ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

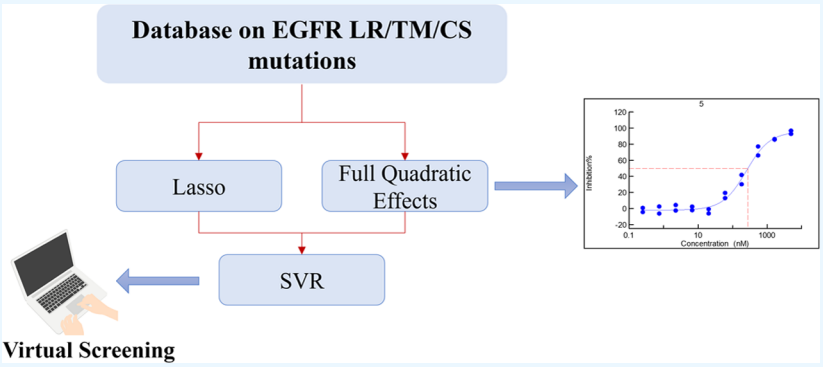
Machine Learning-Based Virtual Screening and Identification of the Fourth-Generation EGFR Inhibitors

Hao Chang,[#] Zeyu Zhang,[#] Jiaxin Tian, Tian Bai, Zijie Xiao, Dianpeng Wang,* Renzhong Qiao,* and Chao Li*

Cite This: *ACS Omega* 2024, 9, 2314−2324

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



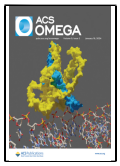
ABSTRACT: Epidermal growth factor receptor (EGFR) plays a pivotal regulatory role in treating patients with advanced nonsmall cell lung cancer (NSCLC). Following the emergence of the EGFR tertiary CIS C797S mutation, all types of inhibitors lose their inhibitory activity, necessitating the urgent development of new inhibitors. Computer systems employ machine learning methods to process substantial volumes of data and construct models that enable more accurate predictions of the outcomes of new inputs. The purpose of this article is to uncover innovative fourth-generation epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs) with the aid of machine learning techniques. The paper's data set was high-dimensional and sparse, encompassing both structured and unstructured descriptors. To address this considerable challenge, we introduced a fusion framework to select critical molecule descriptors by integrating the full quadratic effect model and the Lasso model. Based on structural descriptors obtained from the full quadratic effect model, we conceived and synthesized a variety of small-molecule inhibitors. These inhibitors demonstrated potent inhibitory effects on the two mutated kinases L858R/T790M/C797S and Del19/T790M/C797S. Moreover, we applied our model to virtual screening, successfully identifying four hit compounds. We have evaluated these hit ADME characteristics and look forward to conducting activity evaluations on them in the future to discover a new generation of EGFR-TKI.

1. INTRODUCTION

The epidermal growth factor receptor is a kinase within the human epidermal growth factor receptor (HER) family and holds a pivotal regulatory role across multiple malignancies. This protein family serves as an appealing drug target, particularly within the context of nonsmall cell lung cancer.[1−4] In recent times, with the intensification of research into tumor targets, targeted antitumor medications, typified by epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs), have emerged as the primary treatment option for patients in the advanced stages of nonsmall cell lung cancer (NSCLC). These drugs have demonstrated notable efficacy, minimal adverse effects, and extended periods of disease-free progression survival post treatment.

The EGFR-TKIs, which include the first generation: erlotinib[5] and gefitinib,[6] the second generation: dacotinib[7] and afatinib,[8] and the third generation: osimertinib,[9] are extensively employed in clinical practice, catering to around 90% of patients possessing EGFR-sensitizing mutations. These inhibitors are commonly administered as standalone treatments or in conjunction with monoclonal antibodies to address patient conditions. Unfortunately, prolonged utilization of EGFR inhibitors has encountered substantial challenges due to the emergence of acquired drug resistance. This phenomenon has led to a reduction in clinical effectiveness.[10−12] Recently, approximately 40% of all resistance cases are caused by C797S mutations, making them the primary cause of resistance to third-generation inhibitors.[13,14] Encouraging strides have been made

Received: August 22, 2023
Revised: November 6, 2023
Accepted: November 15, 2023
Published: January 2, 2024



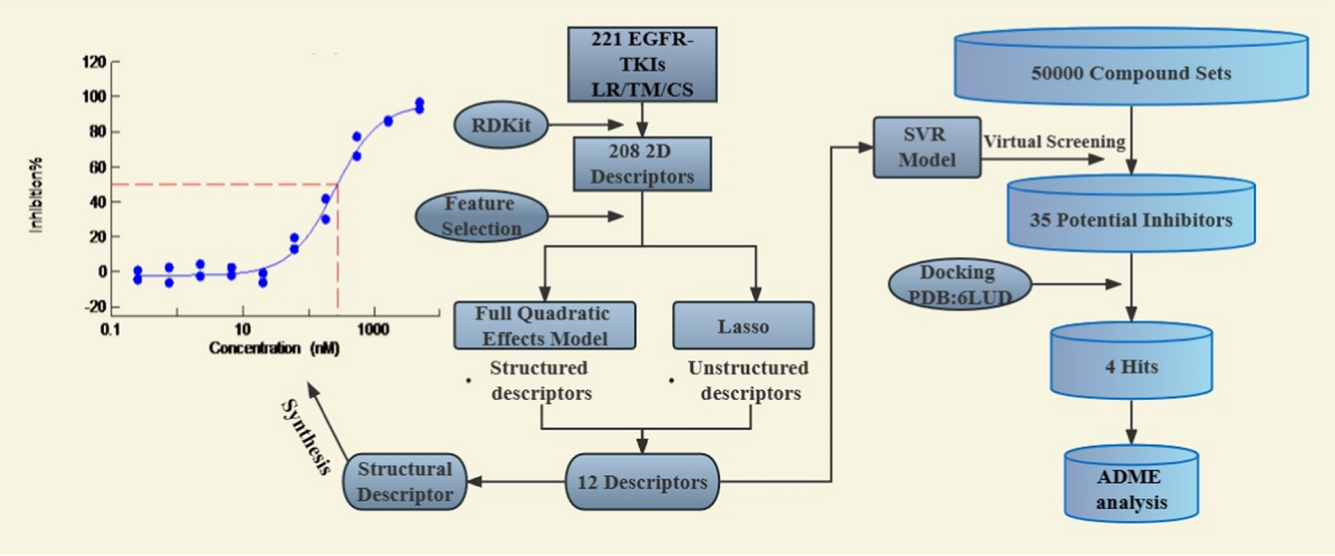


Figure 1. Flowchart of molecular descriptor selection, model construction, and virtual screening.

in the development of fourth-generation EGFR-TKIs. Drugs like BLU-945,[15] BLU-701,[16] TBQ3804,[17] BBT-176,[18] BPI-361175,[19] and others have progressed into the realm of clinical research. However, the advancement of most other drugs to clinical studies has faltered, and none have successfully cleared phase III clinical trials. These drugs exhibit diverse structures, including isoindoline-1-one derivatives, 4-amine-pyrimidine derivatives, quinoline derivatives, 2-amino-pyrimidine derivatives, and more. Hence, further investigation remains imperative to uncover the fourth generation of novel EGFR-TKIs.

Computational modeling brings significant advantages to drug discovery and has found extensive application in computer-aided drug discovery.[20−22] In recent years, numerous QSAR models for EGFR-TKIs have been developed. In the development of wild-type and monomutant EGFR inhibitor models, Chauhan et al.[23] collected 128 quinazoline inhibitors and developed wild-type EGFR and monomutant EGFR (L858R) support vector machine (SVM) models. The coefficients of determination (R^2) for the wild-type and mutant models reached 0.83 and 0.71, respectively. Yan et al. collected a data set of 1248 EGFR inhibitors and built two classification models using the methods of Kohonen's self-organizing map (SOM) and SVM.[24] The SOM model had prediction accuracy rates of 98.5 and 96.3% on the training set and test set, respectively, while the SVM model had rates of 99.0 and 97.0%, respectively. They further developed two-dimensional (2D) classification models and three-dimensional (3D) CoMSIA models for wild-type and L858R/T790M double mutant EGFR-TKI.[25] For 2D models, the accuracy of each model was greater than 0.87.

