

RESEARCH

Open Access



# Discovery of novel VEGFR2 inhibitors against non-small cell lung cancer based on fingerprint-enhanced graph attention convolutional network

Zixiao Wang<sup>1\*†</sup> , Lili Sun<sup>2†</sup>, Yu Xu<sup>3</sup>, Jing Huang<sup>1</sup>, Fang Yang<sup>1</sup> and Yu Chang<sup>1\*</sup>

## Abstract

Despite the proven inhibitory effects of drugs targeting vascular endothelial growth factor receptor 2 (VEGFR2) on solid tumors, including non-small cell lung cancer (NSCLC), the development of anti-NSCLC drugs solely targeting VEGFR2 still faces risks such as off-target effects and limited efficacy. This study aims to develop a novel fingerprint-enhanced graph attention convolutional network (FnGATGCN) model for predicting the activity of anti-NSCLC drugs. Employing a multimodal fusion strategy, the model integrates a feature extraction layer that comprises molecular graph feature extraction and molecular fingerprint feature extraction. The performance evaluation results indicate that the model exhibits high accuracy and stability in predicting activity. Moreover, we explored the relationship between molecular features and biological activity through visualization analysis, thus improving the interpretability of the approach. Utilizing this model, we screened the ZINC database and conducted high-precision molecular docking, leading to the identification of 11 potential active molecules. Subsequently, molecular dynamics simulations and free energy calculations were performed. The results demonstrate that all 11 aforementioned molecules can stably bind to VEGFR2 under dynamic conditions. Among the short-listed compounds, the top six exhibited satisfactory inhibitory activity against VEGFR2 and A549 cells. Especially, compound Z-3 displayed VEGFR2 inhibitory with  $IC_{50}$  values of 0.88  $\mu$ M, and anti-proliferative activity against A549 cells with  $IC_{50}$  values of  $4.23 \pm 0.45$   $\mu$ M. This approach combines the advantages of target-based and phenotype-based screening, facilitating the rapid and efficient identification of candidate compounds with dual activity against VEGFR2 and A549 cell lines. It provides new insights and methods for the development of anti-NSCLC drugs. Furthermore, further biological activity tests revealed that Z1-Z3 and Z6 manifested relatively strong antiproliferative activities against NCI-H23 and NCI-H460, and relatively low toxicity towards GES-1. The hit compounds were promising candidates for the further development of novel VEGFR2 inhibitors against NSCLC.

**Keywords** Vascular endothelial growth factor receptor 2, Non-small cell lung cancer, Multimodal deep learning, Molecular dynamics simulation, Virtual screening

<sup>†</sup>Zixiao Wang and Lili Sun have contributed equally to this work.

\*Correspondence:

Zixiao Wang  
zixiaowang1112@foxmail.com  
Yu Chang  
changyu0552@163.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## Introduction

Lung cancer is the primary cause of cancer-related death worldwide, with an estimated 2.2 million new cases and nearly 1.8 million deaths each year [1]. Approximately 85% of diagnosed cases belong to the non-small cell lung cancer (NSCLC) subtype [2, 3]. Despite the effective application of various targeted therapies and immunotherapies in some populations of patients with advanced NSCLC, these patients almost inevitably develop metastases, chemotherapy resistance and subsequent tumor recurrence [4, 5]. The 5-year overall survival rate for patients afflicted with metastatic NSCLC registers at less than 5% [6–8], underscoring the imperative for the development of novel treatment strategies. Notably, the vascular endothelial growth factor receptor 2 (VEGFR2) plays a critical role in angiogenesis, which is essential for tumor growth and metastasis. NSCLC is a highly vascularized tumor and suppressing angiogenesis has emerged as a promising treatment strategy [9]. Inhibition of VEGFR2 has been shown to significantly impair the proliferation of various cancer cell lines, including A549 lung carcinoma cells [10]. Studies have demonstrated that blocking VEGFR2 signaling pathways leads to reduced endothelial cell proliferation, decreased vascular permeability, and inhibited tumor growth [11, 12]. Agents such as Ramucirumab, Nintedanib, and Anlotinib, targeting VEGFR2, have demonstrated efficacy in inhibiting a spectrum of solid tumors, including NSCLC [13–15]. However, despite the potential effectiveness of VEGFR2 inhibitors in the clinical management of NSCLC, current agents encounter challenges such as limited translatability, moderate efficacy, and the emergence of drug resistance. Moreover, the exclusive focus on VEGFR2-targeted therapies for NSCLC may yield suboptimal outcomes due to off-target effects and resistance mechanisms [16, 17].

