

Modeling and Interpretability Study of the Structure–Activity Relationship for Multigeneration EGFR Inhibitors

Zhiqi Sun, Donghui Huo, Jiangyu Guo, and Aixia Yan*



Cite This: *ACS Omega* 2025, 10, 11176–11187



Read Online

ACCESS |



Metrics & More

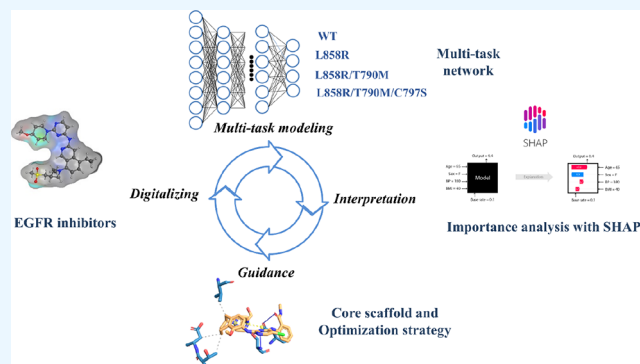


Article Recommendations



Supporting Information

ABSTRACT: The fourth-generation EGFR inhibitors targeting L858R/T790M/C797S mutations are in clinical trials mostly, and it is necessary to develop new inhibitors. In this study, an internal data set containing 2302 multitarget EGFR inhibitors targeting the wild type (83%) and the L858R (92%), L858R/T790M (96%), and L858R/T790M/C797S (60%) mutations was collected. We established a structure–activity relationship model for predicting the bioactivities of multigeneration EGFR inhibitors by a multitask deep neural network (MT-DNN). We also constructed four single-task models on 1384 L858R/T790M/C797S (60%) mutation inhibitors by support vector machine (SVM), random forest (RF), XGBoost (XGB), and single-target neural network (ST-DNN), respectively. The MT-DNN model significantly outperformed single-task models on the external data set of 304 fourth-generation EGFR inhibitors. Furthermore, the combined application of MT-DNN and SHAP/delta-SHAP value interpretability analysis offers rigorous structural information from a global perspective. With SHAP/delta-SHAP methods, the MT-DNN model can mine the core scaffold and important fragments of multigeneration EGFR inhibitors and provide valuable information from a structure–activity relationship perspective to address the resistant mutation problem.



INTRODUCTION

The epidermal growth factor receptor (EGFR) is a member of the tyrosine kinase receptor family and is one of the key targets for treatment of various diseases. It can maintain the growth and dynamic equilibrium of epithelial tissues.¹ The EGFR plays an important role in the development of many cancers, such as lung cancer,² breast cancer,³ and head and neck cancer.⁴ The EGFR is also the first cancer target discovered in nonsmall cell lung cancer overexpressing in more than 60% of patients.⁵ EGFR inhibitors have had great success in clinical cancer medical treatment, which are able to effectively improve the median survival of cancer patients and are widely used in both first-line and second-line treatments.^{6,7}

The first-generation EGFR inhibitors gefitinib and erlotinib were approved in 2003 and 2004, respectively, for the treatment of cancers carrying an abnormal EGFR.^{8,9} These small molecules inhibit both the wild-type (WT) EGFR and single mutations in the EGFR, such as exon 19 deletions or EGFR activating mutation L858R.¹⁰ However, their use in treating cancer patients is limited due to acquired resistance caused by T790M mutation.^{11,12} To overcome this acquired resistance, the second-generation EGFR inhibitor afatinib was developed in 2013;¹³ neratinib¹⁴ and dacomitinib¹⁵ were developed in 2017 and 2018 afterward, respectively. However, they were disrupted due to dose toxicity caused by their irreversible targeting of wild-type EGFR.¹⁶ Subsequently,

emergence of the third-generation inhibitors osimertinib¹⁷ and mobocertinib¹⁸ addressed the toxicity issue and were approved for patients with T790M mutation cancer in 2015 and 2021, respectively. Along with L858R and T790M mutations, the third point mutation C797S prevents the formation of a covalent bond between cysteine and osimertinib, leading to resistance.¹⁹ Currently, there are no available treatments to block disease progress when C797S appears in cis with L858R and T790M²⁰ (Table 1). With clinical trials of the fourth-generation inhibitor BLU-945¹⁹ terminated, it is urgent to develop the fourth-generation EGFR inhibitors.

With the increasing use of machine learning in drug discovery projects, AI-based computational methods have become invaluable tools. These methods can analyze vast amounts of data to identify potential drug candidates more efficiently and accurately than traditional methods. Machine learning algorithms such as deep learning, support vector

Received: November 18, 2024

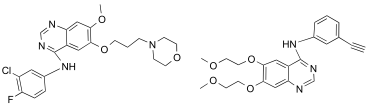
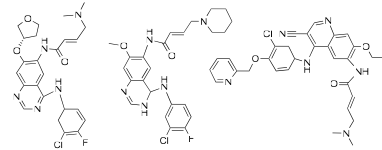
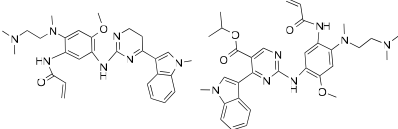
Revised: February 19, 2025

Accepted: February 25, 2025

Published: March 14, 2025



Table 1. Representative Drugs and Inhibitory Targets of First-, Second-, and Third-Generation EGFR Inhibitors^a

Generation of EGFR inhibitors	FDA-approved drugs	time	Targets of WT and resistant mutations				
			WT	L858R	C797S	L858R/T790M	L858R/T790M/C797S
First Generation	 Gefitinib Erlotinib	2003,	✓	✓	✓	×	Trans:✓
		2004					Cis:×
Second Generation	 Afatinib Dacomitinib Neratinib	2013,	Dose toxicity				Trans:✓
		2018,		✓	✓	✓	Cis:×
		2017					
Third Generation	 Osimertinib Mobocertinib	2015,	✓	✓	✓	✓	Trans:✓
		2021					Cis:×

^aTrans: trans mutations on different alleles; Cis: homeopathic mutations on the same allele.

machines, and random forests have been applied to predict drug–target interactions, optimize molecular properties, and design new drug candidates.^{21,22} By leveraging the power of machine learning, researchers can uncover patterns and relationships within data that may not be immediately apparent through conventional analysis techniques. This leads to the development of more effective and targeted therapies. AI-based computational methods can leverage drug design experiences in successful cases to improve the efficiency of developing new EGFR inhibitors.^{23,24} In previous studies, we obtained several highly active inhibitors against the WT EGFR in ligand-based and receptor-based hierarchical virtual screening^{25,26} and used a variety of machine learning methods to explore the 2D and 3D structure–activity relationship of WT and L858R/T790M mutant inhibitors.²⁷ Chang et al. developed a support vector machine (SVM) classification model based on 221 EGFR inhibitors with L858R/T790M/C797S mutations, integrating both structural and nonstructural descriptors. Their SVM classification model determined four potential hits through virtual screening.²⁸ Zhang and Li, using a data set of 539 fourth-generation EGFR inhibitors, established a series of classification/regression models and found that the best SVM and random forest models exhibited optimal predictive capabilities.²⁹ According to the reported EGFR machine learning studies, although there are many studies on single-mutation (L858R) and double-mutation (L858R/T790M) types of EGFR inhibitors (the first-, second-, and third-generations inhibitors),^{30–32} there are still few studies on unsolved triple-mutation (L858R/T790M/C797S) EGFR inhibitors (the fourth-generation inhibitors). The limited application of machine learning in fourth-generation inhibitor development primarily stems from lack of data.

Overcoming EGFR drug resistance typically involves two strategies. One approach is to design multitarget inhibitors against multiple cancer signaling pathways to achieve cooperative inhibition and reduce resistance occurrence.^{33–36}

The other common strategy is to design new inhibitors based on existing EGFR crystal structures, involving structure-based functional fragment modifications and scaffold hopping on existing first-, second-, and third-generation EGFR inhibitors.^{37–39} In addition to the known fourth-generation inhibitors, it is helpful to use the information on the first-, second-, and third-generation EGFR inhibitors to discover the possible fourth-generation EGFR inhibitors. Two most common resistance mutations, T790M and C797S, represent only one amino acid changing in the active binding site of the EGFR, indicating that there are most of the fragments in the existing first-, second-, and third-generation EGFR inhibitors that also work in the fourth-generation inhibitors. From the perspective of protein homology, WT and L858R, L858R/T790M, and L858R/T790M/C797S mutants differ by only few amino acids, which means their similarity is close to 100%. Therefore, this study proposes an approach to the EGFR drug design strategy different from the two strategies above: integrating inhibitors of WT, L858R, L858R/T790M, and L858R/T790M/C797S mutants into a multitask neural network to construct a structure–activity relationship model.

