. To pool the embeddings of all substructures building up molecules, previous approaches depend on a simple summation of substructural embeddings regardless of the fact that their contributions are not equal. Thus, we introduce the NPBERT encoder to address the limitations of existing techniques by using the self-attention mechanism, a recent advancement in language models.

2.3.1. Extraction of Substructural Sentences. Initially, substructural sentences of molecules were extracted using RDKit with two scales: zero-radius and one-radius. For a molecule, at a particular atomic position, the zero-radius scale describes the atom and its bondings. However, for the one-radius scale, it will be the atom, its bondings, and neighboring atoms in contact with those bonds (Figure 2). The zero-radius and one-radius substructures of an atom form an identifier pair identified by RDKit. On the basis of the order of appearance of substructures, the identifier pairs are organized into a "sentence" of symbolic tokens. Since the sequential order of appeared substructural tokens indicates their bonding arrangement in a molecule, substructures identified from neighboring atoms tend to appear next to each other in the sentence to collectively form a rough chain throughout the molecule. Our proposed molecular representation is expected to efficiently approximate the adjacency likelihoods of substructures in a molecule.

2.3.2. Model Architecture. Bidirectional Encoder Representations from Transformers (BERT)[30] is a recent step forward in language modeling enabled by the success of transformers.[43] In Vaswani et al.'s[43] original implementation, the transformer architecture was built for sequence transduction, specifically for the machine translation task. It contains a multilayer transformer encoder and decoder while BERT architecture uses the encoder only. The encoder contains a stack of identical layers in which

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

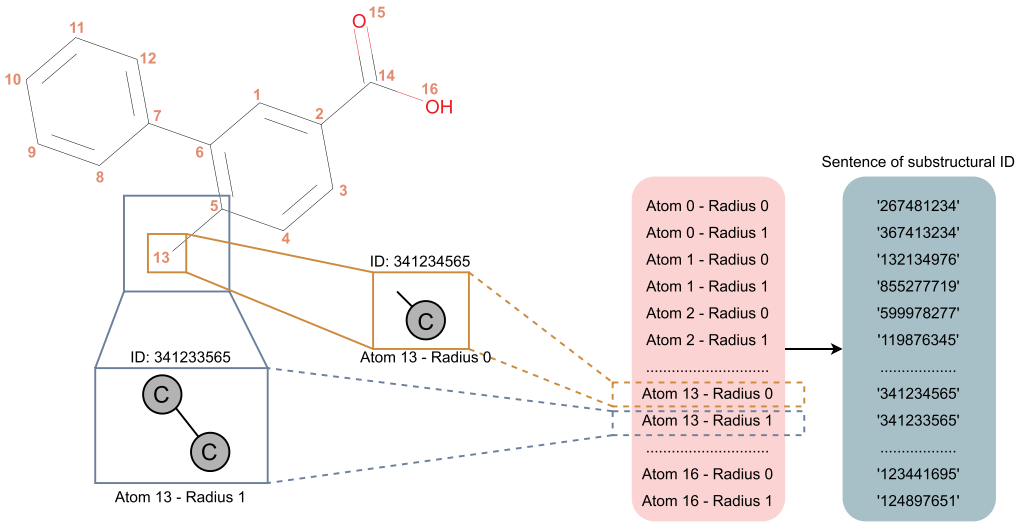


Figure 2. Extraction of substructural sentences of a compound.

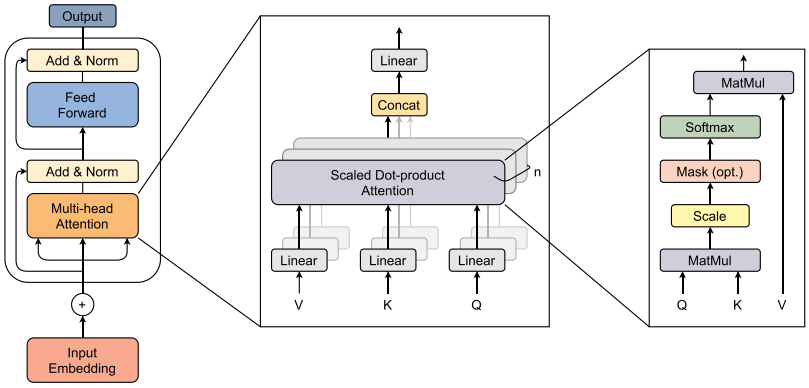


Figure 3. Architecture of bidirectional encoder representations from transformers.

each layer is composed of a multihead self-attention mechanism and a position-wise fully connected feed-forward network. A residual connection is employed around each of the encoder layers, followed by layer normalization (Figure 3).