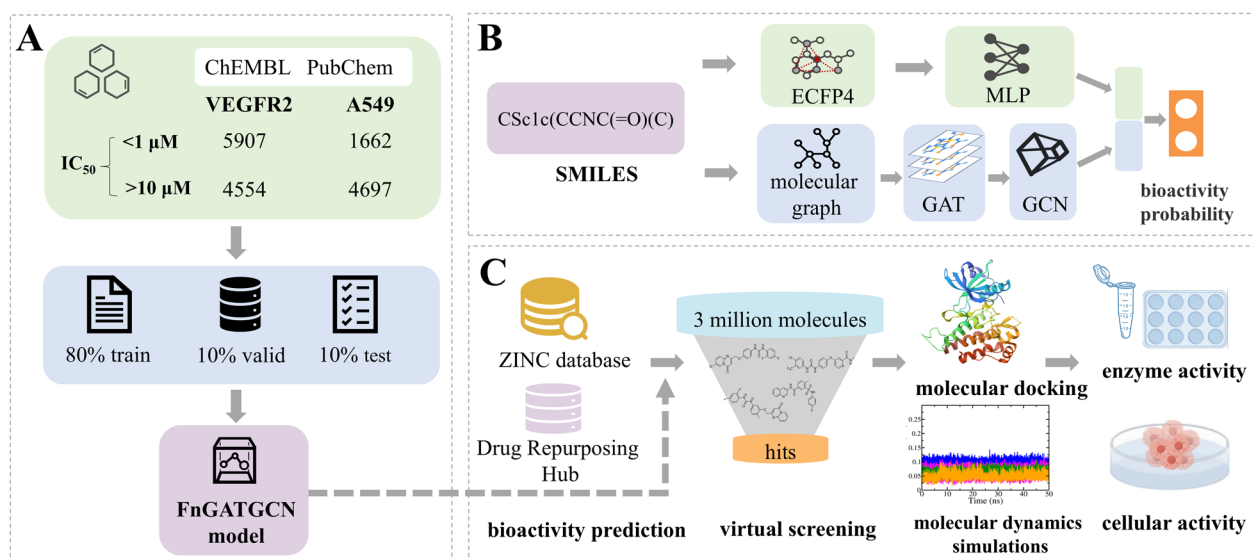
Phenotype-based drug discovery holds promise in addressing this deficiency by enabling a more comprehensive evaluation of drugs [18, 19]. Integrating both phenotype-based screening and VEGFR2 targeting methodologies can enhance the efficiency and success rate of anti-NSCLC drug discovery. However, current methodologies encounter challenges stemming from the heterogeneity of data sources and the intricacies involved in feature extraction [20–22]. Further research is warranted to explore the potential of the phenotype-target combination drug screening model.

The emergence of Artificial Intelligence, particularly Deep Learning, has revolutionized large-scale data mining. Developing reliable deep learning models requires precise algorithms and efficient techniques for extracting molecular features [23]. Molecular structures often involve intricate many-body interactions and complex electronic structures. In contrast to traditional molecular

descriptors, molecular graphs simplify this complexity by representing atoms as nodes and bonds as edges. Recently, graph convolutional network (GCN) have shown promising performance in various aspects of drug design [24], including drug screening [25], prediction of drug-target affinity [26], changes in protein–protein binding affinity [27], and drug toxicity prediction [28]. While GCN demonstrates remarkable molecular representation capability and multi-scale feature integration, its method of directly summing neighbor node features limits model complexity and expressive power [29, 30].

To overcome this limitation, the graph attention network (GAT) was introduced, which adaptively weights neighboring node features and supports multi-head attention mechanisms [31]. The integration of GAT with GCN further amplifies the model's representational capacity and flexibility [20, 32]. For instance, Yu et al. predicted drug-disease associations through layer attention graph convolutional network [33], Sun et al. utilized a graph convolutional attention network to forecast drug similarity [34], and Wang et al. established residue-based graph attention and convolutional network RGN for predicting protein–protein interaction sites [35]. However, the issue of activity cliffs (molecules with highly similar structures but significantly different activities) leads to a decrease in the accuracy of this model [36]. Additionally, the GAT-GCN model still faces limitations in representing chemical molecular structures, including inadequate extraction of local features and loss of chemical information.

Multimodal deep learning is an exciting subfield in the field of artificial intelligence, focusing on developing advanced models that can simultaneously process and learn multiple types of data. In this study, we introduce FnGATGCN, an advanced framework that utilizes a multimodal feature fusion strategy, integrating fingerprint data with graph data. The methodology overview is depicted in Fig. 1. Firstly, we compiled a dataset comprising molecules exhibiting both active and inactive attributes against VEGFR2 and the A549 cell line. This dataset was randomly divided into training, validation, and testing subsets, as illustrated in Fig. 1A. Then, leveraging molecular fingerprints and graphs, we encoded the chemical structures of the compounds. Our FnGATGCN model was then trained and assessed, as demonstrated in Fig. 1B. Subsequently, we employed this model to rapidly screen candidate compounds with dual activity against VEGFR2 and the A549 cell line from the ZINC database (<https://zinc.docking.org/>). Following this, we conducted sequential screening and validation through molecular docking, drug-likeness analysis, and molecular dynamics simulation. Finally, we evaluated the biological activity of the hit compounds, as shown in Fig. 1C. This approach



**Fig. 1** Overview of FnGATGCN model-based screening of potentially active molecules anti-NSCLC. **A** Data Collection and Splitting. **B** The overall architecture of FnGATGCN model. **C** The virtual screening pipeline based on FnGATGCN model.

aims to leverage the strengths of different data modalities to enhance model predictive performance and provide novel insights for NSCLC therapy.