Significant strides have been made in the QSAR modeling of EGFR tyrosine kinase inhibitors. However, there are still gaps when it comes to modeling small-molecule inhibitors targeting the $EGFR^{L858R/T790M/C797S}$ triple mutation. This is due to the short duration of inhibitor studies focused on the C797S mutation, leading to a small sample size of compounds. Because of the resulting structural diversity stemming from the relatively low number of inhibitors targeting triple mutations, it becomes imperative to train models using known activities and structures to propel advancements in medicinal chemistry studies. This paper intends to develop an advanced machine learning model capable of identifying small-molecule inhibitors that zero in on the $EGFR^{L858R/T790M/C797S}$ triple mutation. Furthermore, this model will be applied to virtually screen potential inhibitors, and molecular structure descriptors will be extracted to guide the synthesis of inhibitors (Figure 1). A considerable portion of the existing literature combines all of the descriptors into a single feature selection method for screening purposes, thereby rendering it difficult to acquire precise structural descriptors for guiding compound synthesis. In this context, we introduced an integrated framework that initially employs the full quadratic effect model to screen structural descriptors, followed by utilizing Lasso for screening nonstructural descriptors. This approach guarantees the acquisition of essential structural molecular descriptors crucial for predicting activity, thereby furnishing a broader array of avenues for designing molecular structures. To initiate the process, we amassed a data set comprising 221 small-molecule inhibitors specifically targeting the $EGFR^{L858R/T790M/C797S}$ triple mutation. The open-source toolkit RDKit was employed to condense the SMILES representation of each inhibitor, yielding a comprehensive set of descriptors elucidating structural information. Subsequently, an integrated framework combining full quadratic effect modeling and Lasso modeling was employed to meticulously select pivotal descriptors, encompassing both structural and general descriptors. These acquired descriptors were then employed in training the support vector regression (SVR) model. Concurrently, the structural descriptors played a pivotal role in steering the synthesis of small-molecule inhibitors engineered to target the $EGFR^{L858R/T790M/C797S}$ triple mutation. This was corroborated through kinase inhibition activity experiments. Furthermore, our model was extended to virtual screening, wherein molecular simulation docking techniques were harnessed to identify the potential EGFR-TKIs. The safety of these identified hits was subjected to rigorous analysis via an ADME evaluation. Ultimately, by leveraging the structural descriptors as guidance, three inhibitors were meticulously designed and synthesized. In-depth *in vitro* kinase activity assays were subsequently conducted to validate their efficacy in inhibiting the activity of the triple mutant kinases, with varying degrees of success. Moreover, virtual screening resulted in the identification of four hits with the potential to inhibit triple

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

mutant kinases. These hits are now poised for further in-depth investigation and study.

2. MATERIALS AND METHODS

2.1. Data Preparation. We collected 30 papers published in recent years, which focus on small-molecule inhibitors of EGFR $^{L858R/T790M/C797S}$ triple mutations and compiled a data set containing 221 valuable active structures and their corresponding kinase inhibitory activities (expressed as IC_{50}).[26−55] For screening efficiency, only structures with $IC_{50} < 5\ \mu M$ and with accurate numerical representation have been retained in the collation. The activity values were converted to the $pIC_{50}(-\log_{10}(IC_{50}))$ for the sake of the same magnitude on our data.

2.2. Data Preprocessing. In order to validate the performance of the machine learning models, the whole data set was split into a training and a test set at a 3:1 ratio randomly, whose training set included 165 compounds and the test set included 56 compounds.

An important challenge in computational chemistry was to present molecular structures as pieces of information so that computers can process them and use them to train models. The method we used here is to convert the compound SMILES formula into 2D descriptors via RDKit,[56] which is an open-source toolkit for cheminformatics based on 2D and 3D molecular manipulations of compounds. The generated molecular descriptors capture various features of individual molecules, including molecular properties, connectivity, composition, topological information, and MOE-type (molecular operating environment) information.

In this work, calculated molecule descriptors and the pIC_{50} values of the whole data set were standardized using the following equation

$$x_i^* = \frac{x_i - x_{\text{mean}}}{x_{\text{std}}} \tag{1}$$

where x_i^* represents the scaled value; x_i represents the original value; x_{mean} represents the mean value; and x_{std} represents the standard deviation (SD).

2.3. Feature Selection. There are two types of descriptors, structural and nonstructural descriptors, in the data set. The structural descriptors can help to filter the new drugs, and the nonstructural descriptors can benefit the prediction of activity of new drugs. It is crucial to select both useful descriptors that will be used to train the prediction model. Thus, in this work, we proposed an integrated framework by using the full quadratic effect model and the Lasso model for the selection of important molecular descriptors to develop a machine learning model and guide the subsequent inhibitor synthesis. We used the full quadratic effect model to take the lead in selecting the descriptors with significant effects from the structural molecular descriptors. On the other hand, given that the data set used in this paper was high-dimensional and sparse, where the number of variables (221) and the number of features (208) were close, the Lasso model was used to filter significant nonstructural descriptors from the 208 molecular descriptors calculated by RDKit. Finally, the descriptors screened from the fusion model described above were fitted together to build a machine learning model and guide the synthesis of inhibitors.

2.3.1. Structural Descriptor Selection. The full quadratic effect model is a type of fixed effect model that describes the causal relationship between the factors and the response that persists throughout the experiment. It reveals the causal relationship between the factors and response by employing a regression model to find out the useful first-order effects, second-order effects, and interaction effects between factors (eq 2) that have a significant effect on the dependent variable

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_1}^2 + \beta_4 {x_2}^2 + \beta_5 x_1 x_2 + \varepsilon \tag{2}$$

where Y represents the response, x_1 and x_2 represent the effects of factors, ε represents the random error term, and β_0−β_5 represents the parameters.

In practice, the model can be estimated and extrapolated by statistical software to understand the relationships among the independent variables in order to better understand and interpret the data. In this paper, the OLS model was built using the statsmodels package in Python, containing the first-order and interaction effects of the variables as well as the second-order effects, with the size of the p-value used to determine whether the effect was significant.

2.3.2. Lasso. Lasso regression is a widely used statistical algorithm for variable selection. It is a least-squares with a penalty term (eq 3). Unlike the L2 penalty term in ridge regression, the L1 penalty term in Lasso regression achieves feature selection by compressing the coefficients of insignificant variables to zero without setting any value to zero. Therefore, for high-dimensional sparse data sets, it is more suitable to use Lasso regression to select features

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - -\sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j|,$$
$$\text{for some } t > 0, \sum_{j=0}^{p} |w_j| < t \tag{3}$$

2.4. Activity Prediction Model Development and Evaluation. *2.4.1. SVR Model.* Support vector regression (SVR) model is a widely used supervised learning model for regression analysis, which was first proposed by Drucker et al.[57] It translates the low-dimensional linearly indivisible data set into a high-dimensional space via a nonlinear transform defined by inner product function and then calculates the optimal hyperplane by minimizing the distance to the furthest data point from the hyperplane. The introduction of SVR is available in ref 58.

In this work, we used the RBF kernel function (eq 4), which transforms the input vector from a low-dimensional space to a high-dimensional space and generates outputs through the weighted summation of its hidden units. As a result, it is suitable for tackling nonlinear problems. The optimization of parameters C, ε (eq 5), and γ was taken out by using the lattice search method with 5-fold cross-validation. In 5-fold cross-validation, the training data were randomly divided into a training set and a test set five times, and we scored each test set by the mean-squared error (MSE), with the best average score indicating the best hyperparameters. We repeated the 5-fold cross-validation 10 times and obtained the average score for each set of hyperparameters. Then, we averaged these MSE values, with the minimal mean MSE value corresponding to the optimal hyperparameter

$$K(x, y) = -\gamma \| x - y \|^2 \tag{4}$$

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

$$\min\left[\frac{1}{2}w'w + C\sum_{n=1}^{N}(\xi_n + \xi_n^*)\right]$$
$$\text{s.t. } \forall\, n: y_n - (x_n'w + b) \le \varepsilon + \xi_n,$$
$$(x_n'w + b) - y_n \le \varepsilon + \xi_n^*,$$
$$\xi_n \ge 0,\ \xi_n^* \ge 0 \quad (5)$$

2.4.2. Model Evaluation. In this work, two universal statistic metrics (R^2 and MSE) were applied to assess the model effectiveness. The R^2 and MSE were calculated by eqs 6 and 7, respectively

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \quad (6)$$

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

where n represents the total number of compounds; y represents the observed value of a compound; $\hat{y}$ represents the predicted value of a compound; and $\overline{y}$ represents the average of y.

2.5. General Procedure for Synthesis. The coupling reaction of 2,4-dichloroquinazoline and *tert*-butyl-4-(2-aminoethyl)piperazine-1-carboxylate was catalyzed by triethylamine. The product was coupled with phenylboronic acid with different structures under a palladium catalyst. The tertiary butoxycarbonyl protecting group and methyl group of the compound were removed in the presence of boron tribromide to give the final product. The reaction steps are shown in the S1.

*2.5.1. tert-Butyl-4-(2-((2-chloroquinazolin-4-yl)amino)-ethyl)piperazine-1-carboxylate (**2**).* 1.75 g (7.63 mmol) of *tert*-butyl-4-(2-aminoethyl)piperazine-1-carboxylate was added to a 100 mL round-bottom flask and 30 mL of tetrahydrofuran was dissolved; 1 mL (7.63 mmol) of triethylamine and 1.51 g (7.63 mmol) of 2,4-dichloroquinazoline were added to the reaction system and stirred at 25 °C for 4 h. After the reaction, it was concentrated under reduced pressure, dissolved in 20 mL of DCM, and extracted three times with a saturated sodium carbonate solution. The organic phase was combined and concentrated under reduced pressure. A yellow oily product of 2.5 g was obtained through silica gel column chromatography and purification, with a yield of 83%.

A white solid; yield: 83%; ^{1}H NMR (400 MHz, chloroform-*d*) δ 7.81−7.68 (m, 3H), 7.50 (ddd, J = 8.2, 6.4, 1.8 Hz, 1H), 3.81−3.71 (m, 2H), 3.53 (q, J = 4.7, 4.1 Hz, 4H), 2.78 (t, J = 5.9 Hz, 2H), 2.60−2.49 (m, 4H), 1.49 (s, 9H).

*2.5.2. tert-Butyl-4-(2-((2-(2-hydroxyphenyl)quinazolin-4-yl)amino)ethyl)piperazine-1-carboxylate (**3a**).* 0.50 g (1.28 mmol) of compound **2**, 0.17 g (1.28 mmol) of (2-hydroxyphenyl) boronic acid, and 0.016 g (0.011 mmol) of *tetra* (triphenylphosphine) palladium were added in a 100 mL three-necked flask, placing the reaction system in a nitrogen atmosphere. 10 mL of 2 M sodium carbonate solution, 10 mL of *n*-butanol, and 6 mL of toluene were added in sequence and reacted at 110 °C for 18 h. During this time, TLC monitors the reaction process. After the reaction was completed, heating was stopped and the reaction system was filtered through diatomaceous earth and washed with ethyl acetate; the filtrate was extracted with saturated sodium bicarbonate solution three times, the organic phase was merged, and a white solid product of 0.42 g was obtained by silica gel column chromatography, with a yield of 75%.