The attention mechanism maps a query and a set of key-value pairs to an output, where the query (**Q**), key (**K**), and values (**V**) are all vectors. The query and key vectors are used to compute the compatibility of the input to the output. The final output is an average of all the input values, weighted by the scores provided by the compatibility function. In Vaswani et al.'s[43] implementation, the attention mechanism scales the dot-products of the query and key vectors by $\frac{1}{\sqrt{d_k}}$, where d_k is the dimension of the key vector and uses the result to weight the corresponding value vector. The query, keys, and values vectors are packed together as matrices **Q**, **K**, and **V**, and the outputs are computed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

With a single attention head, averaging prevents the model from jointly attending to different segments of the sequence simultaneously. The transformers overcomes this by using a multiheaded attention layer composed of multiple attention mechanisms (heads), each learning to its own weights separately. The multiheaded attention layer outputs one weighted value vector for each attention head, which all are then concatenated into a single vector used as inputs for subsequent layers of the transformer encoder.

For a molecule, it can be represented as a sequence of substructural tokens. The transformer can jointly attend to each substructure that constitutes the molecule. Each substructural token is first replaced by an input embedding that is unique to it. In order to retain information about the position of each substructural token in the sentence, a positional embedding vector is added to each input embedding vector. The positional embedding has the same dimension as the input embeddings for summation. Sine and cosine functions of different frequencies are used as positional embeddings. For the i^{th} dimension of the positional embedding vector at position pos in the sequence, the positional embedding (PE) is calculated as

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{2}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{3}$$

Our proposed architecture of NPBERT was designed with five hidden encoder layers of four-head attention. The output dimension was set at 512. Gaussian Error Linear Units (GELU) were used as the activation, and a dropout rate of 10% was added

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

in between these layers. The NPBERT model was optimized using the AdamW optimizer. For the pretraining task, Masked Language Model (MLM) was employed. NPBERT was used to predict the tokens in 15% of the sentences. In 15% of randomly chosen sentences for the prediction task, 80% of the tokens were replaced with a "Mask" token, 10% were replaced with a random token, and 10% were unchanged. In our experiments, NPBERT was implemented using PyTorch 1.3.1 and trained on Google Colab equipped with 25 GB of RAM and one NVIDIA Tesla T4 GPU. NPBERT was trained over 20 epochs. It took about 10 hours to complete one epoch.

2.4. Learning Algorithms. *2.4.1. k-Nearest Neighbors.* *k*-Nearest Neighbors (*k*-NN), a simple supervised learning algorithm, was first introduced by Fix and Hodges in 1951[44] and then completely conceptualized by Altman in 1992.[45] *k*-NN can be employed to deal with regression and classification problems. In the *k*-NN algorithm, the number of *k* closest neighbors decides values or classes of unknown samples. As a distance-based measuring method, *k*-NN is highly sensitive to unnormalized samples.

2.4.2. Support Vector Machines. Support Vector Machines (SVM)[46] is one of the most frequently used supervised learning algorithms. In the beginning, SVM was mainly used to address binary classification problems, but it was then developed to effectively deal with various complicated classification problems. It separates multidimensional data by creating multi-hyperplanes of maximum expansion of the margin. This method limits the accuracy collapse in case the number of features far exceeds that of the samples. SVM, thus, is considered to be flexible and smoothly adaptive to immense types of data while still being able to have remarkable accuracy.

2.4.3. eXtreme Gradient Boosting. eXtreme Gradient Boosting (XGB) is a supervised ensemble learning algorithm initially developed by Chen and Guestrin.[47] XGB is a more regularized implementation derived from Gradient Tree Boosting[48,49] combined with Classification and Regression Trees (CART).[50] It is a computationally accelerated algorithm that can effectively control overfitting to produce better performances. In boosting, single classifiers are not constructed using entirely random subsets of data and features but sequentially in which mispredicted instances will be added with more weights.