The multitask deep neural network (MT-DNN) refers to a deep learning model trained together on multiple shared tasks.^{40,41} The MT-DNN has been hugely successful in characterizing the selectivity of kinase inhibitors.^{42–44} MT-DNN models outperform single-task models for predicting the bioactivity of small-molecule inhibitors of protein kinases,^{45,46} and dense and highly related kinase prediction tasks bring more positive transfer.^{46–49} Similar to the advantage of perturbation-theory machine learning in multitarget drug discovery,^{50–54} the MT-DNN can simultaneously predict multiple activities against different targets and offer a direct physicochemical and structural interpretations. Compared with applying the MT-DNN to a large kinome, our model is constructed based on inhibitors against different mutants of the EGFR, achieving at least three objectives: (1) focusing on

the EGFR and its three mutants to reduce computational resource consumption; (2) investigating if the performance of the MT-DNN model is better than that of the single-task model on the same task of EGFR inhibitors; and (3) interpreting the structure–activity relationship of fourth-generation EGFR inhibitors based on the MT-DNN model.

MATERIALS AND METHODS

Data Set Preparation. *Internal Data Set:* 2302 EGFR Inhibitors Targeting the WT and the L858R, L858R/T790M, and L858R/T790M/C797S Mutants. We have collected 2302 inhibitors from the SciFinder,⁵⁵ ChEMBL,⁵⁶ and BindingDB⁵⁷ databases as the internal data set (Table 2 and Table S1).

Table 2. Internal, External, and Decoy Data Sets of EGFR Inhibitors in This Study

data set	type	label 0 ^a	label 1 ^b	sum	percentage ^c
internal set	WT	839	1066	1905	83%
	L858R	592	1528	2120	92%
	L858R/T790M	754	1458	2212	96%
	L858R/T790M/C797S	894	490	1384	60%
external set	L858R/T790M/C797S	191	113	304	
decoy set				1123	

^aInhibitors with $IC_{50} \geq 1 \mu M$ on different EGFR types are regarded as label 0. ^bInhibitors with $IC_{50} < 1 \mu M$ on different EGFR types are regarded as label 1. ^cThe percent of sum of the inhibitors to all the 2302 inhibitors.

There were 1905 (83%), 2120 (92%), 2212 (96%), and 1384 (60%) of 2302 EGFR inhibitors against the WT and the L858R, L858R/T790M, and L858R/T790M/C797S mutants, respectively; the data distribution is shown in Table 2.

External Data Set: 304 EGFR Inhibitors Targeting L858R/T790M/C797S Mutation. We also used SciFinder and Reaxys⁵⁸ to search for the latest journals and patents on EGFR L858R/T790M/C797S mutant inhibitors, collecting a total of 304 inhibitors as the external data set (Table 2 and Table S2).

Decoy Set: 1123 Molecules from the ZINC12 All-Purchasable Subset. To test the performance of our model on the decoy set, we used “MUBD-DecoyMaker 2.0”⁵⁹ to generate a decoy set consisting of 1123 data points (Table 2).

Data Preprocessing. The inhibitors are all small-molecule compounds. Each inhibitor has an inhibitory activity IC_{50} in vitro for at least one of the three EGFR mutants: L858R, L858R/T790M, or L858R/T790M/C797S. To ensure similarity in binding modes for relevant learning, we exclude inhibitors targeting only the allosteric site. We classify inhibitors into highly and lowly active according to an IC_{50} threshold of $1 \mu M$, whose activity values below $1 \mu M$ are considered as highly active inhibitors and those equal to or above $1 \mu M$ are considered as lowly active inhibitors.

We have set a criterion for data wash in order to improve the data quality. If an inhibitor has conflicting activity labels from different sources, we follow the majority rule and remove conflicting data. Furthermore, if the experimental bioassay results from a journal or patent are all highly or lowly active, then we require a clear control group experiment. Otherwise, we do not collect the data.

MT-DNN Model. Compared with single-task neural networks, many studies have found that this shared mode

can significantly improve the model’s predictive performance, especially for single tasks with insufficient data.^{45,46} Here, the MT-DNN model treats L858R, L858R/T790M, and L858R/T790M/C797S EGFR mutants and EGFR WT inhibitor data as four related tasks.

The architecture of the MT-DNN model is illustrated in Figure 1. Its architecture and training process are assisted by

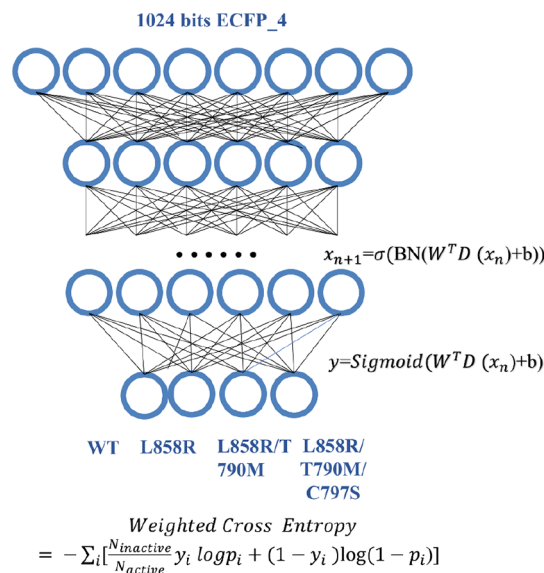


Figure 1. Architecture of the MT-DNN model. The model input is the 1024-bit ECFP₄ fingerprint. The input of the $n+1$ th layer is x_{n+1} , and the last layer consists of four y outputs. The loss function is a balanced cross-entropy with weights.

Python’s PyTorch package (Version 2.0.1).⁶⁰ The input of the model only considers a two-dimensional structure of the molecule, represented by a 1024-bit ECFP₄ fingerprint.⁶¹

Architecture of the MT-DNN Model. The hidden layer in the MT-DNN model is composed of a single modular concatenation. This module sequentially connects a dropout layer, a fully connected layer, an activation function layer, and a normalization layer. The layer is updated using eq 1:

$$X_{i+1} = \sigma(\text{BN}(W_i^T(\text{DO}(X_i)) + b_i)) \quad (1)$$

where X_{i+1} and X_i represent the inputs of the $i+1$ th and i th layers, respectively. W_i and b_i represent the weight matrix and bias of the i th layer, respectively. BN is the normalization layer between the fully connected layer and the activation function. DO is the dropout layer. The activation function layer σ uses the ReLU function. The model’s output layer contains four prediction tasks corresponding to four different types of EGFR inhibitors. They all use the Sigmoid function as activation function. The label prediction probability for all tasks is between 0 and 1. The predicted label probability y_l for the l th task model is computed using eq 1:

$$y_l = \text{sigmoid}((W_{l-1}^T(\text{DO}(X_{l-1})) + b_{l-1})) \quad (2)$$

Due to the imbalance of positive and negative samples for EGFR WT and mutant inhibitors, our model inevitably tends to predict learning for the more numerous samples during training. In the internal data set, the percentages of positive samples for the EGFR WT and the L858R, L858R/T790M, and L858R/T790M/C797S mutants are 56, 72, 66, and 36%, respectively. Given the imbalance of positive and negative

samples, we used weighted binary cross-entropy as a loss function for deep learning. Samples from the less numerous samples are given a higher weight in model learning. The loss function for each inhibitor is calculated using eq 3:

$$\text{weighted cross - entropy} = - \sum_{i=1}^4 \left[\frac{N_{\text{inactive},i}}{N_{\text{active},i}} y_i \log(p_i) + \left(1 - y_i\right) \log(1 - p_i) \right] \quad (3)$$

As mentioned earlier, there are four tasks corresponding to prediction of four types of EGFR inhibitors. $N_{\text{inactive},i}$ and $N_{\text{active},i}$ represent the number of negative and positive samples for the i th task, respectively. y_i represents the true label for the i th task. p_i represents the predicted probability for the i th task. The training is terminated early if the Matthews correlation coefficient (MCC) value does not improve after 40 learning iterations.

Parameter Optimization. We used Bayesian optimization for the Python (pyGPGO) package⁶² to optimize hyperparameters, which include dropout rate, batch size, number of hidden layers, and size of hidden layers. The optimization process is conducted 20 times (Table S3). Each optimization process undergoes 5-fold cluster cross-validation. The Butina clustering algorithm based on Tanimoto similarity⁶³ is used for data division. Two molecules with a Tanimoto similarity of greater than 0.8 are clustered together. The data are roughly divided into five equal parts by the clustering algorithm. Each part runs as a test set five times, and the average of five results is taken as the final performance of the model under that hyperparameter condition. Since similar molecules do not appear in both the training set and the test set, the model cannot simply make predictions based on the similarity between molecules. Only by mining deep structural information can the model improve its final performance. The 5-fold cluster cross-validation groups similar molecules together, reducing the possibility of taking shortcuts of the model. During parameter optimization, if the MCC does not improve after five learning iterations, the training process is terminated early.