A white solid; yield: 75%; ^{1}H NMR (400 MHz, chloroform-*d*) δ 8.54 (dd, J = 7.9, 1.8 Hz, 1H), 7.77−7.67 (m, 2H), 7.40 (m, J = 23.0, 8.6, 7.0, 1.6 Hz, 2H), 7.04 (dd, J = 8.2, 1.2 Hz, 1H), 6.94 (ddd, J = 8.2, 7.1, 1.2 Hz, 1H), 6.73 (t, J = 4.5 Hz, 1H), 3.84 (q, J = 5.7 Hz, 2H), 3.52 (t, J = 5.1 Hz, 4H), 2.79 (t, J = 6.0 Hz, 2H), 2.54 (t, J = 5.1 Hz, 4H), 1.50 (s, 9H).

*2.5.3. tert-Butyl-4-(2-((2-(5-cyano-2-hydroxyphenyl)-quinazolin-4-yl)amino)ethyl)piperazine-1-carboxylate (**3b**).* A white solid; yield: 72%; ^{1}H NMR (400 MHz, chloroform-*d*) δ 8.07 (d, J = 2.2 Hz, 1H), 7.93−7.87 (m, 1H), 7.81−7.71 (m, 2H), 7.67 (dd, J = 8.6, 2.2 Hz, 1H), 7.49 (ddd, J = 8.3, 7.1, 1.2 Hz, 1H), 7.06 (d, J = 8.7 Hz, 1H), 3.91 (s, 3H), 3.75 (q, J = 5.5 Hz, 2H), 3.49 (t, J = 5.0 Hz, 4H), 2.75 (t, J = 6.0 Hz, 2H), 2.51 (t, J = 5.0 Hz, 4H), 1.47 (s, 10H).

*2.5.4. tert-Butyl-4-(2-((2-(5-chloro-2-hydroxyphenyl)-quinazolin-4-yl)amino)ethyl)piperazine-1-carboxylate (**3c**).* A yellow solid; yield: 72%; ^{1}H NMR (400 MHz, chloroform-*d*) δ 7.90 (dd, J = 8.4, 1.2 Hz, 1H), 7.77−7.71 (m, 2H), 7.45 (ddd, J = 8.3, 7.0, 1.2 Hz, 1H), 7.31 (dd, J = 8.8, 2.8 Hz, 1H), 6.93 (d, J = 8.8 Hz, 1H), 6.73 (t, J = 4.7 Hz, 1H), 3.82 (s, 3H), 3.74 (q, J = 5.6 Hz, 2H), 3.48 (t, J = 5.0 Hz, 4H), 2.71 (t, J = 6.0 Hz, 2H), 2.49 (t, J = 5.0 Hz, 4H), 1.47 (s, 9H).

*2.5.5. 2-(4-((2-(Piperazin-1-yl)ethyl)amino)quinazolin-2-yl)phenol (**4a**).* 0.20 g (0.44 mmol) of compound **3a** was added to a 50 mL Schlenk flask. 15 mL of anhydrous dichloromethane was dissolved in a nitrogen atmosphere and stirred at −30 °C for 15 min; then, 514 μL (5.34 mmol) of boron tribromide was added, and after 1 h, the reaction was moved to room temperature. After the reaction, 15 mL of methanol was added to a low-temperature bath at −30 °C to quench the reaction. The solvent was removed by vacuum distillation and dissolved in 20 mL of ultrapure water. The pH was adjusted to 12 or above, extracted three times using a mixture of 18 mL of dichloromethane and 2 mL of methanol, and the organic phase was merged. Vacuum distillation was used to remove solvents, and reverse phase silica gel column chromatography was performed to obtain 102 mg of green solid product with a yield of 65%.

A white solid; yield: 65%; mp: 170.6−173.3 °C; ^{1}H NMR (400 MHz, methanol-d_4) δ 8.44 (dd, J = 7.9, 1.6 Hz, 1H), 7.95 (d, J = 8.1 Hz, 1H), 7.73−7.60 (m, 2H), 7.44−7.27 (m, 2H), 6.94−6.83 (m, 2H), 3.81 (t, J = 6.9 Hz, 2H), 2.86 (t, J = 4.9 Hz, 4H), 2.71 (t, J = 6.9 Hz, 2H), 2.57 (s, 4H). ^{13}C NMR (101 MHz, methanol-d_4) δ 161.34, 160.99, 159.39, 147.16, 132.88, 131.89, 129.07, 125.90, 125.46, 122.00, 119.55, 117.97, 116.87, 113.61, 57.03, 53.62, 44.74, 37.52. HRMS(ESI) m/z: ([M + H]$^+$) calcd For $C_{20}H_{24}N_5O^+$: 375.1976, found: 375.1981.

*2.5.6. 2-(5-Cyano-2-hydroxyphenyl)-N-(2-(piperazin-1-yl)-ethyl)quinazolin-4-aminium (**4b**).* A yellow solid; yield: 20%; mp: 179.2−181.3 °C; ^{1}H NMR (400 MHz, methanol-d_4) δ 8.29 (s, 1H), 8.08 (dd, J = 8.1, 1.1 Hz, 1H), 7.78−7.72 (m, 2H), 7.53−7.43 (m, 2H), 6.85 (d, J = 8.6 Hz, 1H), 3.90 (t, J = 6.9 Hz, 2H), 2.90 (s, 4H), 2.80 (t, J = 6.9 Hz, 2H), 2.64 (s, 4H). ^{13}C NMR (101 MHz, methanol-d_4) δ: 159.55, 159.43, 144.62, 134.98, 134.07, 133.36, 126.31, 125.47, 122.21, 119.79, 119.44, 118.90, 113.81, 102.78, 100.18, 56.84, 53.47, 44.85, 37.67. HRMS(ESI) m/z: ([M + H]$^+$) calcd For $C_{21}H_{23}N_6O^+$: 375.1928, found: 375.1929.

*2.5.7. 2-(5-Chloro-2-hydroxyphenyl)-N-(2-(piperazin-1-yl)-ethyl)quinazolin-4-aminium (**4c**).* A green solid; yield: 22%; mp: 172.4−174.2 °C; ^{1}H NMR (400 MHz, methanol-d_4) δ 8.40

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

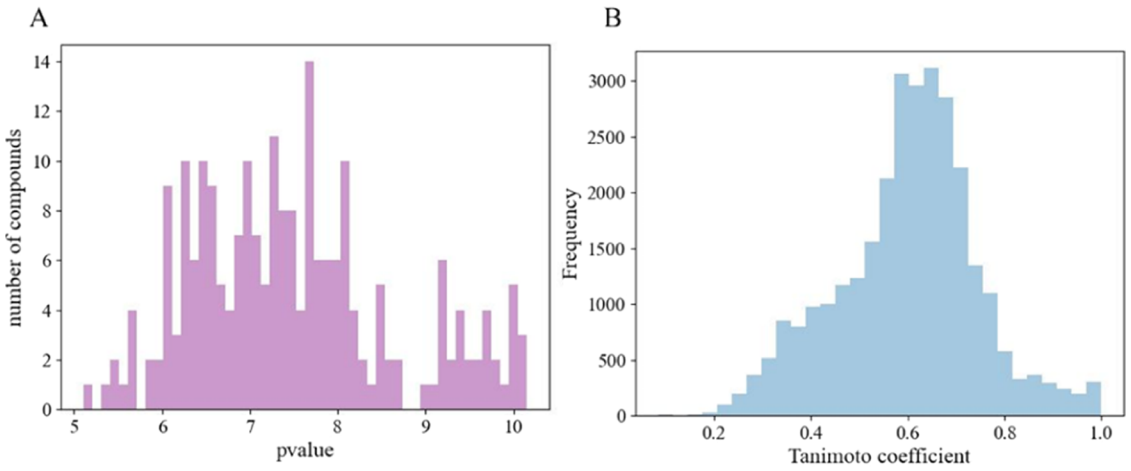


Figure 2. Histograms of the distribution of pIC_{50} values (A) and the Tanimoto coefficient (B) for 221 compounds.

(d, J = 2.7 Hz, 1H), 8.02 (dd, J = 8.2, 1.4 Hz, 1H), 7.79−7.66 (m, 2H), 7.48 (ddd, J = 8.3, 6.9, 1.4 Hz, 1H), 7.27 (dd, J = 8.7, 2.8 Hz, 1H), 6.89 (d, J = 8.8 Hz, 1H), 3.87 (dd, J = 8.3, 5.9 Hz, 2H), 2.91 (t, J = 4.9 Hz, 4H), 2.80−2.63 (m, 6H). ^{13}C NMR (101 MHz, methanol-d_4) δ: 160.19, 159.69, 159.51, 146.64, 133.02, 131.46, 128.17, 126.05, 125.84, 122.63, 121.98, 120.63, 118.55, 113.70, 56.89, 53.67, 44.74, 37.51. HRMS(ESI) m/z: ([M + H]$^+$) calcd. For $C_{20}H_{23}ClN_5O^+$: 384.1586, found: 384.1595.