## Materials and methods

### Problem definition and FnGATGCN Architecture

We consider a problem of drug activity prediction, focusing on a binary classification task with cross-entropy as the loss function. Our training set  $D = (x_i, y_i)$  comprises  $n$  drug molecules, where  $x_i$  denotes the combined fingerprint and graph representation of each molecule, and  $y_i$  denotes its true activity label. In our classification task, our goal is to classify each drug molecule as active or inactive, linking  $f(x)$  to a binary classification output that represents the probability of being active using the *sigmoid* function. We aim to minimize the cross-entropy loss and strive for an area under the curve (AUC) value as close to 1 as possible, indicating outstanding model performance in terms of classification accuracy and class separation.

### Data preprocessing

#### Data description

In this study, compounds exhibiting inhibitory activity against both VEGFR2 and A549 cell lines were sourced from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>) for model training and evaluation. For VEGFR2 inhibitors, following data cleaning procedures including deduplication and removal of missing entries, a total of 5907 compounds with  $IC_{50} < 1 \mu M$  were designated as the active dataset, while 1554 compounds with  $IC_{50} > 10 \mu M$

were identified as the inactive dataset. Given that the number of the collected inactive compounds is lower than that of the active compounds, this scenario does not comply with the natural distribution of active and inactive compounds in the unknown database. Hence, it is highly meaningful to broaden the applicability domain of the model and enrich the inactive compound dataset. To achieve this, approximately 15 million compounds were retrieved from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>). Subsequently, SDF files were processed in Python using RDKit to compute extended connectivity fingerprints 4 (ECFP4, radius=2). Employing the k-means algorithm, compounds were clustered, and 3000 central compounds were selected to represent the negative supplementary dataset, resulting in a total of 4554 compounds within the inactive dataset. For the A549 cell line, to mitigate the influence of varied experimental conditions, compound data were selected based on assays utilizing MTT or CCK8 colorimetric agents, with cells treated for 72 h. This selection process yielded 1662 active compounds and 4697 inactive compounds. During each model training iteration, the dataset was randomly partitioned into training, validation, and test subsets in an 8:1:1 ratio.

### Standardization of SMILES sequences

The simplified molecular input line entry specification (SMILES) of a molecule is simply an ASCII string encoding the molecular structure information. SMILES strings and molecular graphs can be converted bidirectionally using the RDKit package. We standardized the above

SMILES sequences using MolVS (<https://molvs.readthedocs.io/en/latest/index.html>), which involved normalizing structures, desalting, neutralizing charges, and removing duplicate molecules. The inhibitory activity of compounds is measured in binary values of “1” and “0”, where molecules with a value of “1” are considered to have inhibitory effects on VEGFR2 and A549.

### Fingerprint-enhanced graph attention convolutional network

#### Molecular featurization

Prior to graph encoding, defining node features is essential. This study utilizes nine types of atomic features and four types of bond features to characterize atoms and their local environments. The node features have a size of 44, while the edge features have a size of 14. Most of these features are encoded in one-hot form, while formal charge and the number of free radical electrons are encoded as integers due to their additive nature [37]. To create one-hot encoded features, all possible categorical variables for the feature are listed, then matched with these variables and labeled as 1 or 0 (one-hot or null) for encoding. The size of each one-hot encoding equals the desired range of values plus one to accommodate unusual values and increase data sparsity. These details are comprehensively presented in Table 1.

#### GAT-GCN encoding module

In GAT, the hidden layer state of each node is computed based on its features and those of its neighbors. By introducing an attention mechanism, the model discerns the varying importance of neighbors to the target node,

translating this into weight parameters for aggregating neighbor information [38, 39]. The input to a single-head graph attention layer consists of feature vectors of the target node, with the output comprising updated node features. This process entails three main steps: alignment, weighting, and contextualization (Fig. 2B).