Model Evaluation. We used grid search algorithms in Python (SKlearn) to perform hyperparameter optimization for SVMs, random forests, and XGBoost. BA (eq 4), sensitivity (eq 5), specificity (eq 6), MCC (eq 7), and AUC (eq 8) were used to evaluate the model.

$$\text{BA} = \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) / 2 \quad (4)$$

$$\text{SE} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{SP} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (7)$$

$$\text{AUC} = \text{area under the curve of ROC} \quad (8)$$

TP and TN are the true positive rate and true negative rate, respectively. In the MT-DNN model, the TP and TN of

samples are the sum of TP and TN of all task samples, respectively.

SHAP Method. SHAP (SHapley Additive exPlanations) is a machine learning interpretability method based on game theory. SHAP provides a Python package to analyze the contribution value of each feature for each input sample in machine learning.⁶⁴

SHAP Significance Analysis. It is worth exploring whether the addition of three related tasks from ST-DNN to MT-DNN has caused a significant change in the decision-making process for the fourth-generation EGFR inhibitors. We conduct SHAP interpretability analysis on MT-DNN and ST-DNN models with identical training hyperparameters, respectively. The only difference between the two models is the number of tasks. The SHAP value for each fingerprint bit represents its importance in the model. After calculating the SHAP importance for 1024 fingerprint bits for two models, we offer a significance analysis on 1024 SHAP distributions. Figure S2 provides a SHAP significance analysis from two perspectives: the significance of individual fingerprint bit SHAP differences and the trend relationship between fingerprint SHAP differences and fingerprint variance. As shown in Figure S2, there are as many as 931 fingerprint bits with p values less than 0.05 and only 93 fingerprint bits with p values greater than 0.05. This indicates that at a significance level of 0.05, the importance of most fingerprint bits is significantly different between MT-DNN and ST-DNN models (Figure S3). The Spearman correlation coefficient between the variance of fingerprint bits and p values is -0.75 , suggesting that the larger amount of information (larger variance) a fingerprint bit stores, the greater its importance's differences (smaller p value) tend to be. Therefore, the differences between MT-DNN and ST-DNN are mainly affected by those important fingerprint bits.

RESULTS AND DISCUSSION

The Data Distribution of Known EGFR Inhibitors. The numbers of highly and lowly active inhibitors of four different types of EGFR (WT, L858R, L858R/T790M, and L858R/T790M/C797S) and their correlations are shown in Figure S1. One can see that due to the limited amount of data in the public library, there is less data on the fourth-generation EGFR inhibitors targeting L858R/T790M/C797S. We also conducted a correlation analysis between the outputs of each pair of tasks. Apart from WT inhibitors and L858R/T790M/C797S inhibitors, the correlation coefficients between other pairwise type inhibitors' prediction tasks are greater than 0.4 (Figure S1 and Table S4). This indicates that our data has high relevance, making it suitable for establishing a multitask neural network model.

Exploring the Internal Stability of the MT-DNN Model. To investigate the stability of the MT-DNN model, we employ a clustering-based 5-fold cross-validation strategy. Figure 2 demonstrates the 3D T-SNE visualization results for the 5-fold clustering internal data set and external data set. The visualization of EGFR inhibitor data shows that similar molecules were gathered into small but homogeneous groups and divided into one cluster. One can see that the external data set is in the general range of the internal data set but is not very similar to the internal data set, making it a challenging task to predict. Therefore, external tasks are more challenging to predict than internal cross-validation tasks.

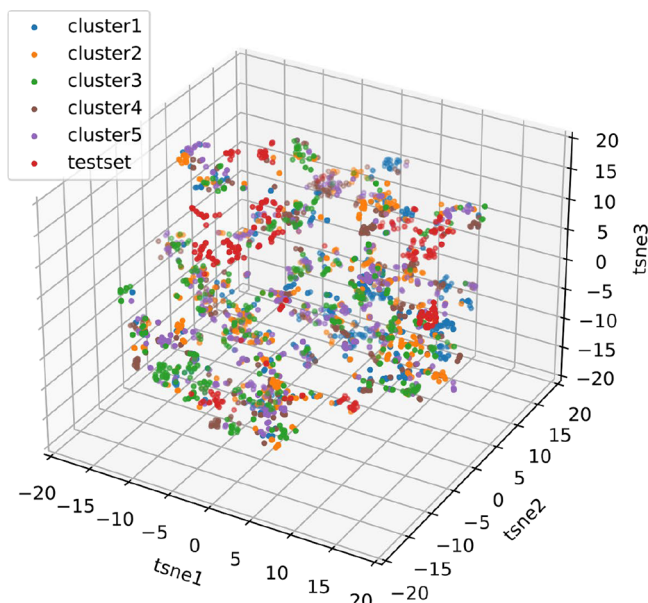


Figure 2. 3D T-SNE visualization results for the 5-fold clustering internal data set and the external data set (test set). The internal data set is split into five clusters (clusters 1–5). Similar molecules with a Tanimoto coefficient less than 0.2 are clustered together, and then all compounds in the internal data set were randomly divided into approximately five equal parts (with similar amounts of compounds). One can see that the distribution of data in five equal parts is similar, but the molecular similarity in different parts is low, and the data in the external data set is within the range of the internal data set.

Table 3 shows the performances of the MT-DNN model based on the 5-fold clustering cross-validation and the mean

Table 3. Mean Results of the MT-DNN Model by 5-Fold Clustering Cross-Validation^a

evaluation	BA	SE	SP	MCC	AUC
Model1	0.87	0.81	0.94	0.76	0.94
Model2	0.84	0.86	0.82	0.66	0.90
Model3	0.76	0.69	0.84	0.53	0.81
Model4	0.80	0.68	0.91	0.62	0.86
Model5	0.85	0.86	0.84	0.68	0.91
mean	0.82	0.78	0.87	0.65	0.88

^aBA: balanced accuracy; MCC: Matthews correlation coefficient; SE: sensitivity; SP: specificity; AUC: area under the ROC curve.

results. Surprisingly, despite using the traditional ECFP₄ fingerprint as input, the average AUC from the 5-fold cross-validation still reaches 0.88. One can see that the MT-DNN model indeed exhibits high internal stability. Additionally, although the collected EGFR data faces class imbalance problems, the average sensitivity and specificity of the 5-fold cross-validation models are quite similar and both exceed 0.75. This can be attributed to the use of weighted balanced cross-entropy as a loss function.

Applicability Domain. Figure 3 illustrates the applicability domain, where black dots represent the internal data set and red dots represent the external data set. By applying principal component analysis (PCA) to ECFP₄ fingerprints, high-dimensional data are reduced to a two-dimensional plane, allowing us to visualize the data distribution intuitively. The maximum PCA radius for the training data is 5.1, indicating the maximum extent of the training data in the PCA space.

Additionally, Figure 3B shows the radius distance from the center of the training data (the internal data set) to the test data (the external data set). It is evident that all test data points have distances less than 5.1, indicating that the test data are entirely within the range of the training data. Further analysis of the distribution of the test data points reveals that most of them are concentrated between 0.5 and 2.0. This indicates a good overlap in the feature space between the training and test data, thereby validating the model's generalization ability on the test set. This visualization helps in understanding the model's performance under different data distributions, evaluating the model's generalization capability, and identifying potential extrapolation risks.

Comparison of the MT-DNN Model with a Single-Task Model. Based on an external data set, we conduct a rigorous comparative evaluation of the MT-DNN with the other four machine learning algorithms (the SVM, RF, XGB and ST-DNN) in predicting structure–activity relationships. Figure 4 shows the balanced accuracy (BA), MCC, and AUC values of the MT-DNN model and four single-task models (SVM, RF, XGB, and ST-DNN models) for the prediction of EGFR L858R/T790M/C797S mutant activity. From Figure 4, one can see that the results of the MT-DNN model are better than those of the other four models. The MCC value for the MT-DNN model is at least 0.2 higher than those of all other single-task models. Since the training data (L858R/T790M/C797S inhibitors) for all the models are conducted on the same data set exactly, such results sufficiently demonstrate that adding multiple related EGFR mutations tasks can greatly enhance model's generalization performance. Even with only four related tasks, the performance of the MT-DNN model is still significantly better than other models.