2.4.4. Random Forest. Random Forest (RF)[51] is a supervised ensemble learning algorithm that incorporates the "bagging" idea[52] and random selection of features[53] to create various decision trees during training. The trees are somewhat different from each other, and the output is returned based on the mode of the classes or the averaged prediction of the single trees for classification or regression problems, respectively. The RF algorithm addresses the shortcomings of decision trees: they have tendencies to overfit the training data.

2.4.5. Model Tuning. To tune the *k*-NN, SVM, XGB, and RF models, we performed exhaustive explorations over a selected grid of parameter values using 5-fold cross-validation. For the *k*-NN models, `k` (the number of neighbors) was tuned. For the SVM models, `C` (error control coefficient) and `gamma` (curvature weight of the decision boundary) were tuned. For XGB, `max_depth` (the maximum depth of the tree), `colsample_bytree` (subsample ratio of columns when building a single tree), and `learning_rate` (model learning speed) were tuned. For RF models, `max_depth` (the maximum depth of the tree), `min_samples_split` (the minimum number of samples needed to split an internal node), and `max_features` (the number of features to examine when deciding the best split) were tuned. The details of searching ranges of parameter values are provided in Table S1 (Supporting Information).

2.5. Molecular Encoding and Representation Schemes. *2.5.1. Extended-Connectivity Fingerprints.* Extended-connectivity fingerprints (ECFPs), also known as circular fingerprints or Morgan fingerprints, belong to a class of topological fingerprints.[38] The ECFP scheme can create various types of circular fingerprints depending on the radius and bit number. The SMILES-formatted compounds were converted to their corresponding 1024-dimensional and 2048-dimensional ECFP binary vectors using a radius of 2.

2.5.2. RDKit Molecular Descriptors. The set of 196 RDKit molecular descriptors was calculated using RDKit,[37] an open-source cheminformatics library, to create 196-dimensional vectors. The molecular descriptors include 106 constitutional descriptors, 58 MOE-type descriptors, 12 connectivity descriptors, seven topological descriptors, four molecular property descriptors, and one CPSA descriptor. RDKit directly converts SMILES-formatted compounds into their corresponding 196-dimensional vectors.

2.5.3. Mol2Vec Encoding. Mol2Vec, a pretrained molecular encoder, was developed and first introduced by Jaeger et al.[39] based on a similar idea of word embedding.[54] Mol2Vec learns substructural sentences of numerous compounds to create molecular representations. The SMILES-formatted compounds were converted to their corresponding 300-dimensional Mol2Vec-encoded vectors.

2.5.4. NPBERT Encoding. NPBERT is our proposed molecular encoding scheme, which is a pretrained model developed using the Bidirectional Encoder Representations from Transformers (BERT) architecture. Similarly to Mol2Vec, NPBERT also learns tokenized substructures of copious compounds, but the training was incorporated with the self-attention mechanism to improve the adaptive molecular embedding. The SMILES-formatted compounds were converted to their corresponding 512-dimensional NPBERT-encoded vectors.

2.6. Evaluation Metrics. To assess the model performance, balanced accuracy (BA), specificity (SP), sensitivity (SN), Cohen's Kappa (CK), and the area under the receiver operating characteristic curve (ROC-AUC) were assessed. TP, FP, TN, and FN are the abbreviated terms of True Positive, False Positive, True Negative, and False Negative values, respectively. The mathematical formulas of these evaluation metrics are expressed below

$$\mathrm{BA} = \frac{(\mathrm{SN} + \mathrm{SP})}{2} \tag{4}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{5}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{CK} = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

where p_o and p_e are the relative observed agreement among raters and hypothetical probability of chance agreement, respectively.

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

Table 2. Performance of Models Trained with Various Feature Schemes (Not Using SMOTE)