Alignment

$$e_{vu} = \text{leaky\_relu}(W \cdot [h_v, h_u]) \quad (1)$$

Weighting

$$a_{vu} = \text{softmax}(e_{vu}) = \frac{\exp(e_{vu})}{\sum_{u \in N(v)} \exp(e_{vu})} \quad (2)$$

Context

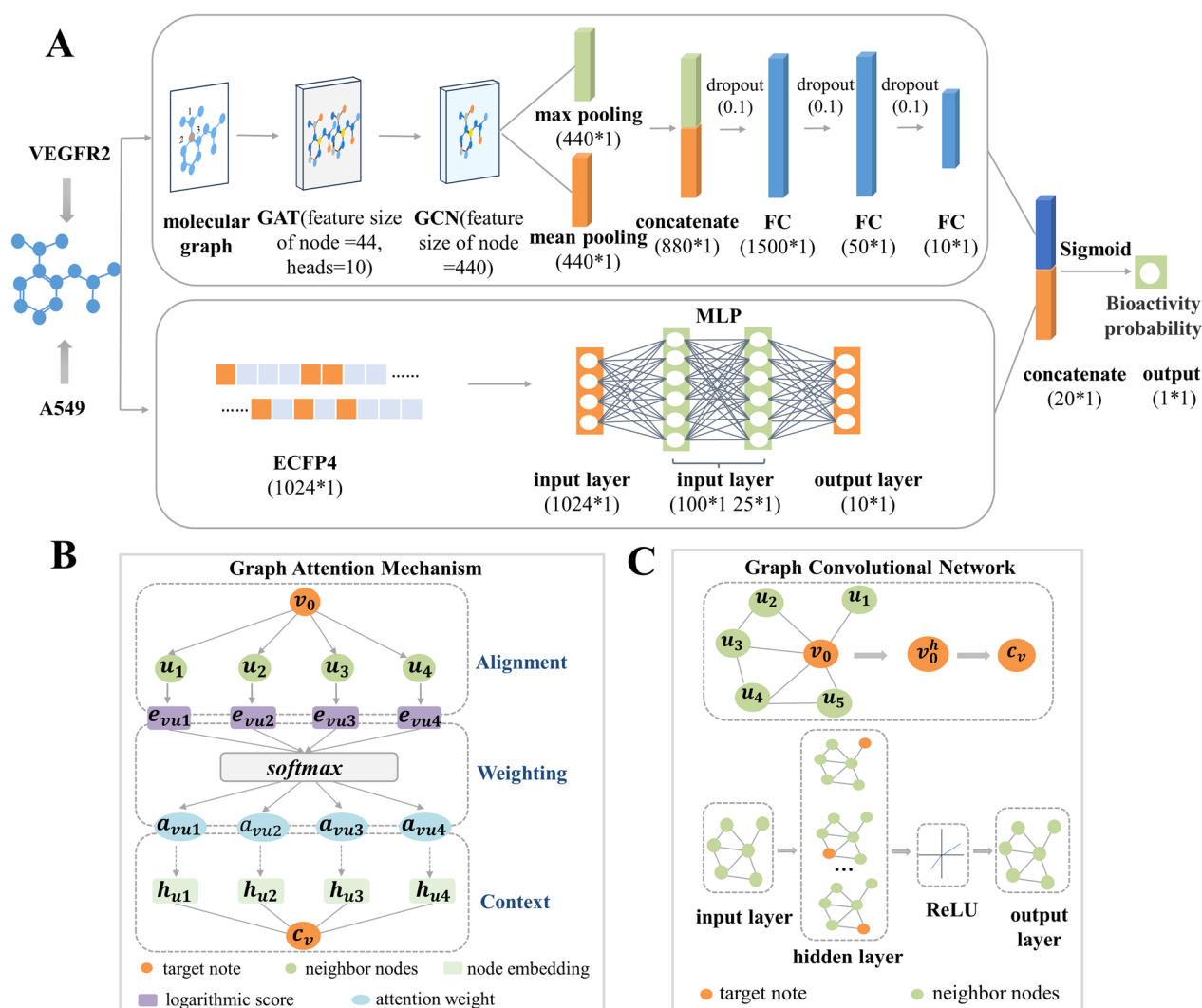
$$C_v = \sigma \left[ \sum_{u \in N(v)} a_{vu} \cdot W \cdot h_u \right] \quad (3)$$

where  $v$  represents the target node,  $h_v$  denotes the state vector of node  $v$ , and  $h_u$  represents the state vector of the neighbor nodes  $u$ . (1) By utilizing a trainable weight matrix  $W$  for linear transformation, the alignment operation produces the output  $e_{vu}$ ; (2) The output  $e_{vu}$  is then normalized using the *softmax* function to obtain the weight  $a_{vu}$  of the neighbor nodes  $u$  with respect to the target node  $v$ . Subsequently; (3) combining the state vector  $h_u$  of the neighbor nodes, the weight  $a_{vu}$ , and a non-linear activation function, the new feature  $C_v$  of the target node  $v$  is obtained.