ROC Curve. Figure 5 displays four distinct ROC curves, showcasing the performance of various models and data sets. Plots A and B represent ROC curves based on leave-one-out cross-validation (LOO-CV) using different similarity-based approaches for the decoy set. Specifically, the “similarity-in-properties-based” and MACCS “similarity-in-structure-based” methods are used. The ROC AUC values close to 0.5 for these plots indicate the unbiased nature of the decoy set. Conversely, the bottom plots illustrate the ROC curves for the multitask deep neural network (MT-DNN) models. Plot C evaluates the models solely on the external test set, while plot D assesses the models on both the external test set and the decoy set. The ROC curves with AUC values approaching 1 in these plots highlight the high generalization capability of the MT-DNN models, demonstrating their effectiveness and robustness across both data sets. By comparing plots C and D, one can see that the performance of the model is significantly improved after adding the decoy set. This reflects the prediction difficulty of the selected external test set. The analysis of activity cliffs (Figure S4) also validates this point, as the data in the test set contains multiple challenging activity cliffs to predict.^{65,66}

Interpretability Analysis of Existing Complexes. To preliminarily explore MT-DNN's interpretability, SHAP interpretability analysis of the fourth-generation inhibitor brigatinib⁶⁷ is shown in Figure 6. Figure 6A presents the results sorted by mean arithmetic SHAP values for each bit of the molecular fingerprint. The top four fingerprints, ECFP193, ECFP602, ECFP757, and ECFP441, which have the most significant positive impact on the model, are circled in red in Figure 6B. They capture the core scaffold structures of pyrimidinamine, phosphorus, phenylphosphine, and piper-

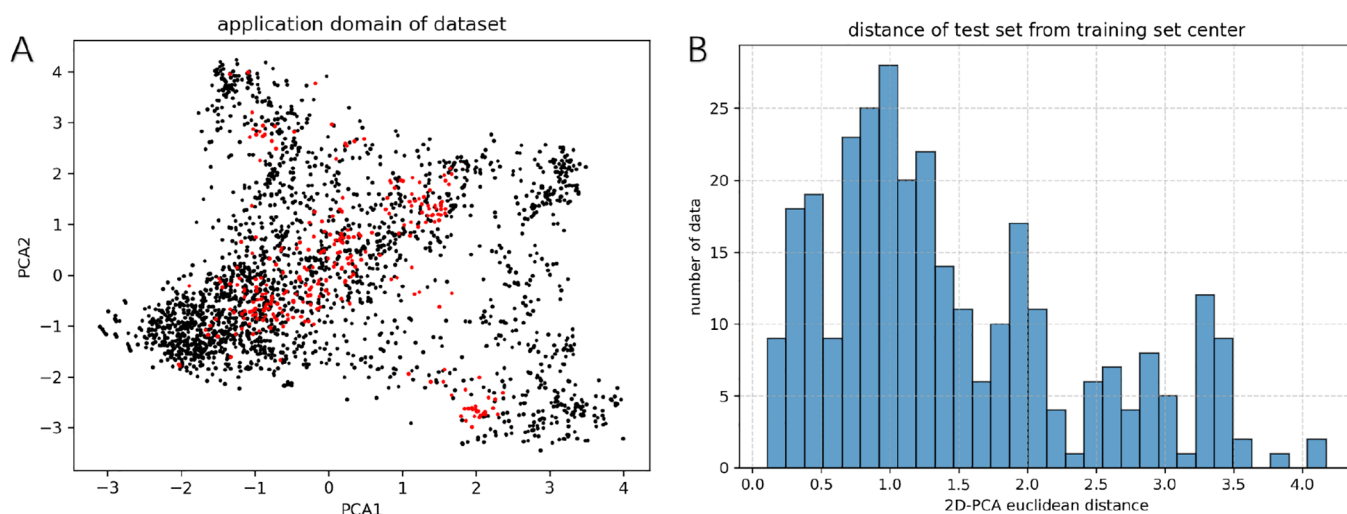


Figure 3. Applicability domain analysis. (A) 2D-PCA plot (black: internal data set; red: external data set). (B) Distance from compounds in the external data set to the center of the internal data set on the PCA plot.

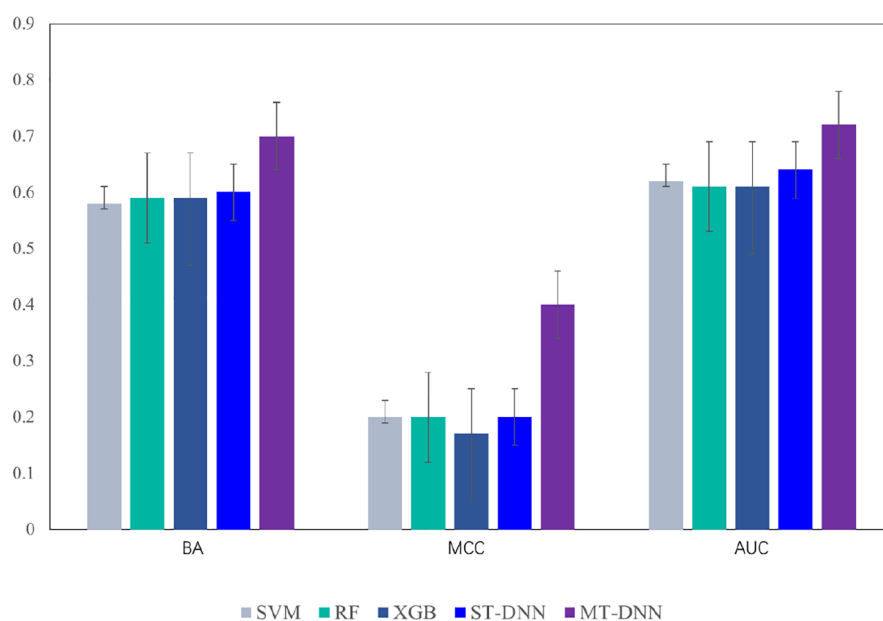


Figure 4. Performance of different models on an external data set. The SVM, RF, XGB, and ST-DNN models are built based on 1384 L858R/T790M/C797S mutation inhibitors in the internal data set. The MT-DNN model is built based on all 2302 compounds containing four different types of EGFR (WT, L858R, L858R/T790M, and L858R/T790M/C797S) inhibitors in the internal data set (SVM: support vector machine; RF: random forest; XGB: XGBoost; ST-DNN: single-target neural network; MT-DNN; multitask neural network).

zine, respectively. Interaction analysis shows that, except for piperazine, other groups can form important hydrogen bonds with EGFR. These preliminary analysis results indeed demonstrate MT-DNN's capability to grasp the core scaffold and important fragment of inhibitors. However, the persuasiveness of this interpretability analysis is questionable due to the lack of high-resolution complex crystal structures for many EGFR inhibitors.

Global Analysis of Model Interpretability. Current interpretability methods still require individual designs for different tasks.⁶⁸ The interpretability of many existing models is often illustrated through case studies,^{43,69} mainly because most small-molecule inhibitors currently lack precise structural information on their interaction with targets. In that case, the interpretability analysis results of structure–activity relationship models can apply only to individual molecules. Such

explanations do not provide an overall situation of the entire model, and the interpretation of a single example also reduces the model's credibility.⁷⁰

To gain a deeper understanding of the decision-making process of the MT-DNN model, we systematically analyze the importance of all substructures within the model. We calculate average SHAP values for the four tasks across 1024 fingerprint bits (Figures S5 and S6).

Unlike other interpretability studies, we focus on Δ SHAP between related tasks (Figure 7 and Table S5). In this study, we proposed the following four interpretable analytical eqs 9–12 to calculate the Δ SHAP values based on the first-, third-, and fourth-generation known inhibitors.