model	feature set	ROC-AUC	BA	SN	SP	KP
k-NN	1024-bit ECFP2	0.7077	0.6350	0.4800	0.7900	0.2700
	2048-bit ECFP2	0.7317	0.6100	0.3200	**0.9000**	0.2200
	RDKit MD	0.7145	0.6500	0.4300	0.8700	0.3000
	Mol2Vec	0.7060	**0.6750**	0.6000	0.7500	**0.3500**
	NPBERT	**0.7393**	0.6400	0.4400	0.8400	0.2800
SVM	1024-bit ECFP2	0.7164	0.6550	0.4800	0.8300	0.3100
	2048-bit ECFP2	0.7096	0.6500	0.4600	**0.8400**	0.3000
	RDKit MD	0.7160	0.6600	0.4900	0.8300	0.3200
	Mol2Vec	0.7290	0.6450	0.4700	0.8200	0.2900
	NPBERT	**0.7749**	**0.7000**	**0.5800**	0.8200	**0.4000**
XGB	1024-bit ECFP2	0.7090	0.6250	0.4700	0.7800	0.2500
	2048-bit ECFP2	0.7312	0.6400	0.4700	0.8100	0.2800
	RDKit MD	0.7069	0.6350	0.4700	0.8000	0.2700
	Mol2Vec	0.7222	**0.6650**	0.4800	**0.8500**	**0.3300**
	NPBERT	**0.7406**	0.6300	**0.5100**	0.7500	0.2600
RF	1024-bit ECFP2	0.7086	0.6200	0.3500	**0.8900**	0.2400
	2048-bit ECFP2	0.7098	0.6000	0.3100	**0.8900**	0.2000
	RDKit MD	0.7132	0.6250	**0.4500**	0.8000	0.2500
	Mol2Vec	0.7114	**0.6300**	0.4000	0.8600	**0.2600**
	NPBERT	**0.7141**	0.6250	0.4400	0.8100	0.2500

Table 3. Performance of Models Trained with Various Feature Schemes (Using SMOTE)

model	feature set	ROC-AUC	BA	SN	SP	KP
k-NN	1024-bit ECFP2	0.6787	0.5850	**0.8100**	0.3600	0.1700
	2048-bit ECFP2	0.6835	0.5650	0.8000	0.3300	0.1300
	RDKit MD	0.6930	0.6250	0.5500	0.7000	0.2500
	Mol2Vec	0.6610	0.6250	0.6500	0.6000	0.2500
	NPBERT	**0.7385**	**0.6550**	0.5500	**0.7600**	**0.3100**
SVM	1024-bit ECFP2	0.6824	0.6500	0.5400	0.7600	0.3000
	2048-bit ECFP2	0.6730	0.6500	0.5500	0.7500	0.3000
	RDKit MD	0.7144	0.6700	0.5800	0.7600	0.3400
	Mol2Vec	0.6985	0.6800	0.6000	0.7600	0.3600
	NPBERT	**0.7696**	**0.7000**	**0.6300**	**0.7700**	**0.4000**
XGB	1024-bit ECFP2	0.7142	**0.6600**	0.5800	0.7400	0.3200
	2048-bit ECFP2	0.6938	0.6550	**0.6100**	0.7000	**0.3100**
	RDKit MD	0.6926	0.6350	0.5200	**0.7500**	0.2700
	Mol2Vec	0.7005	0.6350	0.6000	0.6700	0.2700
	NPBERT	**0.7288**	0.6550	0.5800	0.7300	**0.3100**
RF	1024-bit ECFP2	**0.7428**	**0.6850**	0.6300	**0.7400**	**0.3700**
	2048-bit ECFP2	0.7154	0.6600	**0.6400**	0.6800	0.3200
	RDKit MD	0.7036	0.6600	0.6300	0.6900	0.3200
	Mol2Vec	0.7000	0.6500	0.5700	0.7300	0.3000
	NPBERT	0.6975	0.6600	0.6100	0.7100	0.3200

3. RESULTS AND DISCUSSION

3.1. Model Development and Evaluation. The NPBERT encoder was trained over 20 epochs. After 18 epochs, validation loss continued to decrease, but the loss variation is trivial (less than 5%). Therefore, we selected the pretrained model at epoch 18 to be used as the NPBERT encoder (Figure S1, Supporting Information). For each learning algorithm, five different classifiers were trained using five corresponding molecular encoding schemes and representations. In our experiments, since the training set had imbalanced classes, we employed the Synthetic Minority Oversampling Technique (SMOTE)[40] to rebalance the classes. We decided to test two training scenarios: using SMOTE and not using SMOTE. Tables 2 and 3 provide information on the predictive performance of trained models not using SMOTE and those using SMOTE.

