

# BRAFPred: A Novel Approach for Accurate Prediction of the B-Type Rapidly Accelerated Fibrosarcoma Inhibitor

Ming Zhang, Chaoming Zhang, Keyu Liu, Xibei Yang, Xiaojian Liu, and Fang Ge\*



Cite This: *ACS Omega* 2025, 10, 12170–12184



Read Online

ACCESS |



Metrics & More



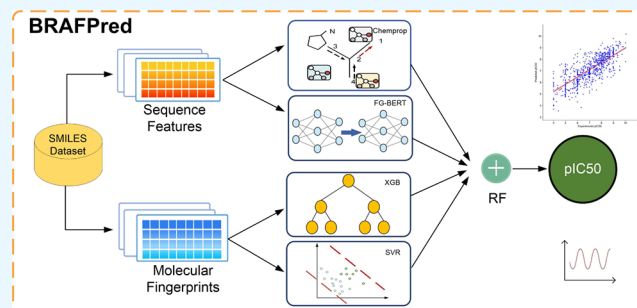
Article Recommendations



Supporting Information

**ABSTRACT:** B-type rapidly accelerated fibrosarcoma (BRAF) is a key oncogene that regulates cell signaling and proliferation, rendering it a crucial target for cancer therapeutics. Traditional QSAR methods are hindered by their reliance on a singular model, their inability to grasp complex nonlinearities, and limited generalization, undermining predictive efficacy. To address these challenges, we introduce BRAFPred, a novel framework that leverages stacked ensemble learning to integrate both classical machine learning and advanced deep learning techniques for the precise prediction of BRAF inhibitors. We utilized 12 handcrafted molecular descriptors derived from PaDeL, in conjunction with small molecule sequence features, as foundational inputs.

Furthermore, we employed extreme gradient boosting (XGB), support vector regression (SVR), and deep learning architectures based on Chemprop and a pretrained BERT model (FG-BERT) to generate additional predictive features. These multisource features were subsequently integrated within a meta-ensemble random forest regression model, which utilized 26 input variables. Empirical results demonstrate that BRAFPred significantly outperforms benchmark models, achieving a mean absolute error (MAE) of 0.383 and a coefficient of determination ( $R^2$ ) of 0.855, surpassing Chemprop (MAE = 0.443,  $R^2$  = 0.803), FG-BERT (MAE = 0.460,  $R^2$  = 0.785), and Stack\_BRAF (MAE = 0.403,  $R^2$  = 0.839). Extensive evaluation on benchmark data sets affirms BRAFPred's superiority over state-of-the-art methodologies, with robust generalization capabilities demonstrated on blind test sets. Additionally, ablation studies and case analyses underscore the robustness of the model's design. The source code, data sets, and prediction results for BRAFPred are available for further research at <https://github.com/EvanZhang1216/BRAFPred>.



## 1. INTRODUCTION

BRAF (B-rapidly accelerated fibrosarcoma) is a proto-oncogene critical to the MAPK/ERK signaling pathway, which transmits signals from the cell surface to the nucleus, influencing cell proliferation and survival.<sup>1,2</sup> Mutations in the BRAF gene, especially the V600E substitution, occur in approximately 50% of melanoma cases, significantly elevating mortality by promoting cancer cell growth and fostering resistance to treatments.<sup>3,4</sup> Although three generations of selective BRAF inhibitors have been developed, their effectiveness as monotherapies is limited due to the rapid emergence of resistance, leading to relapse in most patients within a year.<sup>5</sup> Therefore, the pursuit of more effective BRAF inhibitors continues, with a focus on leveraging advancements in computational techniques to improve therapeutic outcomes.<sup>6,7</sup>

Advancements in machine learning have revolutionized kinase drug discovery, enabling faster identification of potent compounds. These methods have been particularly successful in discovering dual inhibitors for fibroblast growth factor and epidermal growth factor receptors.<sup>8–10</sup> Algorithms such as Random Forest (RF), Support Vector Regression (SVR)<sup>11,12</sup> and Extreme Gradient Boosting (XGB)<sup>13</sup> are widely applied

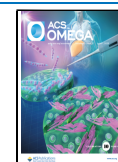
for predicting drug bioactivity, composition, and molecular properties.<sup>14</sup> However, their traditional reliance on single-model predictions often results in high generalization errors. To address this, stacking methods and large-scale ensemble learning have been adopted, integrating outputs from multiple models to improve accuracy.<sup>15</sup> In the realm of molecular property prediction, graph neural networks (GNNs) have made significant strides by modeling structural dependencies through architectures like message-passing neural networks.<sup>16–20</sup> Modern approaches prioritize end-to-end trainable models that extract feature representations directly, making them highly effective when molecular 3D conformations are not well-defined.<sup>21,22</sup> Notably, Transformer models with graph attention mechanisms excel in capturing interactions between both nearby and distant atoms.<sup>23–26</sup> Concurrently, predicting

**Received:** November 18, 2024

**Revised:** March 8, 2025

**Accepted:** March 13, 2025

**Published:** March 21, 2025



properties that rely on molecular 3D conformations, such as quantum mechanical characteristics, has emerged as a key area of research.<sup>27</sup> Additionally, pretrained models that derive rich molecular representations from large unlabeled data sets have demonstrated a remarkable ability to apply this knowledge to specialized tasks, often achieving superior results compared to traditional supervised learning approaches.<sup>28,29</sup>

Despite progress in machine learning for BRAF inhibitor screening, key challenges remain. Traditional molecular fingerprinting combined with conventional models often yields shallow feature representations, failing to capture the complexity of inhibitor molecules and leading to overfitting and low predictive accuracy. Additionally, single-model approaches, whether traditional or deep learning-based, suffer from poor generalization, limiting their ability to address the diverse characteristics of BRAF inhibitors. To address the aforementioned challenges, our research has successfully overcome these obstacles through the following innovative contributions: (1) Ensemble Learning Framework: BRAFPred combines multiple machine learning and deep learning algorithms through stacking, enhancing stability and generalization. (2) Enhanced Feature Generation: Integrating 12 PaDeL molecular descriptors with sequence-based outputs from Chemprop and FG-BERT enriches feature sets, improving prediction accuracy. (3) Superior Predictive Performance: BRAFPred surpasses state-of-the-art methods in accuracy and error rates, showcasing its robust predictive capabilities. (4) Thorough Validation: Benchmarking and ablation studies confirm the model's generalization strength and the significance of each component.

## 2. MATERIALS AND METHODS

**2.1. Benchmark Data Sets.** This study utilizes the data set from Syahid et al.,<sup>8</sup> derived primarily from the ChEMBL database, focusing on compounds targeting BRAF proteins.<sup>30</sup> Chemical features are represented in the standardized isomer SMILES (Simplified Molecular Input Line Entry System) format.<sup>31</sup> As shown in Table 1, entries lacking SMILES or IC<sub>50</sub>

**Table 1. Dataset Sample Distribution**

ChEMBL ID	Training set or Test set	Number of Data set
ChEMBL5145	Training set	2697
ChEMBL5145	Blind test set	1157

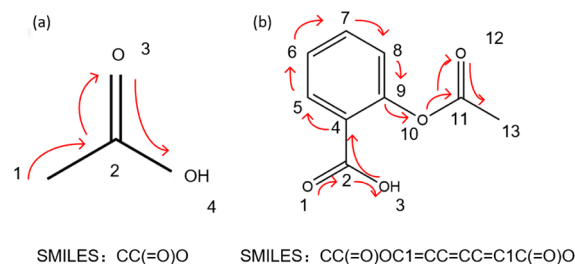
values were excluded. Only compounds with numerical IC<sub>50</sub> values were retained, converted to molar units, and transformed into pIC<sub>50</sub> values using the negative logarithm. Compounds with a molecular weight over 700 Da or LogP greater than 8 were also excluded, as they are unlikely to penetrate cell membranes.<sup>32</sup> The final data set comprises 3854 compounds, divided into two main parts: the training set (2697 compounds) and the blind test set (1157 compounds). The training set was further split using an 8:2 ratio, with the computed results rounded to create the training subset (2157 compounds) and the validation subset (540 compounds),<sup>8</sup> ensuring the robustness of the model's performance.

**2.2. Molecular Descriptors.** Molecular descriptors, also known as molecular fingerprints, are popular in drug discovery and virtual screening for their simplicity, speed, and effective performance in substructure and similarity searches. This study used PaDeL-Descriptor software to generate fingerprints from canonical isomeric SMILES, which offers efficient storage and

processing compared to 3D structures.<sup>29</sup> The use of canonical isomeric SMILES ensures unique identification of compounds, capturing stereochemical details. Fingerprints indicate the presence or absence of specific chemical substructures and are computed after standardizing tautomers, nitro compounds, and removing salts.

We utilized 12 types of molecular fingerprints to extract features from input data, converting each sample into multiple one-dimensional vectors of 0s and 1s, forming a sparse matrix. The fingerprints used include AtomPairs2D, AtomPairs2D-Count (780D),<sup>33</sup> CDK, CDK Extended, CDK Graph Only (1024D),<sup>34</sup> Estate (79D),<sup>35</sup> KlekotaRoth, KlekotaRothCount (4860D),<sup>36</sup> and other open-source fingerprints such as MACCS (166D),<sup>37</sup> PubChem (881D),<sup>38</sup> Substructure (307D), and SubstructureCount (307D).<sup>29</sup>