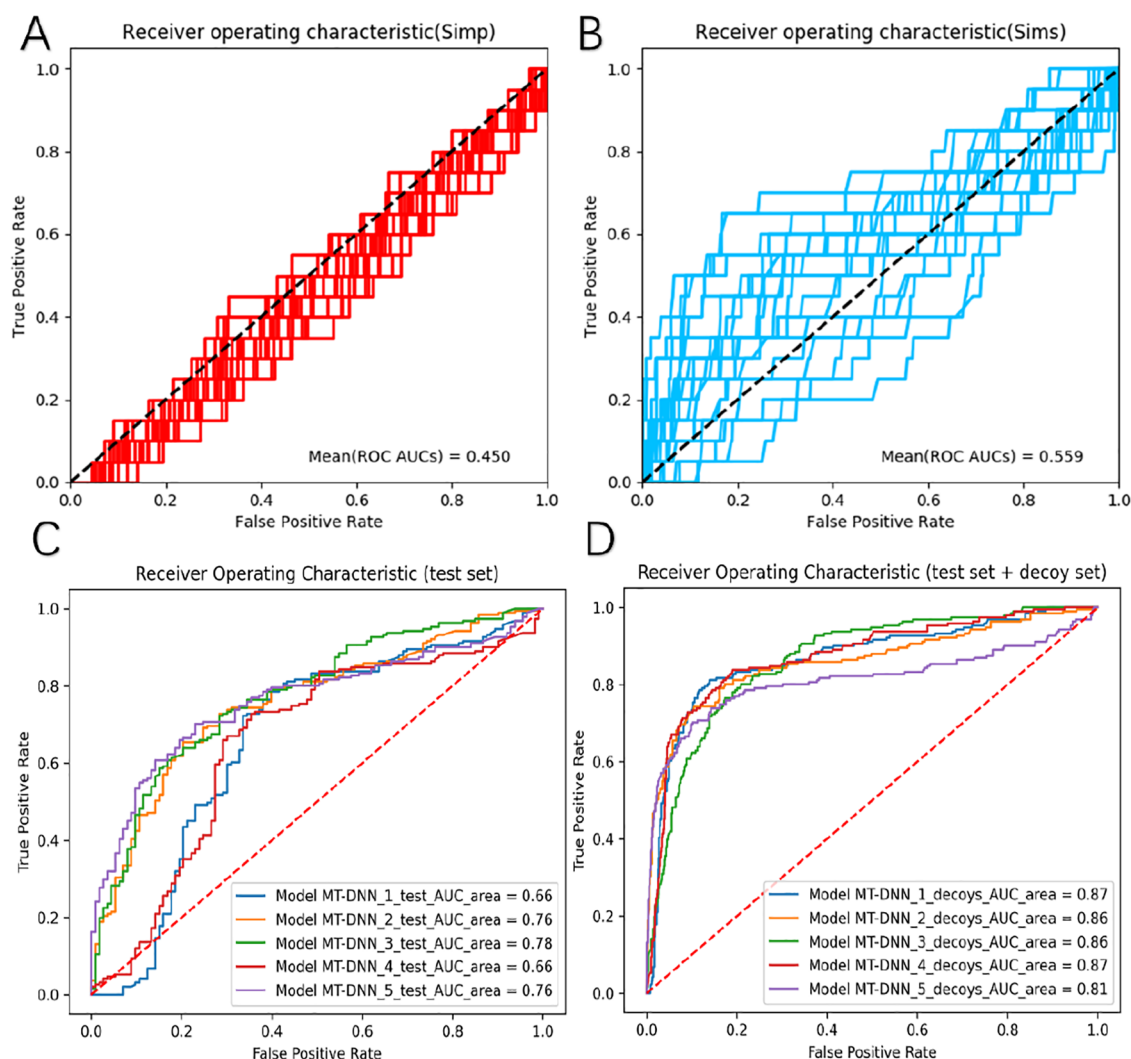


Figure 5. Receiver operating characteristic curve of the external test set and decoy set. (A) ROC based on LOO-CV (leave-one-out cross-validation) using “similarity-in-properties-based”. (B) ROC based on LOO-CV (leave-one-out cross-validation) using MACCS “similarity-in-structure-based” similarity search. (C) ROC for MT-DNN of the external test set only. (D) ROC for MT-DNN of the external test set and decoy set together.).

$$V_{1-3} = [\text{SHAP}(\text{L858R}) + \text{SHAP}(\text{WT})]/2 - \text{SHAP}(\text{L858R}/\text{T790M}) \quad (9)$$

$$V_{3-4} = \text{SHAP}(\text{L858R}/\text{T790M}) - \text{SHAP}(\text{L858R}/\text{T790M}/\text{C797S}) \quad (10)$$

$$V_{3+4} = [\text{SHAP}(\text{L858R}/\text{T790M}/\text{C797S}) + \text{SHAP}(\text{L858R}/\text{T790M})]/2 \quad (11)$$

$$V_{4-1} = \text{SHAP}(\text{L858R}/\text{T790M}/\text{C797S}) - \text{SHAP}(\text{WT}) \quad (12)$$

From Figure 7, one can see that the core scaffold of the first-generation inhibitors, represented by gefitinib and erlotinib, is pyrimidinamine (Figure 7A).³⁷ It can form important water-mediated hydrogen bonds with T790. The T790M resistance mutation disrupts this interaction and creates steric hindrance due to the large side chain of methionine.¹² Additionally, the affinity for ATP increases in the T790M mutant, weakening the competitiveness of the first-generation inhibitors.¹¹ By

calculating eq 9 (Table S6), we could find clues in the TOP20 low ΔSHAP value fingerprints (Figure 7B). Five of the TOP20 low SHAP value fingerprints are highly related to pyrimidinamine. These fingerprints are assigned high SHAP values in WT and L858R inhibitor tasks relatively, and low SHAP values in L858R/T790M inhibitor tasks. ΔSHAP results indicate that the MT-DNN model can identify the deactivation of pyrimidinamine caused by T790M resistance mutation accurately.

The third-generation EGFR inhibitors, represented by osimertinib and olmutinib (Figure 7A), form a covalent bond with EGFR in the active pocket. The acrylamide fragment, acting as an affinity warhead, binds covalently with C797, but this stable mode is disrupted by C797S resistance mutation.^{11,71} It is evident that the development of the fourth-generation EGFR inhibitors should avoid this chemical fragment due to the failure of acrylamide's covalent interaction mode. In the MT-DNN model, ΔSHAP ranking between L858R/T790M/C797S inhibitor tasks and L858R/T790M inhibitor tasks fully reflects the above view (eq 10 and Table S7). Five of the TOP20 low ΔSHAP fingerprints are similar to the acrylamide structure (Figure 7B). It can be learned that

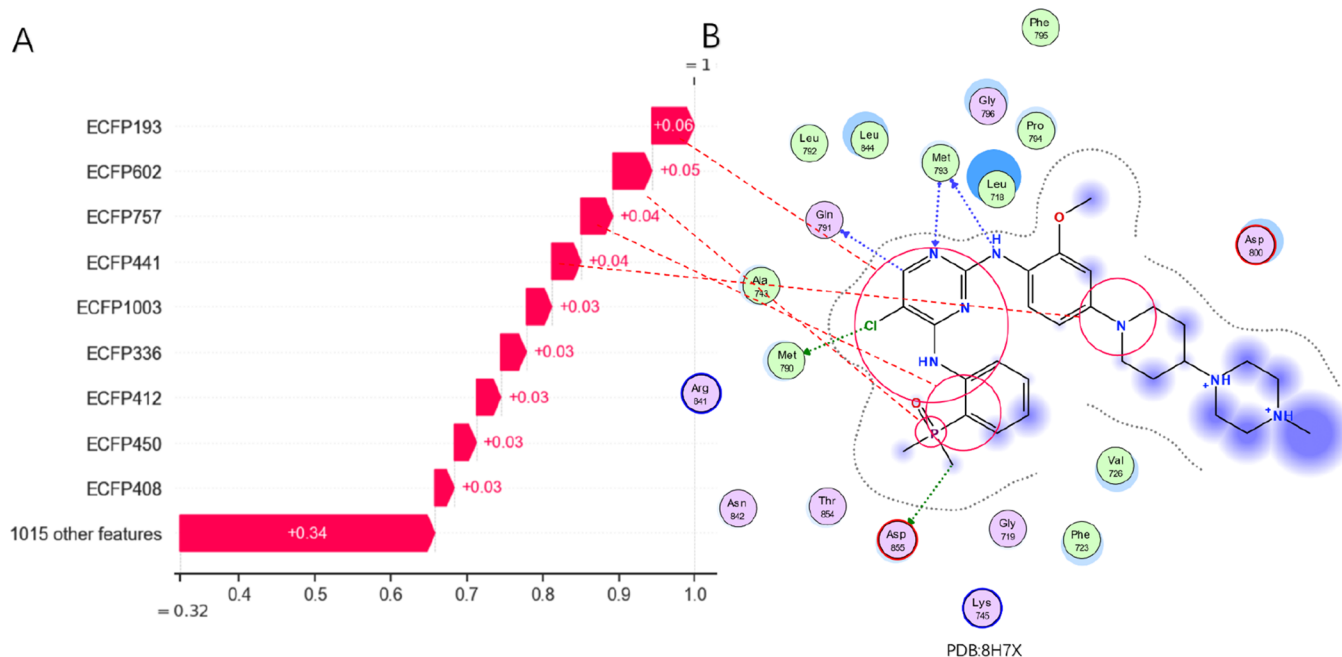


Figure 6. Examples of interpretable analysis of SHAP in the MT-DNN model. (A) Mean arithmetic SHAP values of ECFP_4 fingerprints sorted from the largest to smallest. (B) Interaction analysis plot of the fourth-generation inhibitor brigatinib with EGFR (PDB: 8H7X).

acrylamide is an important fragment in the third-generation EGFR inhibitors but is discarded in the fourth-generation EGFR inhibitors due to the inability to bind S797.