For non-SMOTE models, the use of the NPBERT encoding showed improvements in the predictive performance of models using the *k*-NN, SVM, and XGB algorithms. For models trained with the *k*-NN, SVM, and XGB algorithms, the models using NPBERT encoding showed better performance compared to those using Mol2Vec encoding, ECFP, and RDKit molecular descriptors. For models trained with RF, the performance of the models using RDKit molecular descriptors, Mol2Vec-encoded features, and NPBERT-encoded features are equivalent (Table 2). For SMOTE-used models, the use of the NPBERT encoding presented significant increases in the predictive performance of the models trained with the *k*-NN, SVM, and XGB algorithms. For models trained with RF, the performance of the model using 1024-bit ECFP2 stayed at the top, followed by those using 2048-bit ECFP2 and other encoding schemes. Generally, except for RF models, the models trained with other algorithms using

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

NPBERT encoding work more effectively than corresponding models using other encoding schemes and representations. For RF models, although trained models using NPBERT encoding do not work as efficiently as expected in both scenarios, the superior performance of other models still provides sufficient evidence to confirm the effectiveness of our proposed encoding scheme. Under the scope of this study, the performance of the NPBERT scheme exceeds those of four other state-of-the-art methods in terms of identifying the antimalarial activity of natural products (Table 3). The ROC curves for the *k*-NN, SVM, XGB, and RF models are provided in Figures S2, S3, S4, and S5 (Supporting Information), respectively. The 5-fold cross-validation results are provided in Tables S2 and S3 (Supporting Information).

3.2. Comparison with Relevant Studies. Egieyeh et al. developed various prediction models using NB, VP, RF, and SVM.[11] The best-performing and second best-performing models were RF and SVM models with ROC-AUC values of 0.9100 and 0.8600, respectively. From the original data set, they used 80% and 20% of the data set as their training set and validation set, respectively, using stratified sampling. The ratio of active antimalarial compounds to inactive ones is 7:3. To address the class imbalance problem, they reported that SMOTE had been used, but there was no detailed explanation mentioning how SMOTE was applied. There might be an unexpected scenario in which SMOTE was done before performing cross-validation. In case this assumed scenario occurred, the model might be overfitted. Egieyeh et al. conducted 10-fold cross-validation in which nine-fold of it was used as the training set, while one-fold was used as the validation set. During the *k*-fold cross-validation process, every fold was iteratively treated as the validation set, and the final evaluation was based on the averaged evaluation of all *k* validation sets. When employing SMOTE in *k*-fold cross-validation, it is permitted to be applied on the training set only. Moreover, many duplicates were found in their original data set, and this issue might significantly affect the model's outcomes. Although Egieyeh et al. is the most relevant to our study, we decided not to compare ours to theirs due to these reasons. In our approach, the SVM model was the best model, obtaining ROC-AUC values of 0.7749 and 0.7696 for the not using SMOTE and using SMOTE scenarios, respectively. Unlike Egieyeh et al.'s study, our experiments were carefully designed and performed with a training set and an independent test set. Additionally, the utilization of SMOTE in our study was clearly explained and appropriately processed to avoid overfitting. Therefore, our proposed computational method can be considered a significantly better approach in terms of both study design and experiments.

3.3. Comparative Analysis among Encoding Schemes and Representations. Our experimental results initially showed that the performance of models trained with all encoding schemes and representations varies between two conditions: using SMOTE and not using SMOTE. Generally, while the performance of all models built with distance-based algorithms (*k*-NN and SVM) had downward trends from not using SMOTE to using SMOTE conditions, the performance of all models built with tree-based algorithms (XGB and RF) seem to be relatively stable with minor changes. Compared to other molecular encoding schemes and representations, models trained with the NPBERT feature set have better performance when combining with *k*-NN and SVM. In comparison between the two conditions, the variations in the ROC-AUC values in the *k*-NN and SVM models using the NPBERT encoding are negligible. The RF algorithm seems to not work very effectively with NPBERT, but NPBERT's performance is still considered competitive under the not using SMOTE condition. Under the not using SMOTE condition, Mol2Vec's performance in terms of BA and KP is stronger than that of NPBERT, except for the SVM-based model. Under the using SMOTE condition, NPBERT's performance in terms of BA and KP is higher than that for the *k*-NN, SVM, and XGB models. To fully assess the performance of NPBERT in comparison with each encoding scheme, we used the two-tailed DeLong's test[55] to compare the significant difference in the ROC-AUC values between NPBERT and the others. The results are unsurprisingly anticipated, in that NPBERT can work very well with distance-based algorithms but ineffectively perform in tree-based algorithms. Under the not using SMOTE condition, ROC-AUC values in the SVM models are significantly higher than those of 1024-bit ECFP2 and 2048-bit ECFP2 with p values of 0.0412 (<0.05) and 0.0244 (<0.05), respectively. Under the using SMOTE condition, ROC-AUC values in the SVM models are significantly higher than those of the models using other encoding schemes except for RDKit MD. ROC-AUC values in the *k*-NN model are also significantly greater than those of the models using Mol2Vec. Although NPBERT's performance in the XGB and RF models is not as good as expected, the statistical insignificance (inferred from the test) shows that our proposed method is also competitive when compared to the state-of-the-art methods. Our experimental results and hypothesis testing bring us to a strong conclusion that NPBERT is a robust and effective encoding scheme, especially when being used in combination with distance-based algorithms. Tables S4 and S5 provide information on DeLong's test results for the not using SMOTE and using SMOTE conditions, respectively.