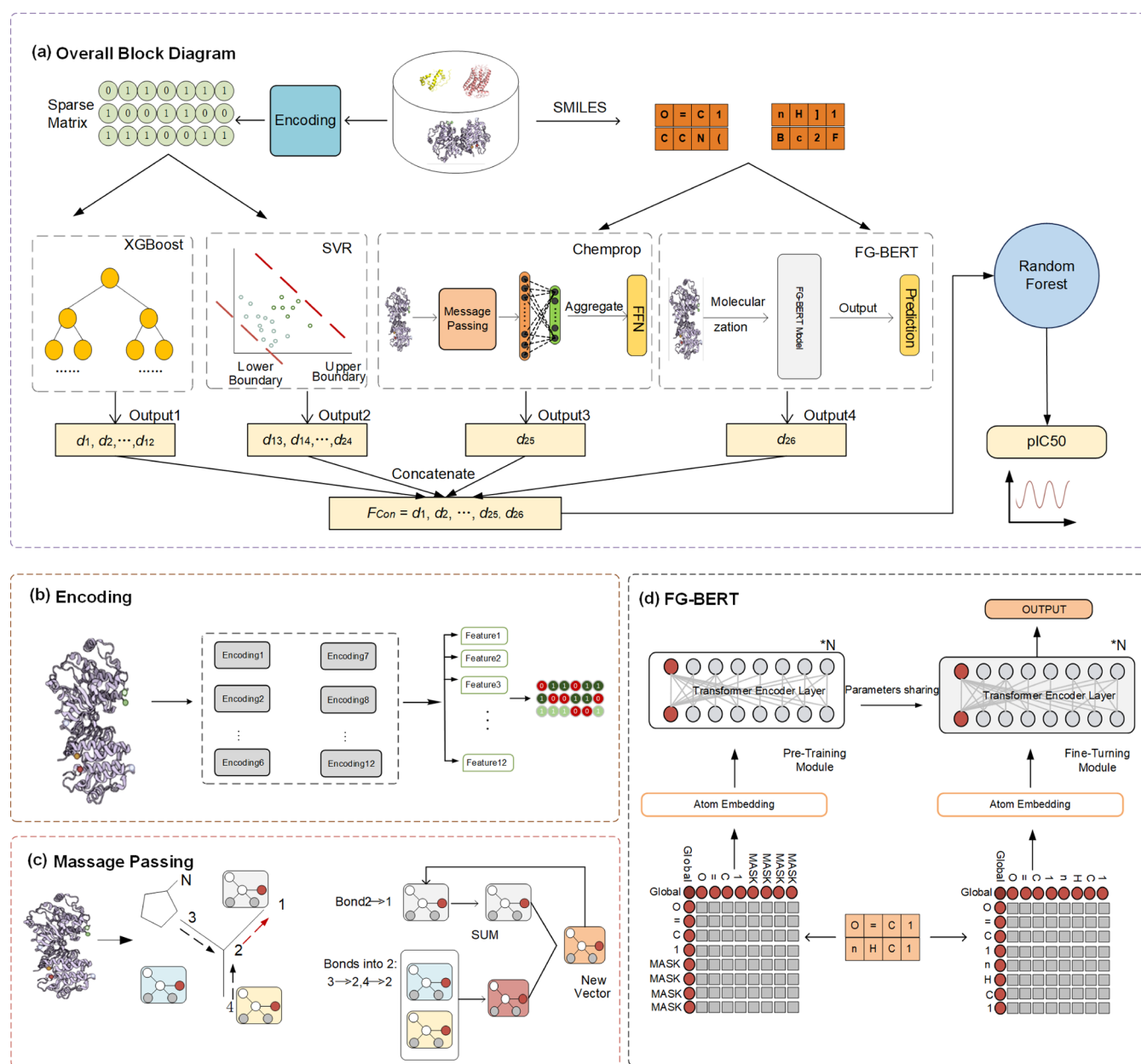
**2.3. Molecular Representation and Embeddings from FG-BERT.** Arthur Weininger et al. developed the SMILES format, which translates complex chemical structures into simple, one-dimensional strings, facilitating molecular modeling and property prediction.<sup>31</sup> For a detailed overview of different molecular representations, including structure, name, and SMILES, refer to Figure 1.



**Figure 1.** Examples of the process involved in generating SMILES representations. (a) acetic acid and (b) aspirin.

Each SMILES string encodes a molecule using elemental symbols (e.g., 'C' for carbon, 'c' for aromatic carbon) and bond types (e.g., '=' for double bonds, '#' for triple bonds), capturing structural features like branches and rings for efficient computational analysis. For example, acetic acid is represented as "CC(=O)O", where '(=O)' indicates a double bond between carbon and oxygen. Similarly, aspirin is denoted as "CC(=O)OC1=CC=CC=C1C(=O)O", with '=' representing double bonds and the number indicating the construction direction of the ring. This simplicity and versatility make SMILES a widely adopted standard in chemical informatics and drug discovery.<sup>39</sup>

Recent advances in pretraining models for molecular property prediction have significantly improved the extraction of representations from large, unlabeled data sets using contrastive and masked language learning, outperforming traditional methods. For example, K-BERT utilizes pretraining on atomic and molecular features to extract chemical insights from SMILES strings. Mole-BERT, with a VQ-VAE-based encoder, encodes atoms into discrete values, expanding atomic vocabulary and reducing disparities between common and rare atoms, thus improving representation accuracy.<sup>28</sup> Expanding on these, Wang et al. introduced FG-BERT, a self-supervised model that captures molecular representations through functional groups, enhancing prediction accuracy and resolving smoothing issues seen in the BERT-based model. FG-BERT extracts embeddings by converting inputs into a 1024D table



**Figure 2.** Architecture of the proposed BRAFPred model. (a) Overall block diagram, (b) feature encoding methods, (c) message passing model, and (d) FG-BERT structure.

with mask information, applying atom embedding, and integrating functional group data through pretraining and fine-tuning to generate molecular graph representations.<sup>40</sup>

**2.4. Embeddings from Chemprop.** Yang et al. introduced Chemprop, a machine learning toolkit for predicting molecular properties using the message passing neural network (MPNN) framework.<sup>41</sup> Chemprop represents molecular structures as graphs, with atoms as nodes and bonds as edges, and employs a message-passing mechanism to learn molecular features. SMILES strings are converted into graphs using RDKit, encoding atomic properties (such as atomic number, degree, formal charge, hydrogen count, hybridization state, aromaticity, and scaled atomic mass) into node features. Bond features are defined by bond type, conjugation, ring status, and stereochemistry. The directed MPNN (D-MPNN) enhances information flow by using directed edges, combining

atomic and bond features for more effective molecular representation, improving property prediction.

**2.5. Performance Evaluation Methods.** The coefficient of determination ( $R^2$  or  $Q^2$ ) and mean absolute error (MAE) are key metrics for assessing the performance of baseline and BRAFPred models. High predictive accuracy is reflected by lower MAE and higher  $R^2$  or  $Q^2$  values. These metrics are computed for the training set (MAE,  $R^2$ ), cross-validation set (MAE,  $Q^2$ ), and external blind test set (MAE,  $Q^2$ ). Additionally, to identify the best molecular fingerprint features, we calculated the Pearson correlation coefficient (PCC) and root-mean-square error (RMSE) for different molecular fingerprints using the same model on both the training data set and the cross-validation data set. By combining these metrics, we can identify the molecular fingerprint features that perform best, allowing us to select more representative features for evaluation in external blind test set.



$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (1)$$

$$R^2 \text{ or } Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

$$\text{PCC} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (4)$$

where  $y_i$  represents the experimental  $\text{pIC}_{50}$  values and  $\hat{y}_i$  represents the corresponding predicted values.  $\bar{y}$  denotes the mean of the experimental values,  $\bar{\hat{y}}$  represents the average of predicted values and  $n$  is the total number of molecules in the data set.

MAE quantifies prediction errors, with lower values indicating higher accuracy, while  $R^2$  and  $Q^2$  measure the proportion of variance in experimental  $\text{pIC}_{50}$  values explained by the model, with values closer to 1 suggesting better model fit and predictability. Generally,  $R^2$  or  $Q^2 > 0.6$  indicates strong model performance,  $R^2 - Q^2 \leq 0.2$  reflects a good fit, and an MAE below 1 suggests high predictive accuracy.

### 3. THE PROPOSED BRAFPRED MODEL

**3.1. Overview of the BRAFPred Framework.** The BRAFPred framework integrates a 26-dimensional predictive feature vector derived from XGB, SVR, Chemprop, and FGBERT, comprising four key components, as illustrated in Figure 2a–d. Initially, 12 molecular fingerprinting techniques are applied as inputs to both the SVR and XGB models, generating outputs denoted as  $F_{\text{output1}}$  and  $F_{\text{output2}}$ , respectively. In parallel, Chemprop processes SMILES strings to produce additional predictive features, referred to as  $F_{\text{output3}}$ , while FGBERT, utilizing the same SMILES strings as input, generates predictive features labeled as  $F_{\text{output4}}$ . Finally, the outputs  $F_{\text{output1}}$ ,  $F_{\text{output2}}$ ,  $F_{\text{output3}}$ , and  $F_{\text{output4}}$  are concatenated to form a comprehensive feature set, designated as  $F_{\text{Com}} = \{d_1, d_2, \dots, d_{26}\}$ , which is subsequently used as input to a RF network within the stacked ensemble learning framework, yielding the final prediction output.

**3.2. XGB.** XGB utilizes extreme gradient boosting through sequential decision trees to enhance predictions. It is a tree-based ensemble algorithm that improves prediction performance by employing both gradient boosting and regularization techniques. The hyperparameters for XGB were configured as follows:  $\text{gamma} = 0$ ,  $\text{reg\_lambda} = 1$ ,  $\text{reg\_alpha} = 0$ ,  $\text{max\_depth} = 6$ ,  $\text{n\_estimators} = 100$ , and  $\text{learning\_rate} = 0.3$ . The input of this XGB consists of 12 different feature matrices derived from SMILES representations of samples (refer to Figure 2b for details on the encoding process). The outputs consist of 12 predicted features, collectively denoted as  $F_{\text{output1}} = \{d_1, d_2, \dots, d_{12}\}$ .

**3.3. SVR.** SVR predicts outcomes by fitting a linear hyperplane within a prescribed margin of tolerance, thereby striving to achieve an equilibrium between model complexity and predictive precision. In the present investigation, the parameter  $C$  as meticulously optimized employing a grid search algorithm coupled with 5-fold cross-validation, with its values

spanning from 1 to 10. The application of SVR was extended across 12 distinct molecular fingerprint features (refer to Figure 2b for encoding details). As a pivotal regression technique within the ensemble methodology framework, SVR processed 12 encoded feature vectors, which were derived from SMILES strings. These vectors were subsequently transformed into multidimensional representations through the application of molecular fingerprinting techniques. Consequently, this processing yielded a collection of predictive outputs, denoted as  $F_{\text{output2}} = \{d_{13}, d_{14}, \dots, d_{24}\}$ .