2.6. Kinase Profiling. All enzymatic reactions were performed at 30 °C for 40 min as the assay method reported by Kashem. 50 μL of reaction mixture contained 40 mM tris, pH 7.4, 10 mM $MgCl_2$, 0.1 mg/mL of bovine serum albumin (BSA), 1 mM of dithiothreitol (DTT), 10 μM of adenosine triphosphate (ATP), 25 ng of kinase, and 0.2 mg/mL of enzyme substrate (Poly (Glu, Tyr)). Compounds were diluted in 10% dimethyl sulfoxide (DMSO), and 5 μL of this dilution was added to 50 μL of the reaction mixture so that the final concentration of DMSO was 1% in all reactions. The assay was performed by using a Kinase-Glo Plus luminescent kinase assay kit. It measures kinase activity by quantifying the amount of ATP remaining in the solution after the kinase reaction. The luminescent signal from the assay correlates with the amount of ATP present and inversely correlates with the increase in kinase activity. The IC_{50} values were calculated using the nonlinear regression of normalized dose responses using graphpad prism 5.0 software.[59,60]

2.7. Virtual Screening. To identify potential tyrosine kinase inhibitors (TKIs) for $EGFR^{L858R/T790M/C797S}$ mutations, we utilized an SVR model to conduct virtual screening experiments on a small-molecule compound data set sourced from the CHEMBL database. Molecular docking was performed by using the CDOCKER module in Discovery Studio. All compounds underwent structure optimization using the MM2 method of ChemDraw 3D and were assigned a Forcefield in the simulation module. The $EGFR^{L858R/T790M/C797S}$ protein (PDB:6LUD)[61] was used as the protein receptor for docking. Before docking, the protein receptor was preprocessed by inserting missing atoms in incomplete residues, modeling the missing loops, and removing cocrystallized water. To generate an aspherical grid with a radius of 10 Å, the binding site was defined around the centroid of the ligand. The resulting protein−drug complexes were then analyzed for their interactions after the docking simulations.

The ADME properties play a critical role in determining the efficacy and safety of drug candidates. Predicting these properties is essential to preventing drug failure during clinical trials. To this end, the Swiss ADME web server, offered by the Swiss Institute of Bioinformatics (SIB), is employed to compute several properties, such as drug-likeness, synthetic accessibility, and PAINS, for the selected compounds.[62] The drug-likeness can be predicted by several rules, such as Lipinski, Ghose, Veber, Egan, and Muegge. Furthermore, the synthetic accessibility of the compounds is evaluated on a scale of 1−10, where a lower score indicates an easier synthetic route and a higher score indicates a more complex route. Compounds that satisfied the criteria for drug-likeness were free of PAINS, demonstrated favorable ADME properties, and were readily synthesized, which are considered the most promising drug candidates.

3. RESULTS AND DISCUSSION

3.1. Diversity of Data Set. The data set of 221 small-molecule inhibitors of $EGFR^{L858R/T790M/C797S}$ triple mutations with various molecular structures was first visualized by the bioactivity (pIC_{50}) distribution. It was generally accepted that a compound was highly active when its IC_{50} < 100 nM, i.e., pIC_{50} > 7, and from Figure 2A, it was found that 58% of the inhibitors were highly active. Then, to measure the similarity of molecular structures, we calculated the Tanimoto coefficients on the basis of MACCS fingerprints. In general, compounds with a Tanimoto coefficient less than 0.7 could be considered to have significant variability. The frequency of distribution histogram (Figure 2B) demonstrated the similarity of each inhibitor. It was obvious that there were 79.5% pairs of inhibitors in the whole data set whose Tanimoto coefficient values were less than 0.70. It can be concluded that the data set employed in this study exhibited a high degree of diversity.

3.2. Performances of Machine Learning Models. *3.2.1. Feature Selection.* In order to select the most pertinent descriptors relating to the bioactivity (pIC_{50}), we initiated the process with 10 molecular structural descriptors extracted from RDKit. Specifically, these descriptors were NOCount, NumHAcceptors, NumAromaticHeterocycles, fr_C_O, fr_Nhpyrrole, fr_amide, fr_aniline, fr_imidazole, fr_piperdine, and fr_piperzine. These descriptors hold substantial influence over EGFR-TKIs and represent fundamental facets of the molecular structure. Both NOCount and NumHAcceptors denote the counts of nitrogen/oxygen and hydrogen bond acceptors, respectively. These counts are pivotal for fostering interactions in the domain structure. The remaining descriptors, which

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

encompass the tally of aromatic heterocycles, ketoaldehydes, pyrroles, amides, anilines, imidazoles, piperidines, and piperazines, are extensively utilized within EGFR-TKIs. Moreover, their structural existence within these compounds has been documented in various reports. A full quadratic effect model was constructed by 10 structural molecule descriptors with an R^2 of 0.750 and a Prob (F-statistics) of 7.33×10^{-26}, which showed the accuracy and effectiveness of the model. Out of 65 variables, 10 are first-order effects, 45 are interaction effects, and 10 are second-order effects. The *p*-values of three first-order effects and two interaction effects are less than 0.05 (Table 1), indicating a significant effect between this variable and the dependent variable pIC_{50}. Among the three first-order effects, aromatic heterocycles and aniline structures are commonly used in the structural design of EGFR-TKIs (Figure 3). Aromatic nitrogen heterocycles, in particular, often act as the parent core structure of inhibitors. For example, Gefitinib is the first-generation EGFR-TKI used for the treatment of nonsmall cell lung cancer caused by primary mutations.[63] Amino-quinazoline can form hydrogen bond interactions with Met793 and Thr790 of tyrosine kinases. To address the continuous occurrence of acquired resistance, the fourth-generation EGFR-TKI Angew-2017 continues the aminoquinoline structure.[64] It can efficiently inhibit the Del19/T790M/C797S mutant kinase (IC_{50} = 17.9 nM). Brigatinib, a fourth-generation EGFR-TKI, has structural similarities to osimertinib. The IC_{50} value for the L858R/T790M/C797S triple mutant kinase is 38.3 nM. In the molecular simulation of brigatinib and tyrosine kinase, the part replaced by piperidine is defined as the part with low affinity and can be modified. This is consistent with the negative correlation results of the piperidine structure in the full quadratic effect model.[65]

Table 1. Results of the Full Quadratic Effect Model

variable	coefficient	standard error	*p*-value
NumAromaticHeterocycles	2.3402	1.030	0.024
fr_piperdine	−3.8413	1.524	0.013
fr_aniline	1.4642	0.552	0.009
NumAromaticHeterocycles* fr_piperdine	−1.2650	0.373	0.001
NumAromaticHeterocycles* fr_aniline	−0.4882	0.186	0.009

Subsequently, we employed the Lasso model to identify significant molecular descriptors from the comprehensive set of descriptors (totaling 208) calculated by using RDKit. The outcomes of this endeavor are presented in Table 2. Ultimately, we amalgamated the 5 variables extracted from the full quadratic effect model with the 7 variables pinpointed by the Lasso model, constituting a holistic collection of pivotal molecule descriptors for the purpose of constructing a support vector regression model.

Table 2. Results of the Lasso Model

descriptor	description
BCUT2D_MWHI	similarity calculations of a property at a certain distance from each atom
BCUT2D_CHGHI	charge distribution calculations for the atoms
Chi2v	the second-order valence bonding index of a molecule
HallKierAlpha	calculation of the similarity of the charge distribution and arrangement of atoms
PEOE_VSA10	atomic partial charges on the van der Waals surface area of a molecule in the range [0.10, 0.15].
EState_VSA7	EState VSA Descriptor 7 ($1.81 \leq x < 2.05$)
VSA_EState2	VSA EState Descriptor 2 ($4.78 \leq x < 5.00$)

3.2.2. SVR Model. For constructing the SVR model with the RBF kernel, we selected three optimal parameters (C, γ, ε) by using the lattice search method with 5-fold cross-validation. As a result (Table 3), the best parameters C = 1.7, γ = 0.11, ε = 0.3 for the SVR model built with 12 important molecule descriptors and an R^2 of 0.745 and an MSE of 0.255 were obtained for test set. The calculated and experimental values of bioactivity (pIC_{50}) on the SVR model are shown in Figure 4.

Table 3. Performance of the SVR Model

	training set		test set	
modeling method	MSE	R^2	MSE	R^2
SVR	0.147	0.853	0.255	0.745

3.3. Synthesis of Inhibitors Based on Structural Descriptors. Following the utilization of a comprehensive quadratic effect model, structural descriptors that offered more robust guidance for inhibitor synthesis were preserved. Grounded in the insight gleaned from these descriptors, we judiciously formulated and synthesized inhibitors aligned with their structural attributes. Subsequently, we validated the inhibitory potential of these inhibitors through a series of *in vitro* experiments.