3.4. Strengths and Limitations. NPBERT, our proposed molecular encoder, was developed using one of the most effective recurrent neural networks incorporated with a self-attention mechanism to help it better learn the substructural characteristics with a focus at specific positions. In comparison with the Mol2Vec encoder, the training of NPBERT used less computing resource with a smaller data set. While NPBERT required only 2 million compounds for its training process, it took nearly 20 million compounds to train Mol2Vec. Additionally, since the NPBERT encoder is designed to cope with molecular embedding for natural products, it can somehow work more efficiently compared to the Mol2Vec encoder in predicting bioactivities and properties of natural products. However, the NPBERT encoder may not always be the best solution in all natural product-related prediction tasks. Presently, many more advanced transformer architectures have better performance compared to BERT. Therefore, creating a different version of NPBERT that varies according to situations is highly recommended instead of sticking with NPBERT only. Additionally, another limitation of NPBERT is that the encoding is limited to small molecules, thus peptides are not covered, though they also play an important role as potential antimalarial agents.[56,57]

Data and Software Availability. Chemical data used in this study were partially collected from Egieyeh et al.[11] and the NPASS database[36] with independent recurration. Source code and data are available at https://github.com/mldlproject/2021-NPBERT-Antimalaria.

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058

4. CONCLUSIONS

Our proposed molecular encoding scheme, NPBERT, confirms its superior performance compared to other state-of-the-art molecular encoding methods. On the other hand, our prediction models for identifying antimalarial natural products obtain ROC-AUC values of up to 0.7749. The application of NPBERT-encoded features for model development significantly improved the predictive power of constructed models to better detect potent antimalarial natural products. Subsequently, NPBERT, as well as its future modified version, can be applied in many other natural product-related prediction tasks.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00584.

Plots presenting changes in training losses versus validation losses of the NPBERT pretrained model, plots displaying ROC curves and AUC values, parameter searching range during model tuning, and 5-fold cross-validated AUC values (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Matthew Chin Heng Chua − *Institute of Systems Science, National University of Singapore, Singapore 119620, Singapore;* orcid.org/0000-0002-5200-5079; Phone: +65 6516 2088; Email: mattchua@nus.edu.sg

Binh P. Nguyen − *School of Mathematics and Statistics, Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand;* orcid.org/0000-0001-6203-6664; Phone: +64 4 886 4489; Email: binh.p.nguyen@vuw.ac.nz

Authors

Thanh-Hoang Nguyen-Vo − *School of Mathematics and Statistics, Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand;* orcid.org/0000-0003-0006-5245

Quang H. Trinh − *Computational Biology Center, International University−VNU HCMC, Ho Chi Minh City 700000, Vietnam;* orcid.org/0000-0001-9724-8405

Loc Nguyen − *Computational Biology Center, International University−VNU HCMC, Ho Chi Minh City 700000, Vietnam;* orcid.org/0000-0003-0561-6659

Trang T. T. Do − *School of Business and Information Technology, Wellington Institute of Technology, Lower Hutt 5012, New Zealand;* orcid.org/0000-0002-1614-4661

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00584

Author Contributions

T.-H.N.-V.: data curation, investigation, formal analysis, validation, visualization, writing—original draft, writing—review and editing. Q.H.T.: investigation, data curation, software. L.N.: investigation, data curation. T.T.T.D.: formal analysis, writing—review and editing. M.C.H.C.: methodology, resources, writing—review and editing, supervision. B.P.N.: conceptualization, methodology, formal analysis, visualization, writing—review & editing, supervision.