There are diverse strategies for design of the fourth-generation EGFR inhibitors, with one of the most straightforward molecular modifications based on the third-generation EGFR inhibitors. The quinazoline amine group in the third-generation EGFR inhibitors remains effective against the C797S mutation. Quinazoline amine can interact with M793 through hydrogen bond and does not have the spatial hindrance that the pyrimidinamine core scaffold of the first-generation inhibitor has when facing T790M. Reassuringly, by calculating the TOP20 fingerprints with high SHAP in both L858R/T790M/C797S and L858R/T790 M inhibitors tasks (eq 11 and Table S8), we found that the MT-DNN model assigns high importance to the quinazoline amine group. In fact, many successful designs of the fourth-generation EGFR inhibitors include the quinazoline amine group, such as brigatinib.⁶⁷

Selectivity is another important issue for fourth-generation EGFR inhibitors. Fourth-generation EGFR inhibitors can be divided into competitive inhibitors and allosteric inhibitors based on the binding site (Figure 7A). Finding lead compounds that target both orthosteric (ATP binding site) and allosteric sites is one of the methods to improve the selectivity of EGFR mutations.⁷² Although the MT-DNN model in this study does not consider inhibitors only in allosteric sites, the features of inhibitors that simultaneously target competitive and allosteric sites are also captured. Δ SHAP ranking between L858R/T790M/C797S inhibitor tasks and WT inhibitor tasks shows that five of the TOP20 highest fingerprints are related to the important indole fragment of the fourth-generation allosteric inhibitors (Figure 7B, eq 12, and Table S9). This fragment is first proposed by designers of allosteric inhibitors EAI001 and EAI045.⁷³ The MT-DNN model provides a design scheme to improve the selectivity of the fourth-generation EGFR inhibitors, suggesting

that we can use the similarity between competitive and allosteric inhibitors to design dual-target inhibitors.

In summary, the MT-DNN model for EGFR inhibitors, based on both WT and mutant types, benefits from the high correlation between different tasks effectively. In interpretability analysis, Δ SHAP values between different tasks provide a reasonable explanation for inhibitor design from a global perspective. The alignment of interpretability analysis with existing knowledge gives high credibility to the MT-DNN model.

CONCLUSIONS

In this study, we established a multitask deep neural network (MT-DNN) model based on the EGFR inhibitors to predict the high and low inhibitory activity of EGFR WT, L858R, L858R/T790M, and L858R/T790M/C797S mutants. First, we have collected an internal data set containing 2302 EGFR inhibitors, which exhibit cross activity against four different types of EGFR. The inherent high correlation between the activity prediction tasks of these EGFR inhibitors allows the model to learn the distinctions among different generations of inhibitors. In internal 5-fold cluster cross-validation, the average balanced accuracy is 0.82, the MCC is 0.65, and the AUC is 0.88. In addition, the MT-DNN model also performs well on an external data set of 304 fourth-generation EGFR inhibitors, with the MCC value surpassing other single-task models by more than 0.2.

The MT-DNN model not only provides reasonable explanation for EGFR and ligand complexes but also systematically identifies core scaffolds and important fragments of the first-, second-, third-, and fourth-generation EGFR inhibitors by using Δ SHAP values from a global perspective. The MT-DNN model not only catches the inactivation of quinolinamine in the first and second generations of EGFR inhibitors and phenylacrylamide in the third-generation EGFR inhibitors due to resistance mutations in T790M and C797S, respectively, but also captures that the core scaffold pyrimidin-

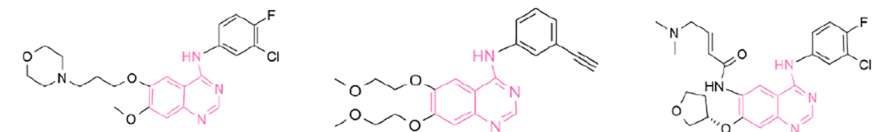
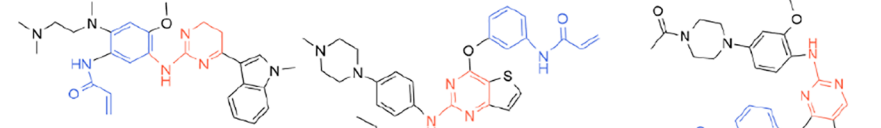
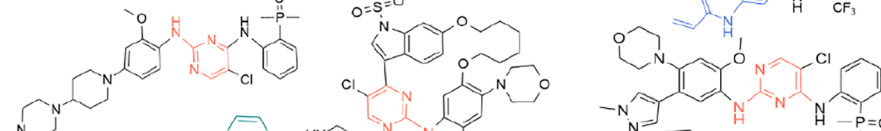
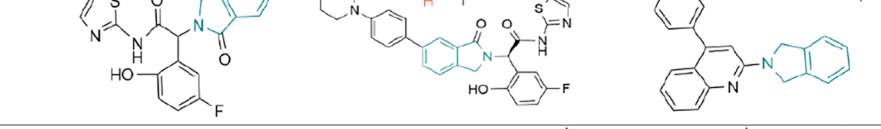
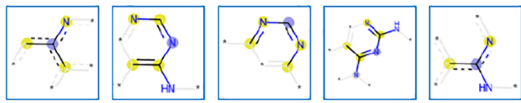
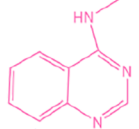

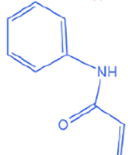

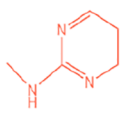
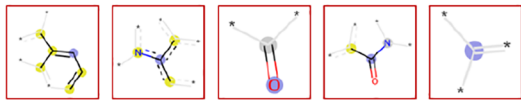
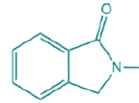
A	Generation of EGFR inhibitors	Representative EGFR inhibitors		
	First and Second generation			
	Third generation			
	Fourth (competitive) generation			
	Fourth (allosteric) generation			
B	Δ SHAP	Top ECFP_4 fingerprint bits	Core scaffold	Source
	SHAP _{L858R/T790M} - (SHAP _{WT} + SHAP _{L858R})/2			First and second generation inhibitors
	SHAP _{L858R/T790M/C797S} - SHAP _{L858R/T790M}			Third generation inhibitors
	(SHAP _{L858R/T790M/C797S} + SHAP _{L858R/T790M})/2			Third and Fourth generation Inhibitors (competitive)
	SHAP _{L858R/T790M/C797S} - SHAP _{WT}			Fourth generation inhibitors (allosteric)

Figure 7. Representative structures of each generation of EGFR inhibitors and global Δ SHAP analysis for the MT-DNN model. (A) Representative drug molecules for each generation of EGFR inhibitors (from left to right and top to bottom respectively: gefitinib; erlotinib; afatinib; osimertinib; olmutinib; rociletinib; brigatinib; compound 18j; compound 12; EAI045; JBJ-04-125-02; compound 4ad). Red and blue highlights on the molecule indicate the core scaffold and important fragments. (B) Global Δ SHAP calculation and analysis of the top-ranked fingerprint segments between tasks of the MT-DNN model. The sum of SHAP values denotes fingerprint segments that are important in both tasks, and the minus of SHAP values denotes fingerprint segments that are important in the former task but not in the latter. Fingerprint segments in red boxes are top-ranked features that have a positive effect on the MT-DNN model; fingerprint segments in blue boxes are top-ranked features that have a negative effect on the MT-DNN model.

amine in the third-generation inhibitors is still important in the setup of the fourth-generation EGFR inhibitors. The MT-DNN model also suggests the design of inhibitors targeting both competitive and allosteric sites for enhancing selectivity of the fourth-generation EGFR inhibitors.

In conclusion, the MT-DNN model distinguishes structural differences among different generations of EGFR inhibitors and provides valuable structure–activity relationship recommendations for the fourth-generation EGFR inhibitors. The combination of the MT-DNN model and interpretability

analysis allows us to discover subtle differences between various types of inhibitors. This research approach is not only applicable to explore the structure–activity relationship among different generations of EGFR inhibitors but also beneficial in drug design scenarios with similar binding sites theoretically. As drug data sharing continues to advance, we believe that this analytical approach will exhibit extensive usage in contexts such as drug repurposing and beyond.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code is available at <https://github.com/ZHIQISUN/MT-DNN-for-EGFR>.

■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.4c10464>.

Internal and external data sets (Tables S1 and S2), Bayesian optimization hyperparameter results for MT-DNN models (Table S3), Tanimoto correlation analysis with activity labels for EGFR and its mutation inhibitors (Table S4), SHAP mean values for 1024 bits of ECFP₄ fingerprints (Table S5), visualization of the first 20 fingerprints for interpretable algorithms (Tables S6–S9), distribution and significance analysis of EGFR inhibitors (Figures S1 and S2), SHAP value analysis in MT-DNN and ST-DNN models (Figures S3, S5, and S6), and scatter plots of ARKA₂ vs ARKA₁ for internal and external data sets (Figure S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Aixia Yan — State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, Beijing University of Chemical Technology, Beijing 100029, China; orcid.org/0000-0003-4811-8510; Phone: +86-10-64455320; Email: yanax@mail.buct.edu.cn; Fax: +86-10-64416428

Authors

Zhiqi Sun — State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Donghui Huo — State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Jiangyu Guo — State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsomega.4c10464>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [22473009] and the “Chemical Grid Project” of Beijing University of Chemical Technology.