To enhance the expressiveness and stability of the model, we utilized Multi-head attention. Specifically,

**Table 1** Initial atomic and bond features

Atom feature	Size	Description
Atom symbol	16	[B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, I, At, other] (one-hot)
Degree	6	number of covalent bonds [0,1,2,3,4,5] (one-hot)
Formal charge	1	electrical charge (integer)
Radical electrons	1	number of radical electrons (integer)
Hybridization	6	[sp, sp <sup>2</sup> , sp <sup>3</sup> , sp <sup>3</sup> d, sp <sup>3</sup> d <sup>2</sup> , other] (one-hot)
Aromaticity	1	whether the atom is part of an aromatic system [0/1] (integer)
Hydrogens	5	number of connected hydrogens [0,1,2,3,4] (one-hot)
Chirality	1	whether the atom is a chiral center [0/1] (integer)
Chirality type	2	[R, S] (one-hot)
Bond feature	Size	Description
Bond type	4	[single, double, triple, aromatic] (one-hot)
Conjugation	1	whether the bond is conjugated [0/1] (integer)
Ring	1	whether the bond is in ring [0/1] (integer)
Stereo	6	[StereoNone, StereoAny, StereoZ, StereoE, StereoIs, StereoTrans] (one-hot)



**Fig. 2** The architecture of FnGATGCN model. **A** Overview of the FnGATGCN network architecture. **B** Molecular feature representation based on attention mechanisms. **C** Molecular feature representation based on graph convolutional networks

we conducted  $K$  independent attention calculations for each target node, resulting in  $K$  vector representations of the target node  $v$ . Then concatenated these  $K$  vectors to obtain the final output vector.

$$h'_v = \parallel_{k=1}^K \sigma \left( \sum_{u \in N(v)} \alpha_{vu}^k \cdot W^k \cdot h_u \right) \quad (4)$$

where  $\alpha_{vu}^k$  is the normalization coefficient of the  $K$ -th attention mechanism,  $W^k$  corresponds to the weight matrix of the  $K$ -th linear transformation.

In GCN, each node aggregates its features with those of neighboring nodes to update the feature representation of the target node [40]. Unlike GAT, these weights are typically determined by the degrees of neighboring nodes

(Fig. 2C). Subsequently, after feature updating, the ReLU activation function is applied to introduce non-linearity, thereby enhancing the model's expressive power. Additionally, pooling layers are utilized to downsample the data, reducing the data volume while retaining key information. Specifically, both average pooling and max pooling strategies are applied to the node features ( $440 * n$ ) processed by GAT-GCN, resulting in vectors of size  $440 * 1$  each. These vectors are then concatenated to obtain a feature vector of size  $880 * 1$ .

#### Molecular fingerprint encoding module

In this section, we initially use the GetMorganFingerprintAsBitVect function from the RDKit package to convert the SMILES of compounds to a 1024-bit ECFP4 with

a radius of 2. Next, the encoded molecular fingerprint features of ECFP4 are fed into the input layer of the multilayer perceptron (MLP) consisting of 1024 neurons. The data is then subjected to processing and feature learning via the ReLU non-linear activation function across two hidden layers, comprising 100 and 25 neurons, respectively. Finally, the outcomes are computed using the 10 neurons in the output layer.

#### Feature fusion module and activity prediction

In this part, we integrate the previously extracted molecular graph features and molecular fingerprint features using a fully connected layer to create an augmented feature vector. Subsequently, the study utilizes an output layer with a sigmoid function for classification. The sigmoid activation function, characterized by its continuous and smooth 'S' shaped curve, is widely used and defined by a straightforward mathematical expression, as illustrated in the formula.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

It maps real number inputs to the (0,1) interval, normalizes the output of each neuron, and is suitable for probability prediction in classification problems.

#### FnGATGCN model training and evaluation

##### Experimental setting

In this study, we utilize the Adam optimizer within the PyTorch 1.13.1 framework to train the FnGATGCN model on the A549 dataset and the VEGFR2 dataset respectively. Model parameters were fine-tuned using the gradient descent optimization algorithm. The binary cross-entropy (BCE) loss function was minimized during training to reduce the disparity between predicted outcomes and actual labels, thereby enhancing the model's performance and accuracy. To mitigate overfitting during model training, we introduced L2 weight decay, adjusted the learning rate, and applied dropout. Furthermore, to prevent overfitting and expedite training, an early-stopping strategy was implemented. The maximum number of training epochs was capped at 800. If a model's AUC does not improve for 18 consecutive training rounds, and the loss of the validation set does not decrease for 28 consecutive rounds, training will be stopped. Then, the model with the lowest validation set loss among all models at the time of stopping is selected as the final model. Throughout the experiments, we utilized an NVIDIA GeForce RTX 3080 graphics card with 10 GB of memory and 8704 CUDA cores.

#### Model evaluation

We assessed our FnGATGCN model alongside several other advanced graph neural networks, including GAT-GCN, GAT, GIN, ECC, and GraphSAGE. To mitigate data imbalance, we repeated the data sampling process for each model 10 times and computed the average results. Performance metrics, including AUC, balanced accuracy (BA), F1-measure (F1) score, and Matthews correlation coefficient (MCC), were calculated for both validation and test sets using a Python script. The calculation formula is detailed in Table 2.