Funding

The work of T.T.T.D. was supported in part by the Whitireia and WelTec Contestable fund.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) White, N. J.; Pukrittayakamee, S.; Hien, T. T.; Faiz, M. A.; Mokuolu, O. A.; Dondorp, A. M. Erratum: Malaria. *Lancet* **2014**, *383*, 696−696.

(2) Talapko, J.; Škrlec, I.; Alebić, T.; Jukić, M.; Včev, A. Malaria: the past and the present. *Microorganisms* **2019**, *7*, 179.

(3) WHO. *World Malaria Report 2018*; World Health Organization, 2016.

(4) WHO. *Epidemiological Update: Increase of Malaria in the Americas*; World Health Organization, 2018.

(5) Dhiman, S. Are malaria elimination efforts on right track? An analysis of gains achieved and challenges ahead. *Infectious Diseases of Poverty* **2019**, *8*, 1−19.

(6) Bloland, P. B. *Drug Resistance in Malaria*; World Health Organization, 2001.

(7) White, N. J. Antimalarial drug resistance. *J. Clin. Invest.* **2004**, *113*, 1084−1092.

(8) Taylor, W. R. J.; White, N. J. Antimalarial drug toxicity. *Drug Saf.* **2004**, *27*, 25−61.

(9) AlKadi, H. O. Antimalarial drug toxicity: a review. *Chemotherapy* **2007**, *53*, 385−391.

(10) Mojab, F. Antimalarial natural products: a review. *Avicenna J. Phytomed.* **2012**, *2*, 52.

(11) Egieyeh, S.; Syce, J.; Malan, S. F.; Christoffels, A. Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. *PLoS One* **2018**, *13*, e0204644.

(12) Mosaddeque, F.; Mizukami, S.; Kamel, M. G.; Teklemichael, A. A.; Dat, T. V.; Mizuta, S.; Toan, D. V.; Ahmed, A. M.; Vuong, N. L.; Elhady, M. T.; Giang, H. T. N.; Dang, T. N.; Fukuda, M.; Huynh, L. K.; Tanaka, Y.; Egan, T. J.; Kaneko, O.; Huy, N. T.; Hirayama, K. Prediction model for antimalarial activities of hemozoin inhibitors by using physicochemical properties. *Antimicrob. Agents Chemother.* **2018**, *62*, DOI: 10.1128/AAC.02424-17.

(13) Mason, D. J.; Eastman, R. T.; Lewis, R. P. I.; Stott, I. P.; Guha, R.; Bender, A. Using machine learning to predict synergistic antimalarial compound combinations with novel structures. *Front. Pharmacol.* **2018**, *9*, 1096.

(14) Danishuddin; Madhukar, G.; Malik, M. Z.; Subbarao, N. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR QSAR Environ. Res.* **2019**, *30*, 543−560.

(15) Ashdown, G. W.; Dimon, M.; Fan, M.; Sánchez-Román Terán, F.; Witmer, K.; Gaboriau, D. C. A.; Armstrong, Z.; Ando, D. M.; Baum, J. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. *Sci. Adv.* **2020**, *6*, eaba9338.

(16) Batista, R.; De Jesus Silva Júnior, A.; De Oliveira, A. Plant-derived antimalarial agents: new leads and efficient phytomedicines. Part II. Non-alkaloidal natural products. *Molecules* **2009**, *14*, 3037−3072.

(17) Mayer, A. M. S.; Rodríguez, A. D.; Berlinck, R. G. S.; Fusetani, N. Marine pharmacology in 2007−8: Marine compounds with anti-bacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; affecting the immune and nervous system, and other miscellaneous mechanisms of action. *Comp. Biochem. Physiol., Part C: Toxicol. Pharmacol.* **2011**, *153*, 191−222.

(18) Davis, R. A.; Buchanan, M. S.; Duffy, S.; Avery, V. M.; Charman, S. A.; Charman, W. N.; White, K. L.; Shackleford, D. M.; Edstein, M. D.; Andrews, K. T.; Camp, D.; Quinn, R. J. Antimalarial activity of pyrroloiminoquinones from the Australian marine sponge Zyzzya sp. *J. Med. Chem.* **2012**, *55*, 5851−5858.

(19) Xu, Y.-J.; Pieters, L. Recent developments in antimalarial natural products isolated from medicinal plants. *Mini-Rev. Med. Chem.* **2013**, *13*, 1056−1072.