**3.4. Chemprop.** Chemprop is a sophisticated toolkit for predicting molecular properties, leveraging the D-MPNN architecture. It is expertly crafted to encode molecular structures with high fidelity and predict a diverse range of chemical properties. The structural blueprint of Chemprop is outlined in Figure 2a. This segment harnesses the D-MPNN to delineate molecules as graphs, with atoms serving as nodes and bonds functioning as directed edges. This graphical representation empowers the model to discern both local and global structural intricacies, thereby capturing a comprehensive view of the molecular architecture. Chemprop processes SMILES strings, parsing each into a molecular graph where nodes correspond to atoms and directed edges represent bonds, as shown in Figure 2c. For each node  $v$ , the initial feature vector  $\{x_v | v \in V\}$  is derived from atomic properties, including atomic number, formal charge, chirality, hydrogen count, hybridization, aromaticity, and a scaled encoding of atomic mass. For each bond (edge)  $e$ , the initial feature vector  $\{e_{vw} | v, w \in E\}$  incorporates bond type, conjugation, ring membership, and stereochemistry. The directed edge features  $e_{vw}^d$  are then initialized by concatenating the atom feature  $x_v$  with the undirected bond feature  $e_{vw}$ , ensuring comprehensive encoding of both atomic and bonding information, as described in eq 5:

$$e_{vw}^d = \text{cat}(x_v, e_{vw}) \quad (5)$$

Building on the initialized directed edge features, D-MPNN updates these features by transforming the initial directed edge feature  $e_{vw}^d$  through a neural network layer with a learnable weight matrix  $W_e$ , generating a hidden edge feature  $h_{vw}^0$ . A nonlinear activation function  $\tau$  is applied to capture complex molecular interactions. Utilizing the initialized directed edge features obtained in the preceding step as a foundation, the D-MPNN refines these features by transforming the initial directed edge feature  $e_{vw}^d$  through a neural network layer equipped with a trainable weight matrix  $W_e$ . This transformation yields a hidden edge feature  $h_{vw}^0$ , where a nonlinear activation function  $\tau$  is subsequently applied, as described in eq 6:

$$h_{vw}^0 = \tau(W_e \cdot e_{vw}^d) \quad (6)$$

This process is designed to capture intricate molecular interactions, thereby enriching the representational capacity of the model.

After initializing these hidden edge features, Chemprop performs a series of  $t$  message-passing iterations, during which the hidden feature  $h_{vw}^{t+1}$  is iteratively updated by aggregating information from neighboring nodes  $\mathcal{N}(v)$ , excluding the central node  $w$ . This update process uses a learnable weight matrix  $W_h$  for each neighboring feature  $h_{vk}^t$  and a bias term  $b$ ,

allowing the network to encode local structural details crucial for molecular property prediction, as illustrated in eq 7:

$$h_{vw}^{t+1} = h_{vw}^t + \sum_{k \in \{N(v)/w\}} (W_h \cdot h_{vk}^t) + b \quad (7)$$

Upon completing the message-passing iterations, the atomic embeddings are aggregated, integrating information from neighboring atoms. This process enables D-MPNN to effectively capture local structural features, resulting in an enhanced molecular representation. Subsequently, all atomic embeddings are combined into a unified molecular embedding, which is passed through a feedforward neural network with two hidden layers, each containing 300 neurons, to learn the target molecular properties. The output generated by Chemprop for a sample is denoted as  $F_{output3} = \{d_{25}\}$ . For regression tasks, RMSE is employed as the loss function. This architecture efficiently captures intricate relationships within molecular structures, significantly improving predictive performance.

**3.5. FG-BERT.** FG-BERT is designed on the foundation of the BERT model, harnessing its robust language representation capabilities and extending them into the realm of molecular graph learning tasks. Initially, BERT was crafted for natural language processing, incorporating two pivotal pretraining tasks: the Masked Language Model and the Next Sentence Prediction task. These tasks empower BERT to capture profound semantic correlations within and across words within a contextual framework. FG-BERT integrates these fundamental attributes of BERT while tailoring the model specifically to molecular graph structures, thereby enabling it to proficiently interpret and predict chemical information within molecules.

The input to the FG-BERT framework comprises SMILES sequences, which serve as representations of molecules. To facilitate the representation of atoms and functional groups (FGs) within a molecular graph, an atom dictionary is constructed, grounded on the frequency of diverse atoms within a pretrained molecular corpus. Common atoms, such as hydrogen ([H]), carbon ([C]), and nitrogen ([N]), are denoted by their respective element symbols, whereas infrequent atoms are labeled as [UNK]. Furthermore, a super node, designated as [GLOBAL], is appended to the molecular graph to bolster downstream tasks, and [MASK] is utilized to signify masked FGs for pretraining purposes. The FG-BERT framework encompasses three integral components: an embedding layer, a transformer layer, and a pretraining/prediction head, as illustrated in Figure 2d and detailed in eqs 8–14.

$$q_i = W_q x_i \quad (8)$$

$$k_i = W_k x_i \quad (9)$$

$$v_i = W_v x_i \quad (10)$$

$$s_{ij} = \frac{\text{dot}(q_i, k_j)}{\sqrt{d_k}}, j \in N_i \quad (11)$$

$$a_{ij} = \text{softmax}(s_{ij}) \quad (12)$$

$$m_i = \sum_{j \in N_i} a_{ij} v_j \quad (13)$$

$$M_i = W_o \text{concat}(m_i^1, m_i^2, \dots, m_i^K) \quad (14)$$

This meticulously crafted architecture allows FG-BERT to adeptly learn and represent complex molecular structures, paving the way for enhanced chemical information interpretation and prediction capabilities.

The embedding layer maps the atom sequence  $w = (a_1, a_2, a_3, \dots, a_N)$  to an embedding vector  $x = (x_1, x_2, x_3, \dots, x_N)$  via an embedding matrix  $D \in \mathbb{R}^{V \times d_{\text{model}}}$ , where each  $x_i \in \mathbb{R}^{d_{\text{model}}}$ ,  $V$  represents the vocabulary size, and  $d_{\text{model}}$  is the dimensionality of the embedding vectors. In the transformer layer, each node aggregates information from its neighbors through attention mechanisms. The query, key, and value vectors for each node  $i$  are computed as  $q_i, k_i$  and  $v_i$ , respectively. The attention score between nodes  $i$  and  $j$  is calculated as  $S_{ij}$  and normalized using a softmax function to yield  $a_{ij}$ . The message for node  $i$  is then given by  $m_i$ . In a multihead attention mechanism, this process is repeated independently across  $H$  heads, after which the outputs are concatenated and linearly transformed as  $M_i$ , where  $W_o$  is a learnable weight matrix. Furthermore, a feed-forward network sublayer, along with connectivity and layer normalization mechanisms, is applied within the transformer layer. The transformer layer in FG-BERT is applied multiple times according to the specified number of layers. The output obtained from a sample processed by FG-BERT is designated as Output4, represented by  $F_{output4} = \{d_{26}\}$ . For regression tasks, FG-BERT's prediction head consists of two fully connected layers, with the Gaussian Error Linear Unit used as the activation function during pretraining and Leaky ReLU for downstream regression tasks, using RMSE as the loss function. Final hyperparameter settings, selected for optimal validation, are detailed in Table 2.

**Table 2. Hyperparameters of the FG-BERT Pretraining Model**

Hyperparameters	Values
Layers	6
Heads	4
Embedding size	256
FFN size	512
Learning rate	$10^{-4}$
Dropout	0.1
Model Params	~3.2M

**3.6. RF Stacked Ensemble.** Stacked ensemble learning enhances predictive performance by effectively combining multiple base models, leveraging their individual strengths while improving generalization and robustness across various machine learning tasks. In this study, we explored various stacking ensemble methodologies, including RF, Gradient Boosting, ExtraTree, Multilayer Perceptron, K-Nearest Neighbors, and Linear Regression, ultimately selecting RF as the framework for our model. This ensemble learning technique integrates predictions from multiple decision trees to enhance accuracy and mitigate overfitting. Employing the aforementioned base models, including XGB, SVR, Chemprop, and FG-BERT, we generated  $F_{output1}$ ,  $F_{output2}$ ,  $F_{output3}$ , and  $F_{output4}$ , culminating in a consolidated 26-dimensional feature vector, denoted as  $F_{Con} = \{d_1, d_2, \dots, d_{26}\}$ . This feature vector  $F_{Con}$  served as the input for the ensemble methodology to yield the final prediction results. To optimize the hyperparameters of the RF model, we implemented a grid search approach coupled with

5-fold cross-validation. This optimization process encompassed a systematic exploration of various combinations of hyperparameters, specifically max\_depth (10, 20, and 50), n\_estimators (10 and 100), and max\_features (2, 3, 4, and 5), with the aim of achieving superior predictive performance.

## 4. RESULTS AND DISCUSSIONS

**4.1. Comparative Analysis of Molecular Fingerprints and Sequence-Based Features.** In this section, we evaluate the predictive performance of BRAFPred models using two feature types: molecular fingerprints and sequence-based representations. We first assess traditional machine learning regressors (XGB and SVR) with molecular fingerprints on the training set and through 10-fold cross-validation. Next, we examine deep learning models (FG-BERT and Chemprop) using sequence features from SMILES strings. Ultimately, ablation studies delve into the respective contributions of FG-BERT and Chemprop when integrated with molecular fingerprints, underscoring the advantageous synergies achieved through the amalgamation of traditional and deep learning paradigms.

**4.1.1. Performance of Molecular Fingerprints with Traditional Machine Learning on the Training Set.** The predictive performance of 24 models on the training set, combining 12 molecular fingerprints with XGB and SVR algorithms, is presented in Table 3 and Figure 3. Notably, the SVR\_CDK

**Table 3. Prediction Performance of XGB and SVR on the Training Set**

Feature	XGB		SVR	
	Q <sup>2</sup>	MAE	Q <sup>2</sup>	MAE
AtomPairs2D	0.501	0.668	0.473	0.696
AtomPair2DCount	0.635	0.580	0.592	0.627
CDK	0.698	0.525	<b>0.743</b>	<b>0.480</b>
CDKextended	0.681	0.540	0.731	0.499
CDKgraphonly	0.662	0.559	0.628	0.584
EState	0.537	0.675	0.501	0.683
KlekotaRoth	0.716	0.525	0.612	0.615
KlekotaRothCount	0.682	0.545	0.622	0.609
MACCS	0.581	0.618	0.614	0.594
PubChem	0.714	0.516	0.696	0.537
Substructure	0.559	0.658	0.516	0.664
SubstructureCount	0.632	0.588	0.583	0.620

model achieved the best results on the training set, with R<sup>2</sup> of 0.743, MAE of 0.480, RMSE of 0.679, and PCC of 0.862, using an 80/20 training-testing split. At the same time, we tested various regressors on the training set, with detailed descriptions provided in Text S1 and the results shown in Tables S1–S5. The experimental results demonstrate the robust performance of SVR and XGB regressors. These outcomes highlight the strong predictive capability of the SVR\_CDK model among the 12 molecular fingerprint features tested on the same regressor. Figure 3 presents baseline comparisons, including RMSE and PCC values for both XGB and SVR models.