Here, TP, FP, TN, and FN represent the counts of true positive, false positive, true negative, and false negative labels, respectively. TPR stands for True Positive Rate, calculated as  $TP/(TP + FN)$ . FPR stands for False Positive Rate, calculated as  $FP/(FP + TN)$ . TNR stands for True Negative Rate, calculated as  $TN/(TN + FP)$ . PPV stands for Positive Predictive Value, calculated as  $TP/(TP + FP)$ .

#### FnGATGCN model and molecular docking co-screening

Firstly, we retrieved 3 million small molecules from the ZINC database and converted them into SMILES format. Each small molecule was then subjected to screening using the FnGATGCN model. Molecules with predicted probabilities exceeding 50% for activity against both VEGFR2 and the A549 cell line were selected for molecular docking analysis.

In the molecular docking section, the ligand's X-ray diffraction structure was obtained from the RCSB PDB (<https://www.rcsb.org/>) and imported into Discovery Studio 2019. Subsequently, it was processed using the "Prepared Protein" module, which involved removing all water molecules, inserting missing loops, and adding hydrogen atoms. The active site was defined as a 12.0 Å radius around the endogenous ligand [41]. The hits underwent energy minimization using the CHARMM force field. Following that, the ligands and prepared protein were utilized for CDOCKER docking with default parameters. The interaction energy values were analyzed to assess the strength of binding between the ligands and proteins, with higher values indicating stronger interactions.

**Table 2** Calculation formulas for evaluation metrics

Evaluation metric	Equation
AUC	$\frac{TPR}{FPR}$
BA	$\frac{TPR + TNR}{2}$
F1-score	$2 * \frac{PPV * TPR}{PPV + TPR}$
MCC	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

### Molecular dynamics simulation

Molecular dynamics simulations were conducted using the GROMACS software employing the AMBER ff99SB-ILDN force field. Atomic charges for the small molecules were determined using Multiwfn 3.8 (dev) and ORCA 5.0.2 software, and then converted into Amber (GAFF) force field files using Sobtop 1.0 (dev3.1) [42–44]. The entire complex system was solvated into a cubic box with TIP3P water model, and appropriate  $\text{Na}^+$  and  $\text{Cl}^-$  ions were added to neutralize the system's charge. A minimum distance of 1.2 nm was maintained between each amino acid and the edges of the cubic box. The simulation system underwent energy minimization with 1000 steps of steepest descent and 5000 steps of conjugate gradient optimization. Pre-equilibration was performed with 100 ps of NVT (constant Number of particles, Volume, and Temperature) and 100 ps of NPT (constant Number of particles, Pressure, and Temperature) simulation at 300 K. Subsequently, each system underwent a 50 ns simulation with a time step of 2 fs, and trajectories were recorded at 10 ps intervals. Root mean square deviation (RMSD) and root mean square fluctuations (RMSF) were analyzed to assess molecular structural deviations and atomic flexibility throughout the simulation [45]. Binding free energy was calculated using the Molecular Mechanics/Poisson-Boltzmann surface area (MM/PBSA) method, employing the specific formula outlined below:

$$\Delta G_{\text{bind}} = \Delta E_{\text{MM}} + \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}} - T\Delta S \quad (6)$$

$\Delta E_{\text{MM}}$  represents the field of force of molecule, including van der Waals ( $\Delta E_{\text{vdw}}$ ) and electrostatic forces ( $\Delta E_{\text{ele}}$ ),  $\Delta G_{\text{polar}}$  denotes the polar solvation-free energy,  $\Delta G_{\text{nonpolar}}$  indicates the nonpolar solvation-free energy, and  $-T\Delta S$  represents the entropy change.

### Biological evaluation

#### *In vitro* VEGFR2 inhibition assay

The inhibitory activity of the hit compounds on the VEGFR2 enzyme *in vitro* was assessed using the Caliper Mobility-Shift method. The compound (3.0 mg, Chemdiv, USA) was accurately weighed, dissolved in DMSO, and diluted to various concentrations using  $1\times$  kinase buffer. A 250 nL aliquot of the solution was transferred to the assay plate via Echo550. A  $2.5\times$  protein solution was prepared with  $1\times$  kinase buffer, and 10  $\mu\text{L}$  was added to the compound, positive control, and negative control wells, followed by 10  $\mu\text{L}$  of  $1\times$  kinase buffer. The plate was centrifuged at 1000 rpm for 30 s and incubated at room temperature for 10 min. A mixture of  $5\times$  ATP and kinase substrate was then added (15  $\mu\text{L}$  per

well), centrifuged at 1000 rpm for 3 s, and incubated at room temperature for 30 min. Finally, 30  $\mu\text{L}$  of detection solution was added, the plate was centrifuged at 1000 rpm for 30 s, and the conversion rate was measured using a microplate reader.  $\text{IC}_{50}$  values were calculated via curve fitting in GraphPad Prism 9.0.