(20) Ehata, M. T.; Lumpu, S. N.; Munduku, C. K.; Kabangu, O. K.; Cos, P.; Maes, L.; Apers, S.; Vlietinck, A. J.; Pieters, L.; Kanyanga, R. C.

Study of Antiparasitic and Cytotoxicity of the Aqueous, the 80% Methanol Extract and Its Fractions, and the Acute Toxicity of the Aqueous Extract of Brucea sumatrana (Simaroubaceae) Leaves Collected in Mai-Ndombe, Democratic Republic of Congo. *Chin. Med. (Irvine, CA, U. S.)* **2016**, *7*, 93.
(21) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(22) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data resources for the computer-guided discovery of bioactive natural products. *J. Chem. Inf. Model.* **2017**, *57*, 2099−2111.
(23) Nguyen-Vo, T.-H.; Nguyen, L.; Do, N.; Nguyen, T.-N.; Trinh, K.; Cao, H.; Le, L. Plant Metabolite Databases: From Herbal Medicines to Modern Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 1101−1110.
(24) Li, G.-H.; Huang, J.-F. CDRUG: a web server for predicting anticancer activity of chemical compounds. *Bioinformatics* **2012**, *28*, 3334−3335.
(25) Al-Fakih, A. M.; Algamal, Z. Y.; Lee, M. H.; Aziz, M.; Ali, H. T. M. A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm. *SAR QSAR Environ. Res.* **2019**, *30*, 403−416.
(26) Nguyen-Vo, T.-H.; Nguyen, L.; Do, N.; Le, P. H.; Nguyen, T.-N.; Nguyen, B. P.; Le, L. Predicting Drug-Induced Liver Injury Using Convolutional Neural Network and Molecular Fingerprint-Embedded Features. *ACS Omega* **2020**, *5*, 25432−25439.
(27) Helm, M.; Motorin, Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.* **2017**, *18*, 275−291.
(28) Nguyen-Vo, T.-H.; Nguyen, Q. H.; Do, T. T. T.; Nguyen, T.-N.; Rahardja, S.; Nguyen, B. P. iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features. *BMC Genomics* **2019**, *20*, 1−11.
(29) Zhang, D.; Xu, Z.-C.; Su, W.; Yang, Y.-H.; Lv, H.; Yang, H.; Lin, H. iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics* **2021**, *37*, 171−177.
(30) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**.
(31) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* **2019**, 429−436.
(32) Honda, S.; Shi, S.; Ueda, H. R. SMILES Transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv* **2019**.
(33) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**.
(34) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(35) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(36) Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **2018**, *46*, D1217−D1222.
(37) Landrum, G. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org.
(38) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(39) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(40) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321−357.
(41) Jeon, W.; Kim, D. FP2VEC: a new molecular featurizer for learning molecular properties. *Bioinformatics* **2019**, *35*, 4979−4985.
(42) Kimothi, D.; Soni, A.; Biyani, P.; Hogan, J. M. Distributed representations for biological sequence analysis. *arXiv* **2016**.
(43) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Proceedings of 31st Conference on Neural Information Processing Systems*; Curran Associates Inc., 2017; pp 5998−6008.
(44) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review/Revue Internationale de Statistique* **1989**, *57*, 238−247.
(45) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175−185.
(46) Cortes, C.; Vapnik, V. Support vector networks. *Machine Learning* **1995**, *20*, 273−297.
(47) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*; ACM, 2016; pp 785−794.
(48) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Math. Stat.* **2001**, *29*, 1189−1232.
(49) Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, M. R. Boosting algorithms as gradient descent. *Advances in Neural Information Processing Systems* **2000**, 512−518.
(50) Steinberg, D.; Colla, P. *CART: Classification and Regression Trees*; CRC Press: New York, 2009; Vol. *9*; p 179.
(51) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(52) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123−140.
(53) Ho, T. K. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*; IEEE-CS Press, 1995; pp 278−282.
(54) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**.
(55) DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **1988**, *44*, 837−845.
(56) Spänig, S.; Heider, D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Min.* **2019**, *12*, 1−29.
(57) Spänig, S.; Mohsen, S.; Hattab, G.; Hauschild, A.-C.; Heider, D. A large-scale comparative study on peptide encodings for biomedical classification. *NAR Genomics and Bioinformatics* **2021**, *3*, lqab039.

https://doi.org/10.1021/acs.jcim.1c00584
J. Chem. Inf. Model. 2022, 62, 5050−5058