**4.1.2. Evaluation of Molecular Fingerprints with Traditional Machine Learning on 10-Fold Cross-Validation.** Table 4 and Figure 4 show the results of 10-fold cross-validation for 24 feature sets derived from traditional molecular fingerprint methods. The SVR\_CDK and SVR\_CDKextended models

achieved the highest Q<sup>2</sup> (0.771 and 0.770), lowest MAE (0.468 and 0.467), lowest RMSE (0.652 and 0.654), and highest PCC (0.878). These findings demonstrate that traditional molecular fingerprints significantly improve the predictive accuracy of BRAFPred models.

**4.1.3. Assessing Sequence-Based Deep Learning Models on the Training Set.** In this study, we evaluated the performance of two sequence-based deep learning models, FG-BERT and Chemprop. Table 5 shows the 10-fold cross-validation results on the training set. Chemprop demonstrated superior predictive performance, achieving R<sup>2</sup> = 0.782 and MAE = 0.449. The FG-BERT model also performed well, with R<sup>2</sup> = 0.763 and MAE = 0.471. As shown in Figure 5, training loss for Chemprop stabilized after 100 epochs, while FG-BERT required 400 epochs. These results indicate that sequence-based deep learning models can perform on par with traditional handcrafted molecular fingerprint models.

**4.1.4. Ablation Study of FG-BERT for Enhanced Sequence Feature Integration.** To validate the reliability of the BRAFPred model, we conducted ablation experiments, as summarized in Table 6. When 12 different molecular encodings were input into the XGB model, the results were stacked as new features and processed by RF, resulting in an R<sup>2</sup> of 0.722 and an MAE of 0.512. A similar approach with the SVR regressor achieved an R<sup>2</sup> of 0.773 and an MAE of 0.482. These results emphasize the positive role of traditional molecular fingerprinting in enhancing the design and performance of BRAFPred models. FG-BERT alone, processing SMILES sequences, achieved an R<sup>2</sup> of 0.763 and an MAE of 0.472, demonstrating strong performance. Integrating FG-BERT with 12 features from the XGB regressor into the RF network improved results, achieving an R<sup>2</sup> of 0.785 and an MAE of 0.457, indicating the positive contribution of XGB and FG-BERT. Further stacking 25 outputs from XGB, FG-BERT, and SVR into the RF network resulted in an R<sup>2</sup> of 0.793 and an MAE of 0.443, confirming that combining these models enhances predictive accuracy. In addition, we evaluated some ensemble regression models on the blind test set, with the results listed in Table S6 and corresponding analysis provided in Text S2. The experimental findings indicate that RF achieved the best performance.

**4.1.5. Ablation Study of Chemprop for Enhanced Sequence Feature Integration.** We evaluated the impact of integrating the Chemprop module within the BRAFPred framework. As shown in Table 6, Chemprop, after processing SMILES sequences, achieved an R<sup>2</sup> of 0.782 and an MAE of 0.449, demonstrating strong predictive capabilities. When Chemprop predictions were combined with outputs from RF\_XGB, SVR, and FG-BERT in the RF network, BRAFPred reached its highest performance, with an R<sup>2</sup> of 0.818 and an MAE of 0.417, as detailed in Table 6. This integration significantly improved predictive accuracy, highlighting Chemprop's contribution to the BRAFPred model.

In summary, the ablation studies and comparative analyses emphasize the importance of integrating diverse molecular fingerprints and advanced models like FG-BERT and Chemprop. The results demonstrate that combining traditional fingerprinting methods with deep learning models significantly enhances the accuracy and reliability of BRAFPred positioning it as a robust tool for BRAF inhibitor prediction.

**4.2. Performance Comparison of BRAFPred with Existing Predictors.** To assess BRAFPred's performance against existing models, we compared various deep learning



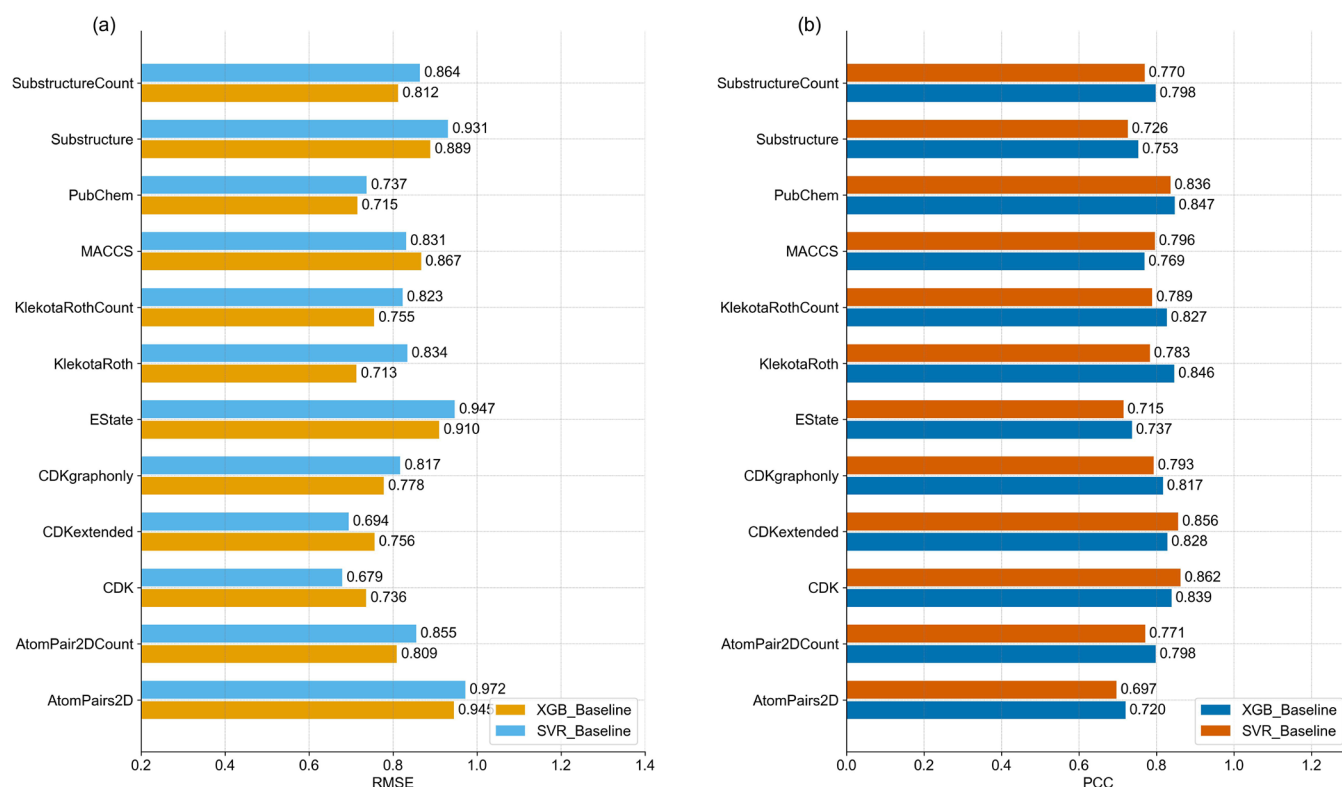


Figure 3. Baseline comparisons. (a,b) RMSE and PCC of XGB and SVR on the training data.

Table 4. Prediction Performances of XGB and SVR on Training Data Using 10-Fold Cross-Validation

Features	XGB		SVR	
	Q <sup>2</sup>	MAE	Q <sup>2</sup>	MAE
AtomPairs2D	0.618	0.606	0.573	0.643
AtomPair2DCount	0.688	0.551	0.572	0.659
CDK	0.715	0.516	0.771	0.468
CDKextended	0.707	0.524	0.770	0.467
CDKgraphonly	0.688	0.537	0.708	0.528
EState	0.558	0.668	0.525	0.683
KlekotaRoth	0.729	0.520	0.759	0.490
KlekotaRothCount	0.735	0.516	0.720	0.529
MACCS	0.646	0.583	0.657	0.576
PubChem	0.727	0.516	0.729	0.514
Substructure	0.607	0.632	0.600	0.623
SubstructureCount	0.669	0.575	0.476	0.747

models on the same training set, as shown in Figure 6. Competitors included graph attention network (GAT),<sup>42</sup> Chemprop and its variants (Chemprop\_RDKit<sup>43</sup> and Chemprop\_RF,<sup>43</sup> FG-BERT<sup>40</sup> and Stack\_BRAF.<sup>8</sup> The results indicated that GAT had the lowest performance ( $R^2 = 0.692$ , MAE = 0.516). Chemprop\_RDKit achieved an  $R^2$  of 0.763 and MAE of 0.473, Chemprop\_RF reached an  $R^2$  of 0.753 and MAE of 0.481, and Chemprop achieved an  $R^2$  of 0.782 and MAE of 0.449. FG-BERT yielded an  $R^2$  of 0.763 and MAE of 0.471, while Stack\_BRAF reached an  $R^2$  of 0.796 and MAE of 0.442. Based on these comparisons, Chemprop was selected for integration into BRAFPred over its variants due to its superior performance. Notably, BRAFPred outperformed Stack\_BRAF in  $pIC_{50}$  prediction, highlighting its enhanced predictive capability.

**4.3. Model Interpretation.** The predictive feature importance of the BRAFPred model is analyzed using Shapley Additive Projection (SHAP) values, which help identify key molecular fingerprints influencing  $pIC_{50}$  predictions. SHAP, rooted in game theory, quantifies the contribution of each feature, effectively distinguishing between positive and negative influences on the model's output. This method provides both local and global interpretability, allowing us to explain individual molecular contributions as well as the overall impact of different features on model performance. Positive SHAP values indicate features that enhance  $pIC_{50}$  predictions, while negative values suggest a decrease in predictive strength.

To better understand the model's behavior, Figure 7 illustrates the diagnostic utility of SHAP values by showing how variations in predictive features affect BRAFPred's outputs (with red indicating high impact and blue indicating low impact). The analysis reveals that the top five most important features in BRAFPred include Chemprop (mean  $|SHAP| = 0.24$ ), SVR\_CDKextended (mean  $|SHAP| = 0.16$ ), SVR\_CDK (mean  $|SHAP| = 0.13$ ), FG-BERT (mean  $|SHAP| = 0.13$ ), and XGB\_KlekotaRothCount (mean  $|SHAP| = 0.075$ ). The analysis highlights that features derived from deep learning models, particularly Chemprop and FG-BERT, play a significant role in driving accurate predictions. The average SHAP values reveal that Chemprop and FG-BERT, along with other impactful features like SVR\_CDKextended, SVR\_CDK, and XGB\_KlekotaRothCount, contribute substantially to the  $pIC_{50}$  predictions. Notably, three of the top five features—SVR\_CDKextended, SVR\_CDK, and XGB\_KlekotaRothCount—are aligned with the baseline models, highlighting the BRAFPred framework's ability to integrate baseline features with novel deep learning-based molecular representations. In contrast, features such as XGB\_MACCS, SVR\_SubstructureCount, XGB\_EState, and SVR\_EState show minimal



Figure 4. Model comparisons with 10-fold cross-validation. (a,b) RMSE and PCC values of XGB and SVR on the training set.