#### *In vitro* cell proliferation inhibition test

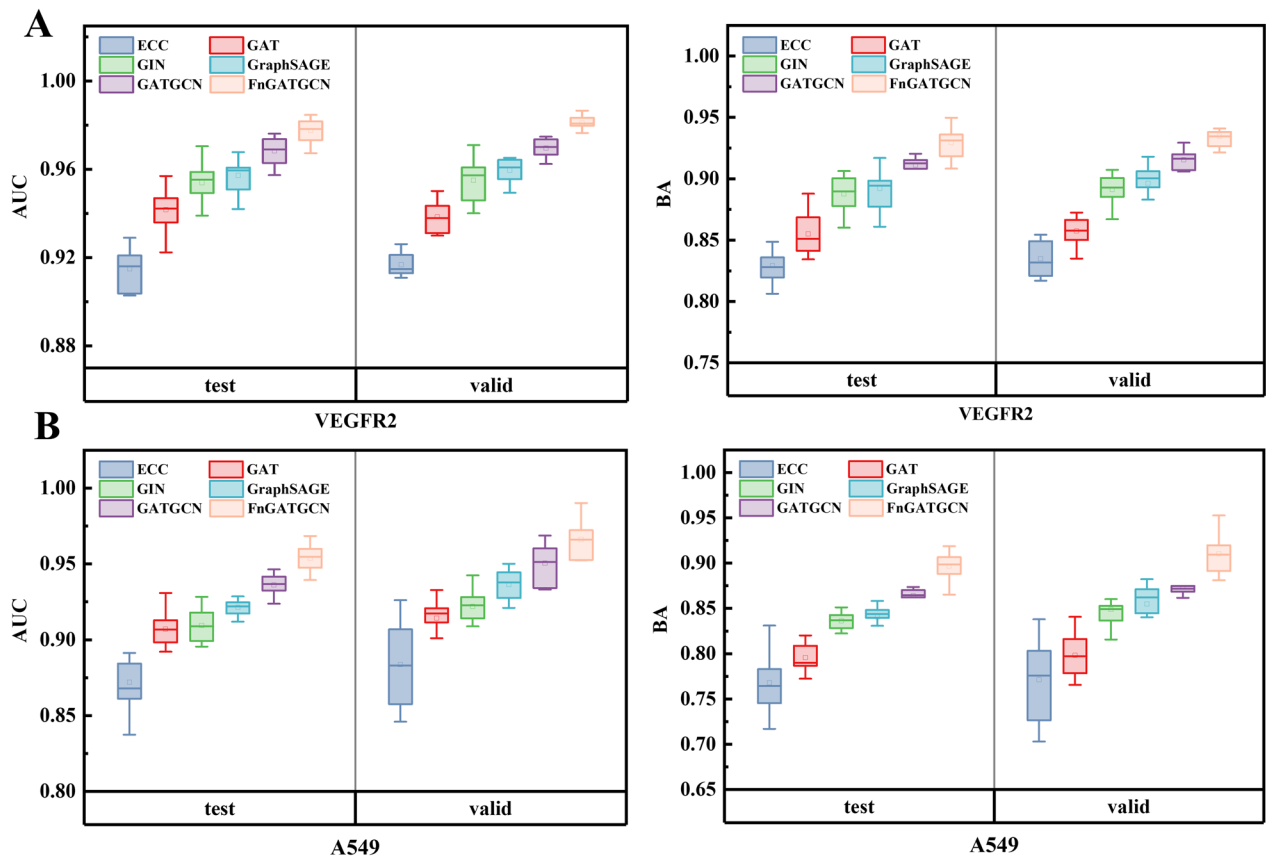
NSCLC cell lines and GES-1 cell ( $5\times 10^3$  per well) were seeded in 96-well plates and incubated for 24 h. The cells were then treated with Sorafenib (Macklin, China) as a positive control and various concentrations of the target compounds for 72 h. After treatment, 5 mg/mL MTT (Sigma, USA) was added, and the cells were incubated for 4 h. The supernatant was then removed, and 100  $\mu\text{L}$  of DMSO was added to dissolve the formazan crystals. Absorbance at 490 nm was measured using a microplate reader, and  $\text{IC}_{50}$  values were calculated using GraphPad Prism 9.0.

## Results and discussion

### Performance of FnGATGCN

In this study, we first developed an activity prediction model, FnGATGCN. The overall workflow and architecture of the model are depicted in Fig. 2.