■ ABBREVIATIONS

EGFR, epidermal growth factor receptor tyrosine kinase; MT-DNN, multitask deep neural network; ST-DNN, single-task deep neural network; WT, wild type; SVM, support vector machine; RF, random forest; XGB, XGboost; BA, balanced accuracy; SE, sensitivity; SP, specificity; MCC, Mathews correlation coefficient; AUC, area under the curve of receiver operating characteristic; T-SNE, t-distributed stochastic

neighbor embedding; SHAP, SHapley Additive exPlanations; PCA, principal component analysis

■ REFERENCES

- (1) Kumagai, S.; Koyama, S.; Nishikawa, H. Antitumour immunity regulated by aberrant ERBB family signalling. *Nat. Rev. Cancer* **2021**, *21* (3), 181–197.
- (2) Eltayeb, K.; La Monica, S.; Tiseo, M.; Alfieri, R.; Fumarola, C. Reprogramming of Lipid Metabolism in Lung Cancer: An Overview with Focus on EGFR-Mutated Non-Small Cell Lung Cancer. *Cells* **2022**, *11* (3), 413.
- (3) Zhu, S.; Wu, Y.; Song, B.; Yi, M.; Yan, Y.; Mei, Q.; Wu, K. Recent advances in targeted strategies for triple-negative breast cancer. *J. Hematol. Oncol.* **2023**, *16* (1), 100.
- (4) Okuyama, K.; Naruse, T.; Yanamoto, S. Tumor micro-environmental modification by the current target therapy for head and neck squamous cell carcinoma. *J. Exp. Clin. Cancer Res.* **2023**, *42* (1), 114.
- (5) Karlsen, E.-A.; Kahler, S.; Tefay, J.; Joseph, S. R.; Simpson, F. Epidermal Growth Factor Receptor Expression and Resistance Patterns to Targeted Therapy in Non-Small Cell Lung Cancer: A Review. *Cells* **2021**, *10*, 1206.
- (6) Xu, J.; Zhang, X.; Yang, H.; Ding, G.; Jin, B.; Lou, Y.; Zhang, Y.; Wang, H.; Han, B. Comparison of outcomes of tyrosine kinase inhibitor in first- or second-line therapy for advanced non-small-cell lung cancer patients with sensitive EGFR mutations. *Oncotarget* **2016**, *7* (42), 68442–68448.
- (7) Nan, X.; Xie, C.; Yu, X.; Liu, J. EGFR TKI as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer. *Oncotarget* **2017**, *8* (43), 75712–75726.
- (8) Herbst, R. S.; Fukuoka, M.; Baselga, J. Gefitinib — a novel targeted approach to treating cancer. *Nat. Rev. Cancer* **2004**, *4* (12), 956–965.
- (9) Tsao, M. S.; Sakurada, A.; Cutz, J. C.; Zhu, C. Q.; Kamel-Reid, S.; Squire, J.; Lorimer, I.; Zhang, T.; Liu, N.; Daneshmand, M.; et al. Erlotinib in Lung Cancer — Molecular and Clinical Predictors of Outcome. *New Engl. J. Med.* **2005**, *353* (2), 133–144.
- (10) Chen, L.; Fu, W.; Zheng, L.; Liu, Z.; Liang, G. Recent Progress of Small-Molecule Epidermal Growth Factor Receptor (EGFR) Inhibitors against C797S Resistance in Non-Small-Cell Lung Cancer. *J. Med. Chem.* **2018**, *61* (10), 4290–4300.
- (11) Ward, R. A.; Fawell, S.; Floc'h, N.; Flemington, V.; McKerrecher, D.; Smith, P. D. Challenges and Opportunities in Cancer Drug Resistance. *Chem. Rev.* **2021**, *121* (6), 3297–3351.
- (12) Song, Z.; Ge, Y.; Wang, C.; Huang, S.; Shu, X.; Liu, K.; Zhou, Y.; Ma, X. Challenges and Perspectives on the Development of Small-Molecule EGFR Inhibitors against T790M-Mediated Resistance in Non-Small-Cell Lung Cancer. *J. Med. Chem.* **2016**, *59* (14), 6580–6594.
- (13) Dunto, R. T.; Keating, G. M. Afatinib: First Global Approval. *Drugs* **2013**, *73* (13), 1503–1515.
- (14) Deeks, E. D. Neratinib: First Global Approval. *Drugs* **2017**, *77* (15), 1695–1704.
- (15) Shirley, M. Dacomitinib: First Global Approval. *Drugs* **2018**, *78* (18), 1947–1953.
- (16) Eskens, F. A. L. M.; Mom, C. H.; Planting, A. S. T.; Gietema, J. A.; Amelsberg, A.; Huisman, H.; van Doorn, L.; Burger, H.; Stopfer, P.; Verweij, J.; et al. A phase I dose escalation study of BIBW 2992, an irreversible dual inhibitor of epidermal growth factor receptor 1 (EGFR) and 2 (HER2) tyrosine kinase in a 2-week on, 2-week off schedule in patients with advanced solid tumours. *Br. J. Cancer* **2008**, *98* (1), 80–85.
- (17) Greig, S. L. Osimertinib: First Global Approval. *Drugs* **2016**, *76* (2), 263–273.
- (18) Markham, A. Mobocertinib: First Approval. *Drugs* **2021**, *81* (17), 2069–2074.
- (19) Li, Y.; Mao, T.; Wang, J.; Zheng, H.; Hu, Z.; Cao, P.; Yang, S.; Zhu, L.; Guo, S.; Zhao, X.; et al. Toward the next generation EGFR inhibitors: an overview of osimertinib resistance mediated by EGFR