In the first-order effect, due to the negative correlation between piperidine and the dependent variable pIC_{50}, the presence of piperidine structures should be avoided. At the same time, the coefficient of aniline and aromatic heterocycles in the interaction is negative, indicating that the number of aniline or aromatic heterocycles present will affect the positive correlation of the activity of other moieties. Therefore, in the design, efforts

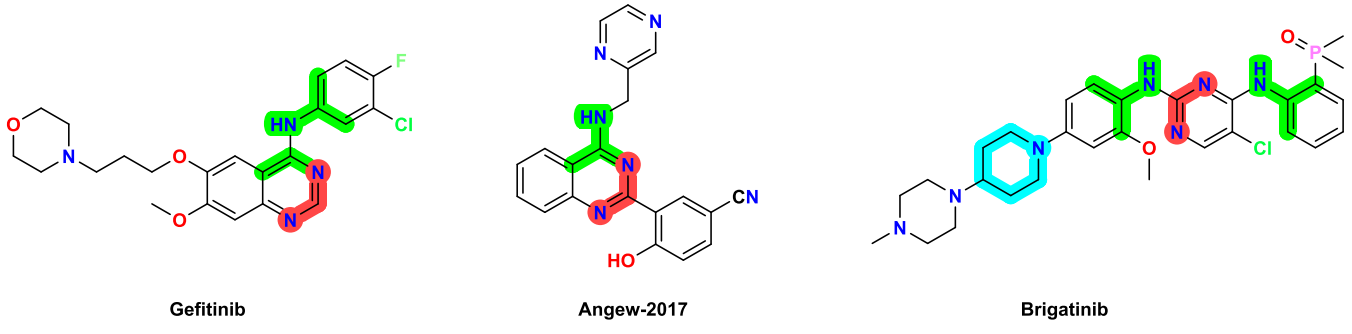


Figure 3. Structural analysis of classical EGFR-TKIs.

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

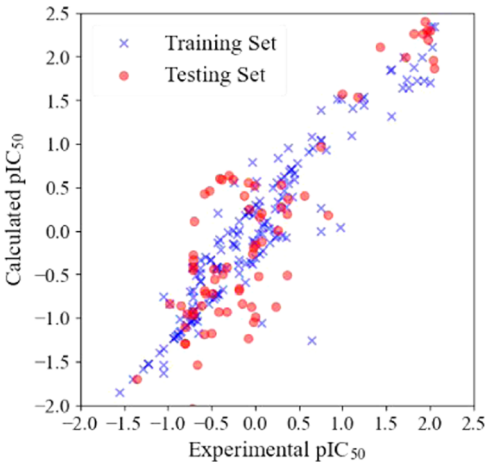


Figure 4. Calculated vs experimental values of bioactivity (pIC_{50}) on the SVR model.

should be made to avoid an overwhelming increase in the number of aniline or aromatic heterocycles. We used 2-aryl-4-aminoquinazole as the starting point. It is not only a classic mother nucleus structure but also contains two relatively balanced structures: aniline and aromatic heterocycles. The regulation of the structure−activity relationship was achieved by replacing different substituents in the aromatic group. To satisfy the interaction between inhibitors and the solvent region outside the tyrosine domain, piperazine structures were connected from the perspective of medicinal chemistry design. A total of 3 inhibitors were synthesized through the above design shown in Figure 5 (**4a−c**).

3.4. *In Vitro* Enzymatic Activity Assay. Angew-2017 containing 2-aryl-4-aminoquinazole structure showed excellent inhibitory activity against Del19/T790M/C797S.[64] Therefore, when *in vitro* enzyme inhibition experiments on inhibitors with 2-aryl-4-aminoquinazole as the mother nucleus were conducted, two triple mutated kinases, L858R/T790M/C797S and Del19/T790M/C797S, were considered. The results are listed in Table 4.

The findings demonstrated that the synthesized compound exhibited concurrent inhibitory activity against both triple mutated kinases L858R/T790M/C797S and Del19/T790M/C797S. Notably, the presence of potent electron-withdrawing substituents within the phenolic hydroxyl group emerged as a critical determinant for the inhibitor's activity. Compound **4c**, which was substituted with chlorine, exhibited the strongest inhibitory effect on both triple mutated kinases ($IC_{50}^{L858R/T790M/C797S}$ = 0.277 μM, $IC_{50}^{Del19/T790M/C797S}$ = 0.089 μM). Compound Angew-2017 demonstrated a stronger inhibitory effect on enzyme activity compared to that of the synthesized compound. This compound contained more aromatic heterocyclic structures in its composition. These observations suggest that the number of aromatic heterocycles could have a positive correlation with activity, which aligns with the findings of the full quadratic effect model. In order to determine the binding effect between the synthesized compound and the active site of the enzyme, molecular docking was performed between **4a−c** and $EGFR^{L858R/T790M/C797S}$ triple mutated kinase (Figures S17−S19). The results showed that **4a−c** can bind to kinase active sites in different forms. The experimental results indicated that the synthesized compounds could be further studied as excellent lead compounds for discovering fourth-generation EGFR-TKIs

3.5. Virtual Screening. In this study, a data set of 50,000 compound molecular structures, sourced from the ChMEBL database, was utilized for virtual screening. A pIC_{50} value threshold of above 7.5 was set to ensure empirical significance. Consequently, 35 compounds, exhibiting pIC_{50} values surpassing 7.5 in the SVR model, were pinpointed as potential inhibitors.

3.5.1. Molecular Docking. Molecular docking is a computational method to identify the bioactive conformations of small molecules within protein binding sites. We employed this method to dock 35 compounds selected through virtual screening with the $EGFR^{L858R/T790M/C797S}$ protein (PDB:6LUD). After the ligands were prepared, a total of 665 conformations were generated. The docking results were sorted by the size of the -CDOCKER ENERGY score, and four hits with a score greater than 30 were obtained.

The visual analysis of the docking results revealed that all four hits were able to interact with the structural domains of the $EGFR^{L858R/T790M/C797S}$ triple mutant proteins in different ways, as shown in Figure 6. Hydrogen-bonding interactions were observed in all compounds, involving amino acids, such as MET793, SER797, LYS745, and PRO794, among others. Aromatic structures also played an important role in the interaction of compounds with tyrosine kinases such as Pi-Alkyl and Alkyl interactions. The potential hit structures discovered above will provide new possibilities for the design of a new generation of EGFR inhibitors.

3.5.2. ADME Analysis. ADME parameters are crucial in determining the potential of a chemical compound to function as an active drug. Solubility, lipophilicity, and bioavailability, among other physicochemical properties, are critical factors in

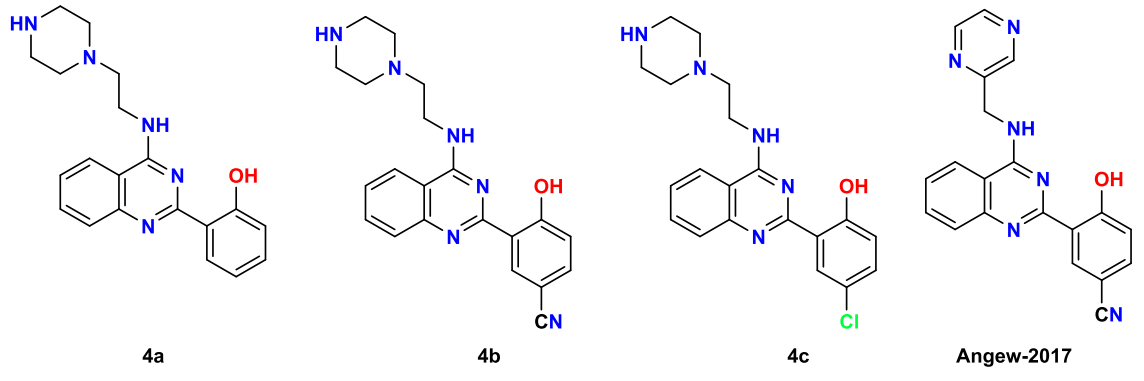


Figure 5. Structure of compounds based on structural descriptors.

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

Table 4. *In Vitro* Enzyme Inhibitory Activity of the Synthesized Compound on the Triple Mutant EGFR and the Number of Structural Descriptors Contained in the Compound[a]

compound	$EGFR^{L858R/T790M/C797S}$ IC_{50} (μM)	$EGFR^{Del19/T790M/C797S}$ IC_{50} (μM)	fr_aniline	NumAromaticHeterocycles
4a	1.221 ± 0.031	1.432 ± 0.022	1	1
4b	0.736 ± 0.044	0.220 ± 0.002	1	1
4c	0.277 ± 0.028	0.089 ± 0.004	1	1
Angew-2017	Nd[b]	0.018	1	2

[a]The data are the mean ± SD of at least three independent experiments. [b]Nd: no data.