Table 5. Prediction Performances of FG-BERT and Chemprop on the Training Set Using 10-Fold Cross-Validation

Model	$R^2$	MAE
FG-BERT	$0.763 \pm 0.022$	$0.471 \pm 0.023$
Chemprop	$0.782 \pm 0.020$	$0.449 \pm 0.021$

impact, indicating that the accuracy of BRAFPred heavily relies on its most predictive features, with a notable emphasis on those derived from deep learning architectures.

These findings highlight the critical role of novel molecular representations, particularly those leveraging deep learning frameworks and BERT-based functional group masking strategies, in improving the predictive performance of BRAFPred. By integrating traditional baseline features with

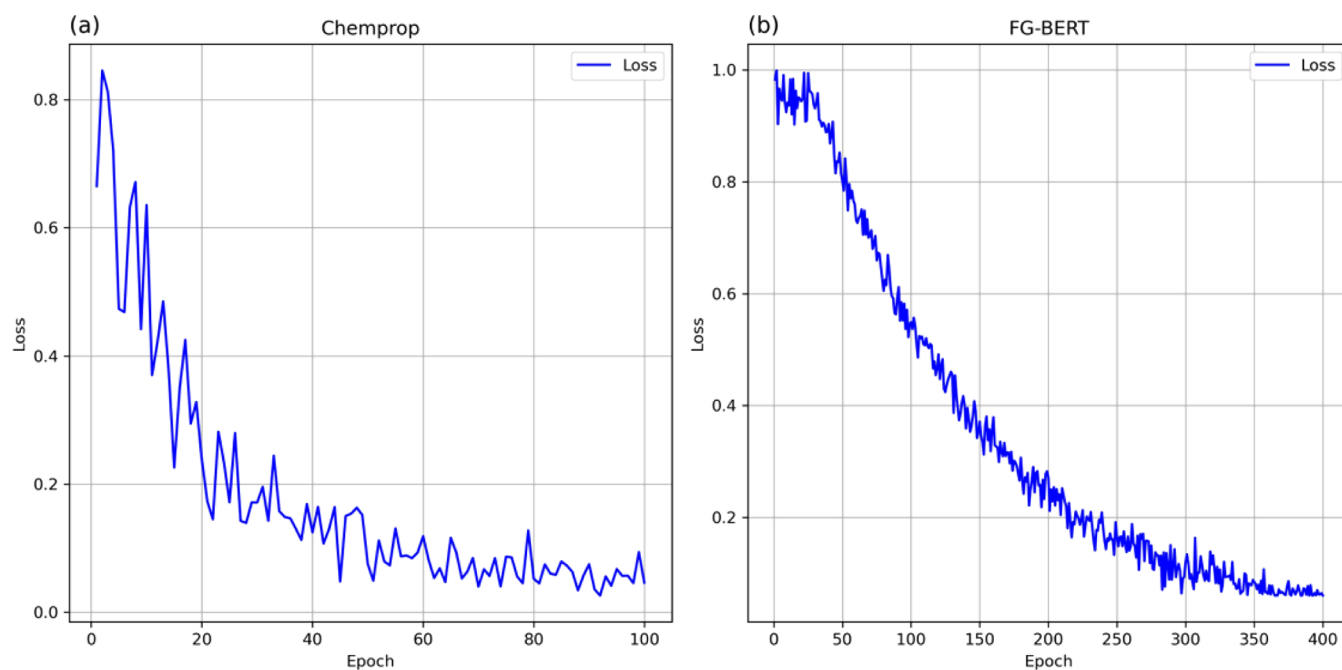


Figure 5. Training loss curves on training set. (a) Chemprop model and (b) FG-BERT model.



**Table 6. Ablation Study Results of the BRAFPred Model on Training Data<sup>a</sup>**

Model	R <sup>2</sup>	MAE
RF_XGB	0.722 ± 0.004	0.512 ± 0.002
RF_SVR	0.773 ± 0.005	0.482 ± 0.004
RF_XGB+FG-BERT	0.785 ± 0.003	0.457 ± 0.002
RF_XGB+SVR+FG-BERT	0.793 ± 0.004	0.443 ± 0.003
<b>BRAFPred</b>	<b>0.818 ± 0.004</b>	<b>0.417 ± 0.004</b>

<sup>a</sup>RF\_XGB is the result of 10-fold cross validation using output 1 as new features on random forest model. RF\_SVR is the result of 10-fold cross validation using output 2 as new features on random forest model. RF\_XGB+FG-BERT is the result of 10-fold cross validation using output 1 and 4 as new features on random forest model. RF\_XGB+SVR+FG-BERT is the result of 10-fold cross validation using outputs 1, 2, and 4 as new features on random forest model. BRAFPred is the 10-fold cross validation result of the model BRAFPred on the training set.

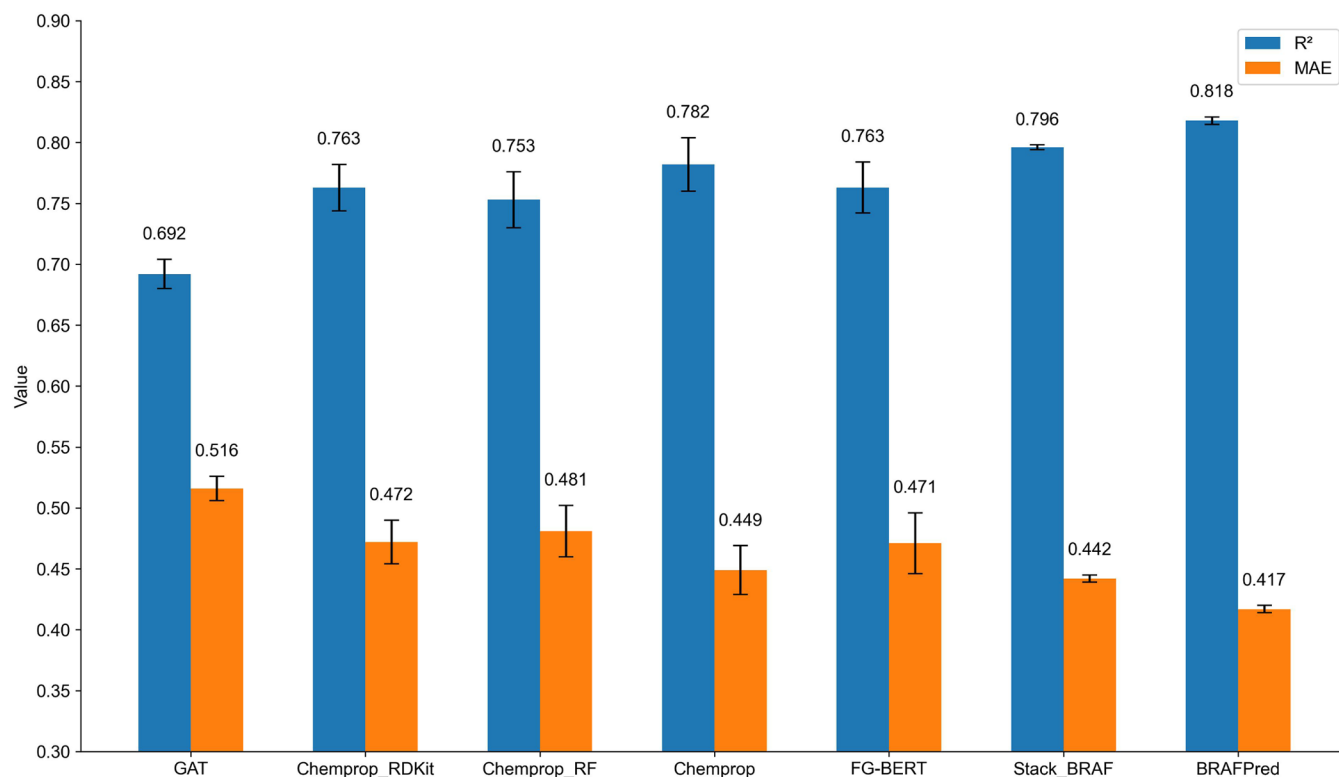
these advanced representations, BRAFPred achieves superior accuracy and robustness in pIC<sub>50</sub> prediction. This hybrid approach underscores the synergy between deep learning techniques and baseline models, which drives the exceptional performance of the framework. Notably, these advanced methodologies have led to a improvement in predictive performance metrics, increasing R<sup>2</sup> from 0.84 (baseline) to 0.855 and reducing MAE from 0.40 to 0.383, Achieving even these modest gains on an already high-performing baseline underscores the effectiveness of the proposed approach.

**4.4. Performance Comparison of BRAFPred with SOTA Methods on the Blind Test Set.** BRAFPred's performance on training and blind test sets is shown in Figure 8. The scatter plot compares predicted pIC<sub>50</sub> values with experimentally observed values for a randomly sampled 30%

subset. Blue dots represent predictions on the blind test set, while the red line indicates the correlation between predicted and actual values. On the training set, BRAFPred achieved an R<sup>2</sup> of 0.818 and an MAE of 0.417, capturing relationships within the data. Its performance improved on the blind test set, with an R<sup>2</sup> of 0.855 and an MAE of 0.383, demonstrating strong generalization.

To further assess model robustness, Table 7 compares BRAFPred against various deep learning models on the blind test set, including GAT, Chemprop and its variants (Chemprop\_RDKit<sup>43</sup> and Chemprop\_RF,<sup>43</sup> FG-BERT<sup>40</sup> and Stack\_BRAF.<sup>8</sup> GAT showed the lowest performance (R<sup>2</sup> = 0.703, MAE = 0.511), while Chemprop\_RDKit, Chemprop\_RF, and Chemprop achieved R<sup>2</sup> values of 0.780, 0.772, and 0.803 with MAEs of 0.463, 0.467, and 0.443, respectively. FG-BERT reached an R<sup>2</sup> of 0.785 and an MAE of 0.460, while Stack\_BRAF achieved R<sup>2</sup> = 0.839 and MAE = 0.403. BRAFPred outperformed all models, including Stack\_BRAF, in pIC<sub>50</sub> prediction, confirming its robustness and suitability for drug discovery applications.