- mutations in non-small cell lung cancer. *Cell Communication and Signaling* **2023**, *21* (1), 71.
- (20) Maity, P.; Chatterjee, J.; Patil, K. T.; Arora, S.; Katiyar, M. K.; Kumar, M.; Samarbakhsh, A.; Joshi, G.; Bhutani, P.; Chugh, M.; et al. Targeting the Epidermal Growth Factor Receptor with Molecular Degradable: State-of-the-Art and Future Opportunities. *J. Med. Chem.* **2023**, *66* (5), 3135–3172.
- (21) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18* (6), 463–477.
- (22) Banerjee, A.; Roy, K.; Gramatica, P. A bibliometric analysis of the Cheminformatics/QSAR literature (2000–2023) for predictive modeling in data science using the SCOPUS database. *Mol. Diversity* **2024**, 1–13.
- (23) Qureshi, R.; Zou, B.; Alam, T.; Wu, J.; Lee, V. H. F.; Yan, H. Computational Methods for the Analysis and Prediction of EGFR-Mutated Lung Cancer Drug Resistance: Recent Advances in Drug Design, Challenges and Future Prospects. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2023**, *20* (1), 238–255.
- (24) Kharkar, P. S. Computational Approaches for the Design of (Mutant-)Selective Tyrosine Kinase Inhibitors: State-of-the-Art and Future Prospects. *Curr. Top. Med. Chem.* **2020**, *20* (17), 1564–1575.
- (25) Huo, D.; Wang, S.; Kong, Y.; Qin, Z.; Yan, A. Discovery of Novel Epidermal Growth Factor Receptor (EGFR) Inhibitors Using Computational Approaches. *J. Chem. Inf. Model* **2022**, *62* (21), 5149–5164.
- (26) Huo, D.; Sun, Z.; Wang, M.; Yan, A. Ligand and structure based hierarchical virtual screening cascade for finding novel epidermal growth factor receptor inhibitors. *Chem. Biol. Drug Des.* **2024**, *103* (1), No. e14375.
- (27) Huo, D.; Wang, H.; Qin, Z.; Tian, Y.; Yan, A. Building 2D classification models and 3D CoMSIA models on small-molecule inhibitors of both wild-type and T790M/L858R double-mutant EGFR. *Mol. Divers.* **2022**, *26* (3), 1715–1730.
- (28) 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.
- (29) Zhang, Y.; Li, Y. Machine learning method aided discovery of the fourth-generation EGFR inhibitors. *New J. Chem.* **2023**, *47* (46), 21513–21525.
- (30) Nada, H.; Gul, A. R.; Elkamhaw, A.; Kim, S.; Kim, M.; Choi, Y.; Park, T. J.; Lee, K. Machine Learning-Based Approach to Developing Potent EGFR Inhibitors for Breast Cancer—Design, Synthesis, and In Vitro Evaluation. *ACS Omega* **2023**, *8* (35), 31784–31800.
- (31) Singh, H.; Singh, S.; Singla, D.; Agarwal, S. M.; Raghava, G. P. S. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* **2015**, *10* (1), 10.
- (32) Chauhan, J. S.; Dhanda, S. K.; Singla, D.; Open Source Drug Discovery Consortium; Agarwal, S. M.; Raghava, G. P. S. QSAR-Based Models for Designing Quinazoline/Imidazothiazoles/Pyrazolopyrimidines Based Inhibitors against Wild and Mutant EGFR. *PLoS One* **2014**, *9* (7), No. e101079.
- (33) Chen, X.; Xie, W.; Yang, Y.; Hua, Y.; Xing, G.; Liang, L.; Deng, C.; Wang, Y.; Fan, Y.; Liu, H.; et al. Discovery of Dual FGFR4 and EGFR Inhibitors by Machine Learning and Biological Evaluation. *J. Chem. Inf. Model* **2020**, *60* (10), 4640–4652.
- (34) Allen, B. K.; Mehta, S.; Ember, S. W. J.; Schonbrunn, E.; Ayad, N.; Schürer, S. C. Large-Scale Computational Screening Identifies First in Class Multitarget Inhibitor of EGFR Kinase and BRD4. *Sci. Rep.* **2015**, *5* (1), 16924.
- (35) Saini, R.; Agarwal, S. M. EGFRisopred: a machine learning-based classification model for identifying isoform-specific inhibitors against EGFR and HER2. *Mol. Divers.* **2022**, *26* (3), 1531–1543.
- (36) Cichońska, A.; Ravikumar, B.; Rahman, R. AI for targeted polypharmacology: The next frontier in drug discovery. *Curr. Opin. Struct. Biol.* **2024**, *84*, No. 102771.
- (37) Pal, R.; Teli, G.; Matada, G. S. P.; Dhiwar, P. S. Designing strategies, structural activity relationship and biological activity of recently developed nitrogen containing heterocyclic compounds as epidermal growth factor receptor tyrosinase inhibitors. *J. Mol. Struct.* **2023**, *1291*, No. 136021.
- (38) Mitrasinovic, P. M. Progress in structure-based design of EGFR inhibitors. *Curr. Drug Targets* **2013**, *14* (7), 817–829.
- (39) Ayati, A.; Moghimi, S.; Toolabi, M.; Foroumadi, A. Pyrimidine-based EGFR TK inhibitors in targeted cancer therapy. *Eur. J. Med. Chem.* **2021**, *221*, No. 113523.
- (40) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv pre-print server* **2015**. DOI: DOI: 10.48550/arXiv.1502.02072
- (41) Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv pre-print server* **2017**. DOI: DOI: 10.48550/arXiv.1706.05098.
- (42) Ren, Q.; Qu, N.; Sun, J.; Zhou, J.; Liu, J.; Ni, L.; Tong, X.; Zhang, Z.; Kong, X.; Wen, Y.; et al. KinomeMETA: meta-learning enhanced kinome-wide polypharmacology profiling. *Briefings Bioinf.* **2023**, *25* (1), No. bbad461.
- (43) Bao, L.; Wang, Z.; Wu, Z.; Luo, H.; Yu, J.; Kang, Y.; Cao, D.; Hou, T. Kinome-wide polypharmacology profiling of small molecules by multi-task graph isomorphism network approach. *Acta Pharmaceutica Sinica B* **2023**, *13* (1), 54–67.
- (44) Hu, J.; Allen, B. K.; Stathias, V.; Ayad, N. G.; Schürer, S. C. Kinome-Wide Virtual Screening by Multi-Task Deep Learning. *Int. J. Mol. Sci.* **2024**, *25* (5), 2538.
- (45) Li, X.; Li, Z.; Wu, X.; Xiong, Z.; Yang, T.; Fu, Z.; Liu, X.; Tan, X.; Zhong, F.; Wan, X.; et al. Deep Learning Enhancing Kinome-Wide Polypharmacology Profiling: Model Construction and Experiment Validation. *J. Med. Chem.* **2020**, *63* (16), 8723–8737.
- (46) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4* (2), 4367–4375.
- (47) de la Vega de León, A.; Chen, B.; Gillet, V. J. Effect of missing data on multitask prediction methods. *J. Cheminform.* **2018**, *10* (1), 26.
- (48) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model* **2017**, *57* (10), 2490–2504.
- (49) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model* **2017**, *57* (8), 2068–2076.
- (50) Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci.* **2018**, *20* (11), 621–632.
- (51) Speck-Planche, A. Combining Ensemble Learning with a Fragment-Based Topological Approach to Generate New Molecular Diversity in Drug Discovery: In Silico Design of Hsp90 Inhibitors. *ACS Omega* **2018**, *3* (11), 14704–14716.
- (52) Speck-Planche, A. Multicellular Target QSAR Model for Simultaneous Prediction and Design of Anti-Pancreatic Cancer Agents. *ACS Omega* **2019**, *4* (2), 3122–3132.
- (53) Speck-Planche, A.; Cordeiro, M. N. D. S. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, *21* (3), 511–523.
- (54) Speck-Planche, A.; Scotti, M. T. BET bromodomain inhibitors: fragment-based in silico design using multi-target QSAR models. *Mol. Divers.* **2019**, *23* (3), 555–572.
- (55) Ridley, D. D. Strategies for Chemical Reaction Searching in SciFinder. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (5), 1077–1084.
- (56) 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* (D1), D1100–D1107.
- (57) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined

protein–ligand binding affinities. *Nucleic Acids Res.* **2006**, 35 (suppl_1), D198–D201.

(58) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model* **2009**, 49 (12), 2897–2898.

(59) Xia, J.; Li, S.; Ding, Y.; Wu, S.; Wang, X. S. MUBD-DecoyMaker 2.0: A Python GUI Application to Generate Maximal Unbiased Benchmarking Data Sets for Virtual Drug Screening. *Mol. Inform.* **2020**, 39 (4), No. 1900151.

(60) Ansel, J.; Yang, E.; He, H.; Gimelshein, N.; Jain, A.; Voznesensky, M.; Bao, B.; Bell, P.; Berard, D.; Burovski, E. et al. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, La Jolla, CA, USA, 2024.

(61) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, 50 (5), 742–754.

(62) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, 104 (1), 148–175.

(63) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, 39 (4), 747–750.

(64) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 30, 4768–4777.

(65) Qin, L.-T.; Zhang, J.-Y.; Nong, Q.-Y.; Xu, X.-C.-L.; Zeng, H.-H.; Liang, Y.-P.; Mo, L.-Y. Classification and regression machine learning models for predicting the combined toxicity and interactions of antibiotics and fungicides mixtures. *Environ. Pollut.* **2024**, 360, No. 124565.

(66) Banerjee, A.; Roy, K. ARKA: a framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data. *Environmental Science: Processes & Impacts* **2024**, 26 (6), 991–1007.

(67) Suzuki, M.; Uchibori, K.; Oh-hara, T.; Nomura, Y.; Suzuki, R.; Takemoto, A.; Araki, M.; Matsumoto, S.; Sagae, Y.; Kukimoto-Niino, M.; et al. A macrocyclic kinase inhibitor overcomes triple resistant mutations in EGFR-positive lung cancer. *npj Precision Oncology* **2024**, 8 (1), 46.

(68) Kırboğa, K. K.; Abbasi, S.; Küçüksille, E. U. Explainability and white box in drug discovery. *Chem. Biol. Drug Des.* **2023**, 102 (1), 217–233.

(69) Qian, X.; Dai, X.; Luo, L.; Lin, M.; Xu, Y.; Zhao, Y.; Huang, D.; Qiu, H.; Liang, L.; Liu, H.; et al. An Interpretable Multitask Framework BiLAT Enables Accurate Prediction of Cyclin-Dependent Protein Kinase Inhibitors. *J. Chem. Inf. Model* **2023**, 63 (11), 3350–3368.

(70) Fan, Y. W.; Liu, W. h.; Chen, Y. T.; Hsu, Y. C.; Pathak, N.; Huang, Y. W.; Yang, J. M. Exploring kinase family inhibitors and their moiety preferences using deep SHapley additive exPlanations. *BMC Bioinformatics* **2022**, 23 (4), 242.

(71) Hossam, M.; Lasheen, D. S.; Abouzid, K. A. M. Covalent EGFR Inhibitors: Binding Mechanisms, Synthetic Approaches, and Clinical Profiles. *Arch. Pharm. (Weinheim)* **2016**, 349 (8), 573–593.

(72) Xu, L.; Xu, B.; Wang, J.; Gao, Y.; He, X.; Xie, T.; Ye, X.-Y. Recent advances of novel fourth generation EGFR inhibitors in overcoming C797S mutation of lung cancer therapy. *Eur. J. Med. Chem.* **2023**, 245, No. 114900.

(73) Zhao, P.; Yao, M.-Y.; Zhu, S.-J.; Chen, J.-Y.; Yun, C.-H. Crystal structure of EGFR T790M/C797S/V948R in complex with EAI045. *Biochem. Biophys. Res. Commun.* **2018**, 502 (3), 332–337.