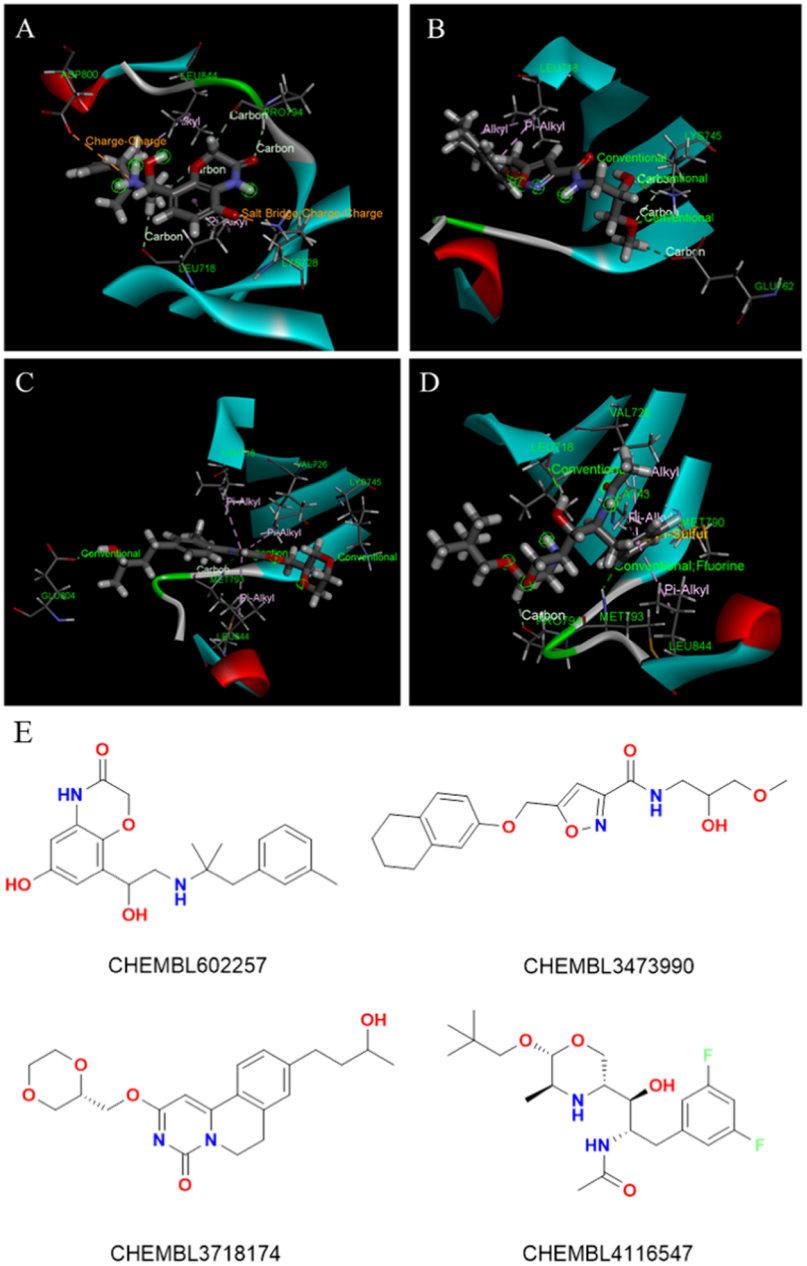


Figure 6. 2D interaction diagram of $EGFR^{L858R/T790MC797S}$ in complex with (A) CHEMBL602257, (B) CHEMBL3473990, (C) CHEMBL3718174, (D) CHEMBL4116547, and (E) the corresponding structures of the compounds.

evaluating the drug-likeness of a compound. All four hits demonstrated excellent physicochemical properties, as highlighted in Table 5. Each hit satisfied specific criteria, including solubility (−6 < Log *S* < 0), lipophilicity (0.7 < Log *P* < 5), molecular weight (150 < MW < 500 g/mol), polarity (20 < TPSA < 130), saturation (0.25 < fraction Csp3 < 1), hydrogen bond acceptors (0 < HBA < 10), and hydrogen bond donors (0 < HBD < 5). Furthermore, drug-likeness was evaluated using important rule-based approaches such as Lipinski's rule of five and Ghose, Veber, Egan, and Muegge (Table 6). All four hits showed a bioavailability score of 0.55, with no PAINS alerts and

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

Table 5. Physicochemical Properties of 4 Hits

molecule	MW	Csp3	HBA	HBD	TPSA	Log *P*	Log *S*
CHEMBL602257	370.44	0.38	5	4	107.74	2.77	−3.44
CHEMBL3473990	360.4	0.47	6	2	94.58	3.55	−3.11
CHEMBL3718174	386.44	0.52	6	1	104.37	3.14	−3.06
CHEMBL4116547	414.49	0.67	7	3	109.3	3.46	−3.59

Table 6. Predicted Drug-Likeness and Synthetic Accessibility of the 4 Hits

molecule	drug-likeness violations[a]	bioavailability score	PAINS alerts	synthetic accessibility
CHEMBL602257	0	0.55	0	3.68
CHEMBL3473990	0	0.55	0	3.79
CHEMBL3718174	0	0.55	0	4.34
CHEMBL4116547	0	0.55	0	4.76

[a]Lipinski's rule of five, Ghose, Veber, Egan, and Muegge.

a plausible synthetic accessibility score ranging from 3.68 to 4.79, as indicated in Table 5.

4. CONCLUSIONS

In this study, we assembled a data set encompassing 221 small-molecule inhibitors targeting $EGFR^{L858R/T790M/C797S}$, a valuable resource for cancer therapy. We employed RDKit to compute 208 molecular descriptors for these compounds. Subsequently, we introduced a fusion framework comprising a full quadratic effect model and a Lasso model. This framework facilitated the selection of both structural and nonstructural molecular descriptors. Utilizing these descriptors, we constructed an SVR model that integrated 12 crucial descriptors, resulting in an R^2 value of 0.745 and an MSE of 0.255 on the test set. Guided by the structural descriptors employed in the modeling process, we synthesized a range of fourth-generation EGFR inhibitors and conducted *in vitro* kinase activity assays. The IC_{50} values of the synthesized compounds reached the nanomolar level. A virtual screening was then conducted on a data set containing around 50,000 compounds, leveraging the SVR model. This screening, combined with molecular simulation docking and ADME analysis, yielded 4 hit compounds. It is expected that activity evaluation on the screened compounds will be conducted to discover a new generation of EGFR-TKI.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c06225

General synthetic route and available copies of ^{1}H NMR, ^{13}C NMR, HRMS, and HPLC chromatograms for compounds **2**, **3a**, **3b**, **3c**, **4a**, **4b**, and **4c**; and molecular docking results of compounds **4a**, **4b**, and **4c** (PDF)

HLPC analysis of all tested compounds (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Dianpeng Wang − *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, P. R. China*; Email: wdp@bit.edu.cn

Renzhong Qiao − *State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China*; orcid.org/0000-0002-6672-9609; Email: qiao_group@163.com

Chao Li − *State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China*; orcid.org/0000-0002-0320-5509; Email: lichao@mail.buct.edu.cn

Authors

Hao Chang − *State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China*

Zeyu Zhang − *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, P. R. China*

Jiaxin Tian − *State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China*

Tian Bai − *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, P. R. China*

Zijie Xiao − *State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c06225

Author Contributions

#Hao Chang and Zeyu Zhang contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (21977012, 21672021, 21572018, and 21372024) and the Joint Project of BRCBC (Biomedical Tralailinslational Engineering Research Center of BUCT-CJFH) (XK2020-06).

■ REFERENCES

(1) Yarden, Y.; Sliwkowski, M.-X. Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell. Biol.* **2001**, *2*, 127−137.

(2) Hynes, N.-E.; Lane, H.-A. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat. Rev. Cancer* **2005**, *5*, 341−354.

(3) Ciardiello, F.; Tortora, G. EGFR antagonists in cancer treatment. *N. Eng. J. Med.* **2008**, *358*, 1160−1174.

(4) Roskoski, R., Jr. Anaplastic lymphoma kinase (ALK): structure, oncogenic activation, and pharmacological inhibition. *Pharmacol. Res.* **2013**, *68*, 68−94.

(5) Westover, D.; Zugazagoitia, J.; Cho, B.; Lovly, C.; Paz-Ares, L. Mechanisms of acquired resistance to first-and second-generation EGFR tyrosine kinase inhibitors. *Ann. Oncol.* **2018**, *29*, i10−i19.