**4.5. Case Study.** In this investigation, we utilized a comprehensive data set encompassing 2,123 drugs approved by the FDA to assess the real-world efficacy of the BRAFPred model. This data set, sourced from the FDA repository, consists of meticulously scrutinized compounds that have undergone rigorous clinical trials. To ensure the integrity of our analysis, the data set was meticulously curated to exclude inorganic substances, composite compounds, and redundant entries.<sup>44</sup> Table 8 lists the top five predicted pIC<sub>50</sub> values from BRAFPred alongside predictions from four baseline models and actual pIC<sub>50</sub> values. The true pIC<sub>50</sub> value of the drug is determined by measuring its IC<sub>50</sub> value in the experimental environment, and the pIC<sub>50</sub> value is further calculated based on

**Figure 6.** Performance comparison of deep learning models on the training set.

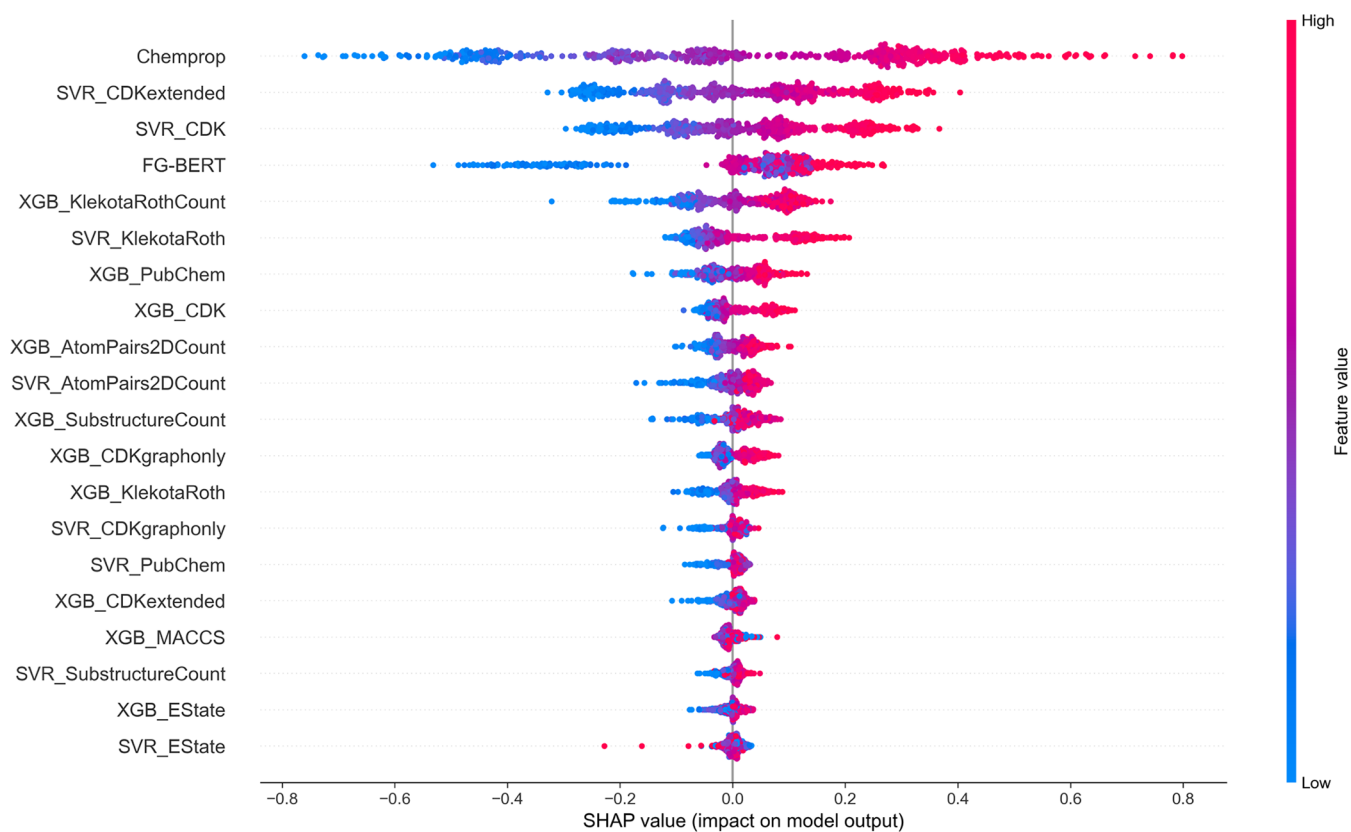


Figure 7. Feature importance of BRAFPred.

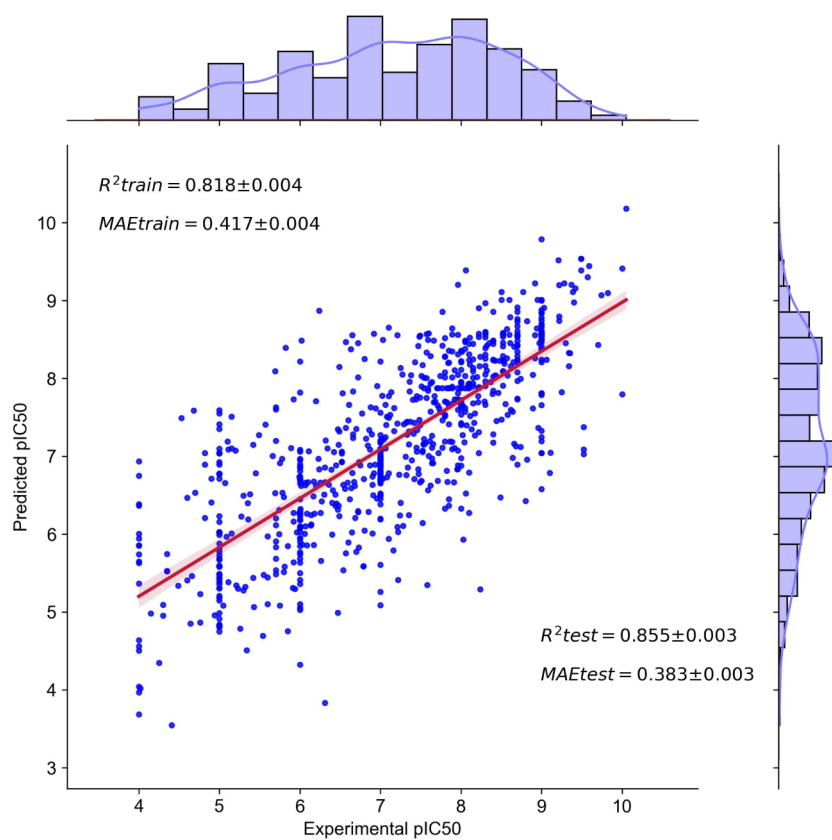


Figure 8. Performance of BRAFPred on training set and blind test set.

**Table 7. Performance Comparison of Deep Learning Models on the Blind Test Set**

Model	R <sup>2</sup>	MAE
GAT <sup>42</sup>	0.703 ± 0.012	0.511 ± 0.010
Chemprop_RDKit <sup>43</sup>	0.780 ± 0.019	0.463 ± 0.018
Chemprop_RF <sup>43</sup>	0.772 ± 0.023	0.467 ± 0.021
Chemprop <sup>41</sup>	0.803 ± 0.022	0.443 ± 0.020
FG-BERT <sup>40</sup>	0.785 ± 0.021	0.460 ± 0.025
Stack_BRAF <sup>8</sup>	0.839 ± 0.002	0.403 ± 0.003
BRAFPred	0.855 ± 0.003	0.383 ± 0.003

the IC<sub>50</sub> value.<sup>45–50</sup> BRAFPred identified ENCORA FENIB, DABRA FENIB, VEMURA FENIB, REGORA FENIB, and COBIMETINIB as the top candidates, with predicted pIC<sub>50</sub> values of 8.24, 8.02, 7.98, 7.86, and 6.85, respectively, compared to their actual values of 8.40, 9.30, 7.59, 7.72, and 8.38. BRAFPred achieved an MAE of 0.70, outperforming Chemprop and FG-BERT, which had MAEs of 0.80, 1.06, respectively.

Figure 9 illustrates the absolute differences between predicted and true values, comparing BRAFPred with other models for drugs with the highest pIC<sub>50</sub> predictions. These results highlight BRAFPred's superior performance, achieving the lowest MAE among the evaluated models. Notably, BRAFPred not only accurately identified FDA-approved inhibitors such as ENCORA FENIB,<sup>51</sup> DABRA FENIB,<sup>52</sup> and VEMURA FENIB,<sup>44</sup> which target the BRAF V600E mutation for melanoma treatment, but also demonstrated its ability to identify inhibitors targeting RAF-1 and other kinase targets. Specifically, it predicted REGORA FENIB, a multitarget kinase inhibitor approved for targeting both BRAF and RAF-1 proteins, as well as COBIMETINIB, a MEK inhibitor used in combination with BRAF inhibitors like vemurafenib to treat BRAF V600E or V600K mutation-positive metastatic melanoma.<sup>53</sup> BRAFPred's precise identification of compounds targeting diverse RAF-related proteins underscores its potential as a robust tool for ligand-based drug design. Further case study experiments are provided in Text S3. The results

emphasize the superior accuracy and stability of BRAFPred, showcasing its excellent generalization performance.

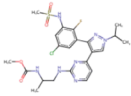
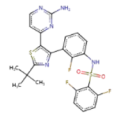
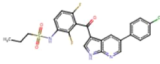
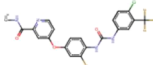
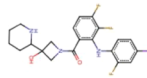
**4.6. Applicability Evaluation of BRAFPred Framework.** To further clarify the practical utility and applicability domain of the BRAFPred framework, we evaluated its performance in a novel scenario by predicting fourth-generation EGFR inhibitors.<sup>54</sup> This experiment was designed to assess whether BRAFPred could generalize beyond BRAF inhibitors and effectively identify structurally diverse kinase inhibitors, thereby demonstrating its broader applicability in early-stage drug discovery. Unlike traditional models that primarily focus on numerical IC<sub>50</sub> prediction, BRAFPred offers distinct advantages in early hit identification and compound ranking, making it particularly valuable for virtual screening campaigns.

To maintain consistency with previous evaluations, we utilized a data set composed entirely of experimentally validated EGFR inhibitors from existing research.<sup>54</sup> The training set was constructed by randomly sampling 70% of the data, ensuring a robust learning foundation. The test set comprised the remaining 30% of the EGFR data set, supplemented with an additional 500 randomly selected samples from the test set of StackBRAF. This augmentation strategy aimed to enhance molecular diversity, allowing us to assess the model's performance across a wider chemical space.