(6) Maemondo, M.; Inoue, A.; Kobayashi, K.; Sugawara, S.; Oizumi, S.; Isobe, H.; Gemma, A.; Harada, M.; Yoshizawa, H.; Kinoshita, I.; et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N. Engl. J. Med.* **2010**, *362*, 2380−2388.
(7) Juan, O.; Popat, S. Treatment choice in epidermal growth factor receptor mutation-positive non-small cell lung carcinoma: latest evidence and clinical implications. *Ther. Adv. Med. Oncol.* **2017**, *9*, 201−216.
(8) Lavacchi, D.; Mazzoni, F.; Giaccone, G. Clinical evaluation of dacomitinib for the treatment of metastatic non-small cell lung cancer (NSCLC): current perspectives. *Drug. Des. Dev. Ther.* **2019**, *13*, 3187−3198.
(9) Le, T.; Gerber, D.-E. Newer-generation EGFR inhibitors in lung cancer: how are they best used? *Cancers* **2019**, *11*, 366.
(10) Kosaka, T.; Yatabe, Y.; Endoh, H.; Yoshida, K.; Hida, T.; Tsuboi, M.; Tada, H.; Kuwano, H.; Mitsudomi, T. Analysis of Epidermal Growth Factor Receptor Gene Mutation in Patients with Non-Small Cell Lung Cancer and Acquired Resistance to Gefitinib. *Clin. Cancer Res.* **2006**, *12*, 5764−5769.
(11) Oxnard, G.-R.; Arcila, M.-E.; Chmielecki, J.; Ladanyi, M.; Miller, V.-A.; Pao, W. New Strategies in Overcoming Acquired Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors in Lung Cancer. *Clin. Cancer Res.* **2011**, *17*, 5530−5537.
(12) Lynch, T. J.; Bell, D. W.; Sordella, R.; et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **2004**, *350*, 2129−2139.
(13) Yang, Z.; Yang, N.; Ou, Q.; Xiang, Y.; Jiang, T.; Wu, X.; Bao, H.; Tong, X.; Wang, X.; Shao, Y. Investigating Novel Resistance Mechanisms to Third-Generation EGFR Tyrosine Kinase Inhibitor Osimertinib in Non-Small Cell Lung Cancer PatientsNovel Resistance Mechanisms to Osimertinib in NSCLC Patients. *Clin. Cancer Res.* **2018**, *24*, 3097−3107.
(14) Oxnard, G.-R.; Hu, Y.; Mileham, K.-F.; Husain, H.; Costa, D.; Tracy, P.; Feeney, N.; Sholl, L.; Dahlberg, S.; Redig, A.; et al. Assessment of resistance mechanisms and clinical implications in patients with EGFR T790M-positive lung cancer and acquired resistance to osimertinib. *JAMA Oncol.* **2018**, *4*, 1527−1534.
(15) Eno, M.-S.; Brubaker, J.-D.; Campbell, J.-E.; De Savi, C.; Guzi, T.-J.; Williams, B.-D.; Wilson, D.; Wilson, K.; Brooijmans, N.; Kim, J.; et al. Discovery of BLU-945, a reversible, potent, and wild-type-sparing next-generation EGFR mutant inhibitor for treatment-resistant non-small-cell lung cancer. *J. Med. Chem.* **2022**, *65*, 9662−9677.
(16) Tavera, L.; Zhang, Z.; Wardwell, S.; Job, E.; McGinn, K.; Chen, M.; Iliou, M.; Albayya, F.; Campbell, J.; Eno, M.; et al. BLU-701 tumour suppression and intracranial activity as a single agent and in combination with BLU-945 in models of non-small cell lung cancer (NSCLC) driven by EGFR mutations. *Mol. Cell. Biol.* **2022**, *165*, S37.
(17) Das, D. EGFRC797S Mutation and Fourth-Generation EGFR Tyrosine Kinase Inhibitors. In *Protein Kinase Inhibitors*; Elsevier, 2022; pp 689−709.
(18) Lim, S.-M.; Ahn, J.-S.; Hong, M.-H.; Kim, T.-M.; Jung, H.-A.; Ou, S.-H.; Jeong, S.; Lee, Y.-H.; Yim, E.; Jung, S.; et al. MA07. 09 BBT-176, a 4th generation EGFR TKI, for Progressed NSCLC after EGFR TKI Therapy: PK, Safety and Efficacy from Phase 1 Study. *J. Thorac. Oncol.* **2022**, *17*, S70−S71.
(19) Liu, L.; Qiu, C.; Liu, X.; Lian, Y.; Chen, H.; Song, X.; Shen, Q.; Du, G.; Guo, J.; Yan, D. BPI-361175, a 4th generation EGFR-TKI for the treatment of non-small cell lung cancer (NSCLC). *Cancer. Res.* **2022**, *82*, 5462.
(20) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.
(21) Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug. Discovery Today* **2015**, *20*, 318−331.
(22) Lo, Y.-C.; Rensi, S.-E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.
(23) Chauhan, J.-S.; Dhanda, S.-K.; Singla, D.; Agarwal, S. M.; Raghava, G. P. QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. *PLoS One* **2014**, *9*, No. e101079.
(24) Kong, Y.; Qu, D.; Chen, X.; Gong, Y.-N.; Yan, A. Self-organizing map (SOM) and support vector machine (SVM) models for the prediction of human epidermal growth factor receptor (EGFR/ErbB-1) inhibitors. *Comb. Chem. High Throughput Screening* **2016**, *19*, 400−411.
(25) 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. Diversity* **2022**, *26*, 1715−1730.
(26) Zhang, H.; Wang, J.; Zhao, H.-Y.; Yang, X.-Y.; Lei, H.; Xin, M.; Cao, Y.-X.; Zhang, S.-Q. Synthesis and biological evaluation of irreversible EGFR tyrosine kinase inhibitors containing pyrido [3, 4-d] pyrimidine scaffold. *Bioorg. Med. Chem.* **2018**, *26*, 3619−3633.
(27) Zhang, H.; Wang, J.; Shen, Y.; Wang, H.; Duan, W.; Zhao, H.; Hei, Y.; Xin, M.; Cao, Y.; Zhang, S. Discovery of 2,4,6-trisubstitued pyrido 3,4-d pyrimidine derivatives as new EGFR-TKIs. *Eur. J. Med. Chem.* **2018**, *148*, 221−237.
(28) Xia, Z.; Huang, R.; Zhou, X.; Chai, Y.; Chen, H.; Ma, L.; Yu, Q.; Li, Y.; Li, W.; He, Y. The synthesis and bioactivity of pyrrolo [2, 3-d] pyrimidine derivatives as tyrosine kinase inhibitors for NSCLC cells with EGFR mutations. *Eur. J. Med. Chem.* **2021**, *224*, No. 113711.
(29) Wittlinger, F.; Heppner, D.; To, C.; Günther, M.; Shin, B.; Rana, J.; Schmoker, A.; Beyett, T.; Berger, L.; Berger, B.; et al. Design of a "two-in-one" mutant-selective epidermal growth factor receptor inhibitor that spans the orthosteric and allosteric sites. *J. Med. Chem.* **2022**, *65*, 1370−1383.
(30) Wang, C.; Wang, X.; Huang, Z.; Wang, T.; Nie, Y.; Yang, S.; Xiang, R.; Fan, Y. Discovery and structural optimization of potent epidermal growth factor receptor (EGFR) inhibitors against L858R/T790M/C797S resistance mutation for lung cancer treatment. *Eur. J. Med. Chem.* **2022**, *237*, No. 114381.
(31) Su, Z.; Yang, T.; Wang, J.; Lai, M.; Tong, L.; Wumaier, G.; Chen, Z.; Li, S.; Li, H.; Xie, H.; Zhao, Z. Design, synthesis and biological evaluation of potent EGFR kinase inhibitors against 19D/T790M/C797S mutation. *Bioorg. Med. Chem.* **2020**, *30*, No. 127327.
(32) Shen, J.; Zhang, T.; Zhu, S. J.; Sun, M.; Tong, L.; Lai, M.; Zhang, R.; Xu, W.; Wu, R.; Ding, J.; Yun, C.-H.; Xie, H.; Lu, X.; Ding, K. Structure-Based Design of 5-Methylpyrimidopyridone Derivatives as New Wild-Type Sparing Inhibitors of the Epidermal Growth Factor Receptor Triple Mutant (EGFR(L858R/T790M/C797S)). *J. Med. Chem.* **2019**, *62*, 7302−7308.
(33) Liu, Y.; Lai, M.; Li, S.; Wang, Y.; Feng, F.; Zhang, T.; Tong, L.; Zhang, M.; Chen, H.; Chen, Y.; Song, P.; Li, Y.; Bai, G.; Ning, Y.; Tang, H.; Fang, Y.; Chen, Y.; Lu, X.; Geng, M.; Ding, K.; Yu, K.; Xie, H.; Ding, J. LS-106, a novel EGFR inhibitor targeting C797S, exhibits antitumor activities both in vitro and in vivo. *Cancer Sci.* **2022**, *113*, 709−720.
(34) Li, S.; Zhang, T.; Zhu, S.; Lei, C.; Lai, M.; Peng, L.; Tong, L.; Pang, Z.; Lu, X.; Ding, J.; Ren, X.; Yun, C.; Xie, H.; Ding, K. Optimization of Brigatinib as New Wild-Type Sparing Inhibitors of EGFR(T790M/C797S) Mutants. *ACS Med. Chem. Lett.* **2022**, *13*, 196−202.
(35) Li, Q.; Zhang, T.; Li, S.; Tong, L.; Li, J.; Su, Z.; Feng, F.; Sun, D.; Tong, Y.; Wang, X.; Zhao, Z.; Zhu, L.; Ding, J.; Li, H.; Xie, H.; Xu, Y. Discovery of Potent and Noncovalent Reversible EGFR Kinase Inhibitors of EGFR(L858R/T790M/C797S). *ACS Med. Chem. Lett.* **2019**, *10*, 869−873.
(36) Lei, H.; Fan, S.; Zhang, H.; Liu, Y.-J.; Hei, Y.-Y.; Zhang, J.-J.; Zheng, A. Q.; Xin, M.; Zhang, S.-Q. Discovery of novel 9-heterocyclyl substituted 9H-purines as L858R/T790M/C797S mutant EGFR tyrosine kinase inhibitors. *Eur. J. Med. Chem.* **2020**, *186*, No. 111888.
(37) Lategahn, J.; Keul, M.; Klövekorn, P.; Kloevekorn, P.; Tumbrink, H.-L.; Niggenaber, J.; Müller, M. P.; Mueller, M.-P.; Hodson, L.; Flaßhoff, M.; Flasshoff, M.; Hardick, J.; Grabe, T.; Engel, J.; Schultz-Fademrecht, C.; Schultz, C.; Baumann, M.; Ketzer, J.; Mühlenberg, T.; Muehlenberg, T.; Hiller, W.; Günther, G.; Guenther, G.; Unger, A.; Müller, H.; Mueller, H.; Heimsoeth, A.; Golz, C.; Blank-