The BRAFPred framework demonstrated strong predictive performance, achieving an R<sup>2</sup> value of 0.895 and a MAE of 0.289 on the training set. For the test set, we ranked the predicted pIC<sub>50</sub> values in descending order and analyzed the top 15 predictions. Remarkably, 14 out of these 15 predictions were confirmed as EGFR inhibitors, yielding a precision of 93%. The top five predicted values and their corresponding experimental confirmations are summarized in Table 9.

To further understand the model's practical utility in drug discovery and its performance in early hit identification for discovering novel chemical scaffolds, we used the same EGFR training set strategy to test the Baseline model (StackBRAF) as well as other common machine learning models (XGB, SVR, MLP, DT, KNN) combined with 12 molecular fingerprints

**Table 8. Top Five Predicted pIC<sub>50</sub> Values for FDA-Approved Drugs Identified by Different Models**

Name	Structure	Chemprop	FG-BERT	Stack_BRAF	BRAFPred	True pIC <sub>50</sub>
ENCORA FENIB		7.94	8.42	8.19	8.24	8.40
DABRA FENIB		8.17	7.53	6.32	8.02	9.30
VEMURA FENIB		8.08	7.98	8.00	7.98	7.59
REGORA FENIB		6.94	6.70	6.26	7.86	7.72
COBIMETINIB		7.26	6.27	6.59	6.85	8.38

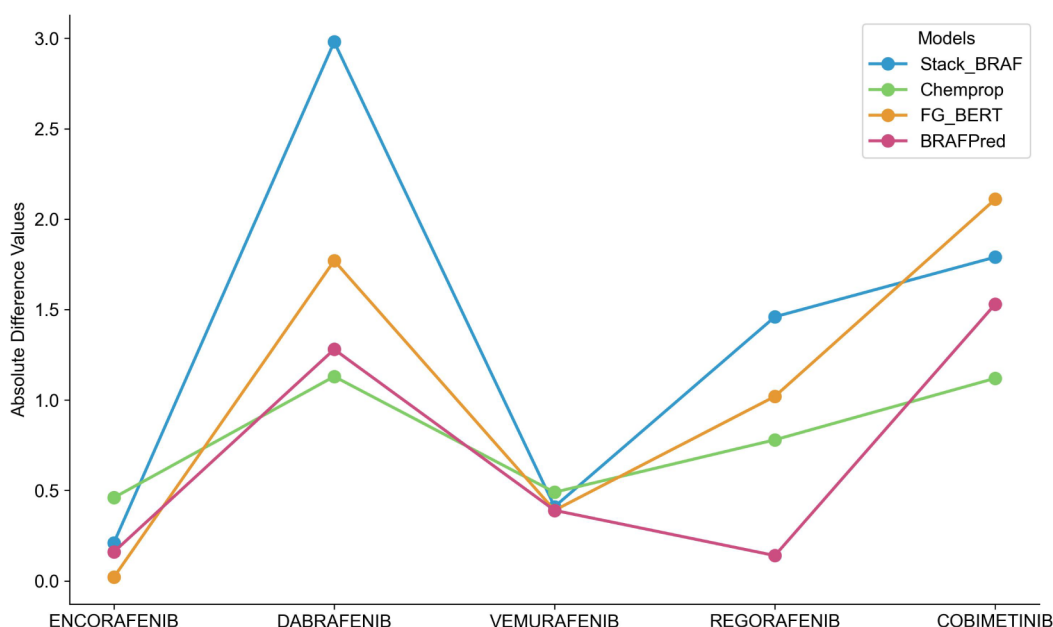


Figure 9. Absolute differences between predicted and true  $pIC_{50}$  values for key BRAF inhibitors.

Table 9. Top Five Predicted  $pIC_{50}$  Values for EGFR Inhibitors Identified by Different Models

NUM	SMILES	Chemprop	FG-BERT	BRAFPred	True $pIC_{50}$
1	<chem>CP(C)(C1=C2C(OCCO2)=CC=C1NC3=NC(NC4=CC(C)=C(N5CCC(N6CC(COC)C6)CC5)C=C4OC)=NC=C3Br)=O</chem>	9.79	9.59	9.94	10.05
2	<chem>F[C@@]([C@H](O)CC1(C)CN1C2=NC=CC(NC3=NC=C(C(N4C[C@H](CS(C)(=O)=O)[C@H]4C)=NC=C5C(C)C)C5=C3)=N2</chem>	9.90	9.82	9.88	9.70
3	<chem>CP(C)(C1=C2N=CC=NC2=CC=C1NC3=NC(NC4=CC(CC)=C(N5CCC(N6CCN(C)CCC6)CC5)C=C4OC)=NC=C3Br)=O</chem>	9.39	9.97	9.84	9.57
4	<chem>O[C@@H]1[C@@](F)(CC)CN(C2=NC=CC(NC3=CC4=C(C=N3)C(N5[C@H](C)[C@@H](CS(=O)(C)=O)C5)=CC=C4C(C)=N2)CC1</chem>	9.62	9.79	9.81	9.40
5	<chem>CP(C)(C1=C2C(OCCO2)=CC=C1NC3=NC(NC4=CC(C)=C(N5CCC(N6CCN(C)CCC6)CC5)C=C4OC)=NC=C3Br)=O</chem>	9.71	9.65	9.80	10.00

using random forest stacking. The results showed that BRAFPred achieved the highest  $R^2$ , outperforming the Baseline model (StackBRAF,  $R^2=0.889$ ) and other machine learning stacking models. Table 10 provides a detailed summary of

Table 10. Performance of Different Models on the EGFR Training Set

Model	$R^2$	MAE
RF_XGB	0.864	0.299
RF_SVR	0.877	0.333
RF_MLP	0.845	0.313
RF_DT	0.880	0.293
RF_KNN	0.881	0.322
StackBRAF	0.889	0.302
BRAFPred	0.895	0.289

different models' performance on the EGFR trainingset. Additionally, when we applied the Baseline model to the test set in this new application scenario, 13 out of the top 15 predicted high  $pIC_{50}$  compounds were confirmed as EGFR inhibitors, yielding a prediction accuracy of 87%.

These results highlight BRAFPred's ability to generalize beyond BRAF inhibitors and effectively identify highly potent inhibitors in early stage drug discovery. Unlike other conventional models, BRAFPred integrates a broader ensemble

learning framework that leverages diverse feature representations, enhancing its ability to generalize across novel chemical scaffolds. This capability is particularly crucial in virtual screening campaigns, where accurately prioritizing novel hit compounds significantly influences downstream experimental validation efforts.

Furthermore, BRAFPred's robust performance across different validation sets, including the newly added EGFR prediction experiment and the case study experiment, demonstrates its ability to mitigate experimental inconsistencies. By providing consistent and reliable activity ranking, the model aids in early hit identification, scaffold selection, and prioritization of promising compounds, ultimately supporting more informed decision-making in the field of drug discovery.

## 5. CONCLUSIONS

This study introduces a robust approach for designing BRAF inhibitors using a stacked ensemble model that combines traditional machine learning and deep learning techniques. SMILES structures are processed through 12 feature extraction networks, generating predictive features that are further refined using XGB and SVR. In parallel, deep learning models like Chemprop and the fine-tuned FG-BERT enhance the feature set, resulting in 26 predictive features. These are then fed into a Random Forest regressor, delivering highly accurate predictions with low error rates. BRAFPred's performance demon-



strates its potential as a powerful tool for evaluating and developing BRAF inhibitors. The integration of advanced modeling techniques underscores its value in targeted cancer therapy, offering a promising direction for future drug design.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The code and data set can be downloaded at <https://github.com/EvanZhang1216/BRAFPred>.

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.4c10367>.

Text S1: explanation of experimental results using different primary regressors on the training set; Text S2: explanation of experimental results using different secondary regressors on the blind test set; Text S3: further validation of model performance; Table S1: the Adaboost Model in first-level regressor performances on the training set; Table S2: the decision tree model in first-level regressor performances on the training set; Table S3: the extra trees model in first-level regressor performances on training set; Table S4: the KNN model in first-level regressor performances on the training set; Table S5: the PLS model in first-level regressor performances on the training set; Table S6: results of replacing different secondary regressors on the blind test set; Table S7: top ten predicted pIC<sub>50</sub> values for casestudy extend data set identified by different models (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Fang Ge** – State Key Laboratory of Flexible Electronics (LoFE) & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications, Nanjing 210023, China; Smart Health Big Data Analysis and Location Services Engineering Research Center of Jiangsu Province, Nanjing 210023, China; [orcid.org/0000-0001-5792-5379](https://orcid.org/0000-0001-5792-5379); Email: [gfang0616@njupt.edu.cn](mailto:gfang0616@njupt.edu.cn)

### Authors

**Ming Zhang** – School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China; [orcid.org/0009-0004-3197-3523](https://orcid.org/0009-0004-3197-3523)  
**Chaoming Zhang** – School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China  
**Keyu Liu** – School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China  
**Xibei Yang** – School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China  
**Xiaojuan Liu** – Department of Obstetrics and gynecology, The Affiliated People's Hospital, Jiangsu University, Zhenjiang 212103, China

Complete contact information is available at:  
<https://pubs.acs.org/doi/10.1021/acsomega.4c10367>

### Author Contributions

F.G. and M.Z. conceived and designed the study. C.Z. conducted the experiments. M.Z., F.G., and C.Z. performed the analyses and wrote the manuscript. K.L., X.L., and X.Y. revised the manuscript. M.Z. and F.G. are the cocorresponding authors.

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This study was funded by the National Science Foundation of China (no. 62076111), the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (grant no. NY223062), and the Natural Science Foundation of Nanjing University of Posts and Telecommunications (grant no. NY224158).