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324

Landeshammer, B.; Blank, B.; Kollipara, L.; Zahedi, R.; Strohmann, C.; Hengstler, J.; van Otterlo, W.; Bauer, S.; Rauh, D. Inhibition of osimertinib-resistant epidermal growth factor receptor EGFR-T790M/C797S. *Chem. Sci.* **2019**, *10*, 10789−10801.
(38) Kim, S. L.; Yang, Y.-S.; Lee, S.; Kim, N.-J. Synthesis and biological evaluation of anilide derivatives as epidermal growth factor receptor L858R/T790M and L858R/T790M/C797S inhibitors. *Bull. Korean Chem. Soc.* **2022**, *43*, 1032−1036.
(39) Karnik, K.-S.; Sarkate, A.-P.; Tiwari, S.-V.; Azad, R.; Wakte, P.-S. Design, synthesis, biological evaluation and in silico studies of EGFR inhibitors based on 4-oxo-chromane scaffold targeting resistance in non-small cell lung cancer (NSCLC). *Med. Chem. Res.* **2022**, *31*, 1500−1516.
(40) Karnik, K.-S.; Sarkate, A.-P.; Tiwari, S.-V.; Azad, R.; Burra, S.; Wakte, P.-S. Computational and Synthetic approach with Biological Evaluation of Substituted Quinoline derivatives as small molecule L858R/T790M/C797S triple mutant EGFR inhibitors targeting resistance in Non-Small Cell Lung Cancer (NSCLC). *Bioorg. Chem.* **2021**, *107*, No. 104612.
(41) Juchum, M.; Guenther, M.; Doering, E.; Sievers-Engler, A.; Laemmerhofer, M.; Laufer, S. Trisubstituted Imidazoles with a Rigidized Hinge Binding Motif Act As Single Digit nM Inhibitors of Clinically Relevant EGFR L858R/T790M and L858R/T790M/C797S Mutants: An Example of Target Hopping. *J. Med. Chem.* **2017**, *60*, 4636−4656.
(42) Hu, X.; Xun, Q.; Zhang, T.; Zhu, S.-J.; Li, Q.; Tong, L.; Lai, M.; Huang, T.; Yun, C.-H.; Xie, H.; Ding, K.; Lu, X. 2-Oxo-3,4-dihydropyrimido 4,5-d pyrimidines as new reversible inhibitors of EGFR C797S (Cys797 to Ser797) mutant. *Chin. Chem. Lett.* **2020**, *31*, 1281−1287.
(43) Heppner, D.-E.; Guenther, M.; Wittlinger, F.; Laufer, S.-A.; Eck, M. J. Structural Basis for EGFR Mutant Inhibition by Trisubstituted Imidazole Inhibitors. *J. Med. Chem.* **2020**, *63*, 4293−4305.
(44) Hei, Y.-Y.; Shen, Y.; Wang, J.; Zhang, H.; Zhao, H.-Y.; Xin, M.; Cao, Y.-X.; Li, Y.; Zhang, S.-Q. Synthesis and evaluation of 2,9-disubstituted 8-phenylthio/phenylsulfinyl-9H-purine as new EGFR inhibitors. *Bioorg. Med. Chem.* **2018**, *26*, 2173−2185.
(45) Guo, Y.; Gao, B.; Gao, P.; Fang, L.; Gou, S. Novel anilinopyrimidine derivatives as potential EGFR(T790M/C797S) Inhibitors: Design, Synthesis, biological activity study. *Bioorg. Med. Chem.* **2022**, *70*, No. 116907.
(46) Günther, M.; Lategahn, J.; Juchum, M.; Doring, E.; Keul, M.; Engel, J.; Tumbrink, H.-L.; Rauh, D.; Laufer, S. Trisubstituted Pyridinylimidazoles as Potent Inhibitors of the Clinically Resistant L858R/T790M/C797S EGFR Mutant: Targeting of Both Hydrophobic Regions and the Phosphate Binding Site. *J. Med. Chem.* **2017**, *60*, 5613−5637.
(47) Günther, M.; Juchum, M.; Kelter, G.; Fiebig, H.; Laufer, S. Lung Cancer: EGFR Inhibitors with Low Nanomolar Activity against a Therapy-Resistant L858R/T790M/C797S Mutant. *Angew. Chem., Int. Ed.* **2016**, *55*, 10890−10894.
(48) Ferlenghi, F.; Scalvini, L.; Vacondio, F.; Castelli, R.; Bozza, N.; Marseglia, G.; Rivara, S.; Lodola, A.; La Monica, S.; Minari, R.; Petronini, P.-G.; Alfieri, R.; Tiseo, M.; Mor, M. A sulfonyl fluoride derivative inhibits EGFR(L858R/T790M/C797S) by covalent modification of the catalytic lysine. *Eur. J. Med. Chem.* **2021**, *225*, No. 113786.
(49) Fang, H.; Wu, Y.; Xiao, Q.; He, D.; Zhou, T.; Liu, W.; Yang, C.-H.; Xie, Y. Design, synthesis and evaluation of the Brigatinib analogues as potent inhibitors against tertiary EGFR mutants (EGFR(del19/T790M/C797S) and EGFR(L858R/T790M/C797S)). *Bioorg. Med. Chem.* **2022**, *72*, No. 128729.
(50) Eno, M.-S.; Brubaker, J.-D.; Campbell, J.-E.; De Savi, C.; Guzi, T.-J.; Williams, B.-D.; Wilson, D.; Wilson, K.; Brooijmans, N.; Kim, J.; et al. Discovery of BLU-945, a Reversible, Potent, and Wild-Type-Sparing Next-Generation EGFR Mutant Inhibitor for Treatment-Resistant Non-Small-Cell Lung Cancer. *J. Med. Chem.* **2022**, *65*, 9662−9677.
(51) Engelhardt, H.; Böse, D.; Petronczki, M.; Scharn, D.; Bader, G.; Baum, A.; Bergner, A.; Chong, E.; Döbel, S.; Egger, G.; et al. Start selective and rigidify: the discovery path toward a next generation of EGFR tyrosine kinase inhibitors. *J. Med. Chem.* **2019**, *62*, 10272−10293.
(52) Ding, S.; Gao, Z.; Hu, Z.; Qi, R.; Zheng, X.; Dong, X.; Zhang, M.; Shen, J.; Long, T.; Zhu, Y.; Tian, L.; Song, W.; Liu, R.; Li, Y.; Sun, J.; Duan, W.; Liu, J.; Chen, Y. Design, synthesis and biological evaluation of novel osimertinib derivatives as reversible EGFR kinase inhibitors. *Eur. J. Med. Chem.* **2022**, *238*, No. 114492.
(53) De Clercq, D.-J.-H.; Heppner, D.-E.; To, C.; Jang, J.; Park, E.; Yun, C.-H.; Mushajiang, M.; Shin, B.-H.; Gero, T.-W.; Scott, D.-A.; Janne, P.-A.; Eck, M.-J.; Gray, N.-S. Discovery and Optimization of Dibenzodiazepinones as Allosteric Mutant-Selective EGFR Inhibitors. *ACS Med. Chem. Lett.* **2019**, *10*, 1549−1553.
(54) Chen, Z.; Huang, K.-Y.; Ling, Y.; Goto, M.; Duan, H.-Q.; Tong, X.-H.; Liu, Y.-L.; Cheng, Y.-Y.; Morris-Natschke, S.-L.; Yang, P.-C.; Yang, S.-L.; Lee, K.-H. Discovery of an Oleanolic Acid/Hederagenin-Nitric Oxide Donor Hybrid as an EGFR Tyrosine Kinase Inhibitor for Non-Small-Cell Lung Cancer. *J. Nat. Prod.* **2019**, *82*, 3065−3073.
(55) 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: miniperspective. *J. Med. Chem.* **2018**, *61*, 4290−4300.
(56) Landrum, G. RDKit: Open-Source Cheminformatics, v. 2019, GitHub, 2019. https://githubcom/rdkit/rdkit.
(57) Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process.* **1996**, 9.
(58) Smola, A. J.; Schlkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199−222.
(59) Lei, H.; Fan, S.; Zhang, H.; Liu, Y.-J.; Zhang, S.-Q. Discovery of novel 9-heterocyclyl substituted 9H-purines as L858R/T790M/C797S mutant EGFR tyrosine kinase inhibitors. *Eur. J. Med. Chem.* **2019**, *186*, No. 111888.
(60) Kashem, M. A.; Nelson, R.; Yingling, J.-D.; Pullen, S.-S.; Prokopowicz, A.-S.; Jones, J.-W.; Wolak, J.-P.; Rogers, G.-R.; Morelock, M.-M.; Snow, R.-J.; et al. Three mechanistically distinct kinase assays compared: Measurement of intrinsic ATPase activity identified the most comprehensive set of ITK inhibitors. *SLAS Discovery* **2007**, *12*, 70−83.
(61) Kashima, K.; Kawauchi, H.; Tanimura, H.; Tachibana, Y.; Chiba, T.; Torizawa, T.; Sakamoto, H. CH7233163 Overcomes Osimertinib-Resistant EGFR-Del19/T790M/C797S Mutation. *Mol. Cancer Ther.* **2020**, *19*, 2288−2297.
(62) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, No. 42717.
(63) Mitsudomi, T.; Morita, S.; Yatabe, Y.; Negoro, S.; Okamoto, I.; Tsurutani, J.; Seto, T.; Satouchi, M.; Tada, H.; Hirashima, T.; Asami, K.; Katakami, N.; Takada, M.; Yoshioka, H.; Shibata, K.; Kudoh, S.; Shimizu, E.; Saito, H.; Toyooka, S.; Nakagawa, K.; Fukuoka, M. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *Lancet Oncol.* **2010**, *11*, 121−128.
(64) Park, H.; Jung, H. Y.; Mah, S.; Hong, S. Discovery of EGFR(d746-750/T790M/C797S) Mutant-Selective Inhibitors via Structure-Based de Novo Design. *Angew. Chem., Int. Ed.* **2017**, *56*, 7634−7638.
(65) Uchibori, K.; Inase, N.; Araki, M.; Kamada, M.; Sato, S.; Okuno, Y.; Fujita, N.; Katayama, R. Brigatinib combined with anti-EGFR antibody overcomes osimertinib resistance in EGFR-mutated non-small-cell lung cancer. *Nat. Commun.* **2017**, *8*, No. 14768.

https://doi.org/10.1021/acsomega.3c06225
ACS Omega 2024, 9, 2314−2324