## ■ ABBREVIATIONS

SMILES, simplified molecular input line entry system; SVR, support vector regression; XGB, eXtreme Gradient Boosting; CV, cross-validation; FDA, Food and Drug Administration; IC<sub>50</sub>, inhibitor concentration at 50%; MAE, mean absolute error; R<sup>2</sup>, coefficient of determination; QSAR, quantitative structure–activity relationship; RF, random forest; SHAP, SHapley Additive exPlanations

## ■ REFERENCES

- (1) Catalanotti, F.; Reyes, G.; Jesenberger, V.; Galabova-Kovacs, G.; de Matos Simoes, R.; Carugo, O.; Baccarini, M. A Mek1-Mek2 heterodimer determines the strength and duration of the Erk signal. *Nat. Struct. Mol. Biol.* **2009**, *16* (3), 294–303.
- (2) Zhang, M.; Xu, Y.; Li, L.; Liu, Z.; Yang, X.; Yu, D. J. Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. *Anal. Biochem.* **2018**, *550*, 41–48.
- (3) Safaei Ardekani, G.; Jafarnejad, S. M.; Tan, L.; Saeedi, A.; Li, G. The prognostic value of BRAF mutation in colorectal cancer and melanoma: A systematic review and meta-analysis. *PLoS One* **2012**, *7* (10), No. e47054.
- (4) Lavoie, H.; Therrien, M. Regulation of RAF protein kinases in ERK signalling. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (5), 281–298.
- (5) Rizos, H.; Menzies, A. M.; Pupo, G. M.; Carlino, M. S.; Fung, C.; Hyman, J.; Haydu, L. E.; Mijatov, B.; Becker, T. M.; Boyd, S. C. BRAF inhibitor resistance mechanisms in metastatic melanoma: Spectrum and clinical impact. *Clin. Cancer Res.* **2014**, *20* (7), 1965–1977.
- (6) Agianian, B.; Gavathiotis, E. Current Insights of BRAF Inhibitors in Cancer. *J. Med. Chem.* **2018**, *61* (14), 5775–5793.
- (7) Ai, Y.; Wang, S.-T.; Tang, C.; Sun, P.-H.; Song, F.-J. 3D-QSAR and docking studies on pyridopyrazinones as BRAF inhibitors. *Med. Chem. Res.* **2011**, *20* (8), 1298–1317.
- (8) Syahid, N. F.; Weerapreeyakul, N.; Srisongkram, T. StackBRAF: A Large-Scale Stacking Ensemble Learning for BRAF Affinity Prediction. *ACS Omega* **2023**, *8* (23), 20881–20891.
- (9) Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26* (1), 80–93.
- (10) Chen, X.; Xie, W.; Yang, Y.; Hua, Y.; Xing, G.; Liang, L.; Deng, C.; Wang, Y.; Fan, Y.; Liu, H. Discovery of Dual FGFR4 and EGFR Inhibitors by Machine Learning and Biological Evaluation. *J. Chem. Inf. Model.* **2020**, *60* (10), 4640–4652.
- (11) Hu, J.; Li, Y.; Zhang, M.; Yang, X.; Shen, H. B.; Yu, D. J. Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2017**, *14* (6), 1389–1398.
- (12) Cherkassky, V. The nature of statistical learning theory. *IEEE Trans. Neural Networks* **1997**, *8* (6), 1564.
- (13) Chen, T.; Guestrin, C. 2016. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 785–794; Association for Computing Machinery.
- (14) Suvannang, N.; Preeyanon, L.; Malik, A. A.; Schaduengrat, N.; Shoombuatong, W.; Worachartcheewan, A.; Tantimongkolwat, T.; Nantasenamat, C. Probing the origin of estrogen receptor alpha

- inhibition via large-scale QSAR study. *RSC Adv.* **2018**, *8* (21), 11344–11356.
- (15) Wang, S.; Di, J.; Wang, D.; Dai, X.; Hua, Y.; Gao, X.; Zheng, A.; Gao, J. State-of-the-Art Review of Artificial Neural Networks to Predict, Characterize and Optimize Pharmaceutical Formulation. *Pharmaceutics* **2022**, *14* (1), 183.
- (16) Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.* **2019**, *6* (1), 11.
- (17) Ben, J.; Sun, Q.; Liu, K.; Yang, X.; Zhang, F. Multi-head multi-order graph attention networks. *Appl. Intell.* **2024**, *54* (17–18), 8092–8107.
- (18) Zhang, J.; Liu, K.; Yang, X.; Ju, H.; Xu, S. Multi-label learning with Relief-based label-specific feature selection. *Appl. Intell.* **2023**, *53* (15), 18517–18530.
- (19) Guo, Q.; Yang, X.; Li, M.; Qian, Y. Collaborative graph neural networks for augmented graphs: A local-to-global perspective. *Pattern Recognit.* **2025**, *158*, 111020.
- (20) Zhang, M.; Gong, C.; Ge, F.; Yu, D. J. FCMSTrans: Accurate Prediction of Disease-Associated nsSNPs by Utilizing Multiscale Convolution and Deep Feature Combination within a Transformer Framework. *J. Chem. Inf. Model.* **2024**, *64* (4), 1394–1406.
- (21) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30* (8), 595–608.
- (22) Zhang, M.; Wang, X.; Xu, S.; Ge, F.; Paixao, I. C.; Song, J.; Yu, D. J. MetalTrans: A Biological Language Model-Based Approach for Predicting Disease-Associated Mutations in Protein Metal-Binding Sites. *J. Chem. Inf. Model.* **2024**, *64* (15), 6216–6229.
- (23) Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.
- (24) Wang, Y.; Yang, X.; Sun, Q.; Qian, Y.; Guo, Q. Purity Skeleton Dynamic Hypergraph Neural Network. *Neurocomputing* **2024**, *610*, 128539.
- (25) Guo, Q.; Yang, X.; Zhang, F.; Xu, T. Perturbation-augmented Graph Convolutional Networks: A Graph Contrastive Learning architecture for effective node classification tasks. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107616.
- (26) Qian, D.; Liu, K.; Zhang, S.; Yang, X. Semi-supervised feature selection by minimum neighborhood redundancy and maximum neighborhood relevancy. *Appl. Intell.* **2024**, *54* (17–18), 7750–7764.
- (27) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchol, M. Mol-CycleGAN: A generative model for molecular optimization. *J. Cheminf.* **2020**, *12*, 2.
- (28) Wu, Z.; Jiang, D.; Wang, J.; Zhang, X.; Du, H.; Pan, L.; Hsieh, C. Y.; Cao, D.; Hou, T. Knowledge-based BERT: A method to extract molecular features like computational chemists. *Brief. Bioinform.* **2022**, *23* (3), bbac131.
- (29) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474.
- (30) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100–D1107.
- (31) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
- (32) Matsson, P.; Kihlberg, J. How Big Is Too Big for Cell Permeability? *J. Med. Chem.* **2017**, *60* (5), 1662–1664.
- (33) Awale, M.; Reymond, J. L. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *J. Chem. Inf. Model.* **2014**, *54* (7), 1892–1907.
- (34) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliakova, N.; Kuhn, S.; Pluskal, T.; Rojas-Cherto, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **2017**, *9* (1), 33.
- (35) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039–1045.
- (36) Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24* (21), 2518–2525.
- (37) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.
- (38) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–1213.
- (39) Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. **2018**.
- (40) Li, B.; Lin, M.; Chen, T.; Wang, L. FG-BERT: A generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Brief. Bioinform.* **2023**, *24* (6), bbad398.
- (41) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388.
- (42) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv*. **2017**.
- (43) Swanson, K.; Liu, G.; Catacutan, D. B.; Arnold, A.; Zou, J.; Stokes, J. M. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* **2024**, *6* (3), 338–353.
- (44) Kim, G.; McKee, A. E.; Ning, Y. M.; Hazarika, M.; Theoret, M.; Johnson, J. R.; Xu, Q. C.; Tang, S.; Sridhara, R.; Jiang, X. FDA approval summary: Vemurafenib for treatment of unresectable or metastatic melanoma with the BRAFV600E mutation. *Clin. Cancer Res.* **2014**, *20* (19), 4994–5000.
- (45) Hao, C.; Li, X.; Wang, Z.; Liu, L.; He, F.; Pan, Z. Optically activated MEK1/2 inhibitors (Opti-MEKi) as potential antimelanoma agents. *Eur. J. Med. Chem.* **2023**, *251*, 115236.
- (46) Ammar, U. M.; Abdel-Maksoud, M. S.; Mersal, K. I.; Ali, E. M. H.; Yoo, K. H.; Choi, H. S.; Lee, J. K.; Cha, S. Y.; Oh, C. H. Modification of imidazothiazole derivatives gives promising activity in B-Raf kinase enzyme inhibition; synthesis, in vitro studies and molecular docking. *Bioorg. Med. Chem. Lett.* **2020**, *30* (20), 127478.
- (47) Abdel-Maksoud, M. S.; Ali, E. M. H.; Ammar, U. M.; Mersal, K. I.; Yoo, K. H.; Oh, C. H. Design and synthesis of novel pyrrolo[2,3-b]pyridine derivatives targeting (V600E)BRAF. *Bioorg. Med. Chem.* **2020**, *28* (11), 115493.
- (48) Kopetz, S.; Grothey, A.; Yaeger, R.; Van Cutsem, E.; Desai, J.; Yoshino, T.; Wasan, H.; Ciardiello, F.; Loupakis, F.; Hong, Y. S. Encorafenib, Binimetinib, and Cetuximab in BRAF V600E-Mutated Colorectal Cancer. *N. Engl. J. Med.* **2019**, *381* (17), 1632–1643.
- (49) Strumberg, D.; Schultheis, B. Regorafenib for cancer. *Expert Opin. Invest. Drugs* **2012**, *21* (6), 879–889.
- (50) Cheng, H.; Chang, Y.; Zhang, L.; Luo, J.; Tu, Z.; Lu, X.; Zhang, Q.; Lu, J.; Ren, X.; Ding, K. Identification and optimization of new dual inhibitors of B-Raf and epidermal growth factor receptor kinases for overcoming resistance against vemurafenib. *J. Med. Chem.* **2014**, *57* (6), 2692–2703.
- (51) Davis, J.; Wayman, M. Encorafenib and Binimetinib Combination Therapy in Metastatic Melanoma. *J. Adv. Pract. Oncol.* **2022**, *13* (4), 450–455.
- (52) Bouffet, E.; Hansford, J. R.; Garre, M. L.; Hara, J.; Plant-Fox, A.; Aerts, I.; Locatelli, F.; van der Lugt, J.; Papusha, L.; Sahm, F. Dabrafenib plus Trametinib in Pediatric Glioma with BRAF V600 Mutations. *N. Engl. J. Med.* **2023**, *389* (12), 1108–1120.
- (53) Ascierto, P. A.; Dreno, B.; Larkin, J.; Ribas, A.; Liszkay, G.; Maio, M.; Mandal, M.; Demidov, L.; Stroyakovskiy, D.; Thomas, L.

5-Year Outcomes with Cobimetinib plus Vemurafenib in BRAFV600 Mutation-Positive Advanced Melanoma: Extended Follow-up of the coBRIM Study. *Clin. Cancer Res.* **2021**, 27 (19), 5225–5235.

(54) Chang, H.; Zhang, Z.; Tian, J.; Bai, T.; Xiao, Z.; Wang, D.; Qiao, R.; Li, C. Machine Learning-Based Virtual Screening and Identification of the Fourth-Generation EGFR Inhibitors. *ACS Omega* **2024**, 9 (2), 2314–2324.