We assessed the efficacy of the FnGATGCN model in comparison to five other graph neural network models. Notably, to mitigate the effects of data imbalance and ensure fair model comparisons within the same partition, we randomly generated 10 sets of training, validation, and test data. These models were then trained and evaluated on each of these partitions. The value range for AUC, BA, F1 is 0 to 1, and for MCC, it is -1 to 1, with higher values indicating better prediction performance. For simplicity, we present these values as percentages in this paper. As illustrated in Fig. 3, FnGATGCN surpassed all five alternative graph neural network models across both the A549 and VEGFR2 datasets, with GATGCN closely trailing behind. Specifically, as detailed in Table 3, on the A549 dataset, FnGATGCN demonstrated notable enhancements relative to the suboptimal GATGCN in terms of AUC, BA, F1, and MCC metrics, achieving increases of 1.76% (1.57%), 3.28% (3.68%), 3.83% (4.82%), and 5.23% (6.14%) on the test set (validation set), respectively. Likewise, on the VEGFR2 dataset, FnGATGCN outperformed GATGCN with relative improvements in AUC, BA, F1, and MCC metrics, exhibiting increases of 0.93% (1.14%), 1.84% (1.93%), 1.55% (1.69%), and 3.61% (3.94%) on the test set (validation set), respectively. These findings underscore the superior performance of the FnGATGCN model, which integrates molecular



**Fig. 3** Performance comparison of FnGATGCN with other models. **A** Performance evaluation of the FnGATGCN model in the VEGFR2 dataset. **B** Performance evaluation of the FnGATGCN model in the A549 dataset

**Table 3** Performance evaluation of the FnGATGCN model

Sets	Models	Validation set				Test set			
		AUC (%)	BA (%)	F1(%)	MCC (%)	AUC (%)	BA (%)	F1(%)	MCC (%)
A549	ECC	88.36 ± 2.74	77.12 ± 4.19	67.16 ± 6.44	58.71 ± 5.73	87.20 ± 2.14	76.79 ± 3.33	66.56 ± 5.35	57.69 ± 6.33
	GAT	91.42 ± 1.19	79.82 ± 2.30	72.31 ± 3.06	65.63 ± 3.06	90.70 ± 1.03	79.57 ± 1.57	72.14 ± 1.89	66.07 ± 2.26
	GIN	92.20 ± 0.98	84.89 ± 2.01	77.77 ± 2.88	70.46 ± 3.71	90.94 ± 1.06	83.63 ± 0.89	76.64 ± 1.94	69.16 ± 3.05
	GraphSAGE	93.64 ± 1.01	85.50 ± 2.39	78.51 ± 2.80	71.66 ± 2.81	92.14 ± 0.51	84.03 ± 1.73	76.32 ± 2.17	68.86 ± 2.47
	GATGCN	95.04 ± 1.28	87.33 ± 1.19	82.05 ± 2.26	76.35 ± 2.84	93.62 ± 0.69	86.38 ± 1.45	81.00 ± 1.57	75.12 ± 2.04
	<b>FnGATGCN</b>	<b>96.61 ± 1.14</b>	<b>91.01 ± 2.13</b>	<b>86.87 ± 2.91</b>	<b>82.49 ± 3.65</b>	<b>95.38 ± 0.83</b>	<b>89.66 ± 1.40</b>	<b>84.83 ± 1.87</b>	<b>80.35 ± 3.28</b>
VEGFR2	ECC	91.68 ± 0.50	83.67 ± 1.68	87.16 ± 1.24	68.96 ± 2.37	91.49 ± 0.96	83.30 ± 1.61	86.83 ± 1.21	67.73 ± 1.84
	GAT	93.84 ± 0.72	85.74 ± 1.10	88.98 ± 0.86	73.78 ± 1.73	94.18 ± 1.04	85.50 ± 1.60	88.89 ± 1.16	73.37 ± 2.60
	GIN	95.50 ± 0.90	89.13 ± 1.17	91.18 ± 0.94	79.37 ± 2.31	95.40 ± 1.09	88.76 ± 1.39	91.00 ± 1.08	78.79 ± 2.44
	GraphSAGE	95.96 ± 0.50	89.63 ± 1.61	91.28 ± 1.03	79.77 ± 2.28	95.73 ± 0.71	89.21 ± 1.63	91.00 ± 0.94	78.94 ± 2.31
	GATGCN	96.96 ± 0.41	91.55 ± 0.73	92.76 ± 0.44	83.29 ± 1.00	96.84 ± 0.63	91.11 ± 1.07	92.44 ± 0.67	82.47 ± 1.58
	<b>FnGATGCN</b>	<b>98.10 ± 0.36</b>	<b>93.48 ± 0.99</b>	<b>94.45 ± 0.68</b>	<b>87.23 ± 1.33</b>	<b>97.77 ± 0.52</b>	<b>92.95 ± 1.16</b>	<b>93.99 ± 0.86</b>	<b>86.08 ± 1.95</b>

Bold indicates the evaluation parameters of the optimal model



fingerprinting into the GATGCN framework, thereby offering the potential for enhanced screening of active molecules.

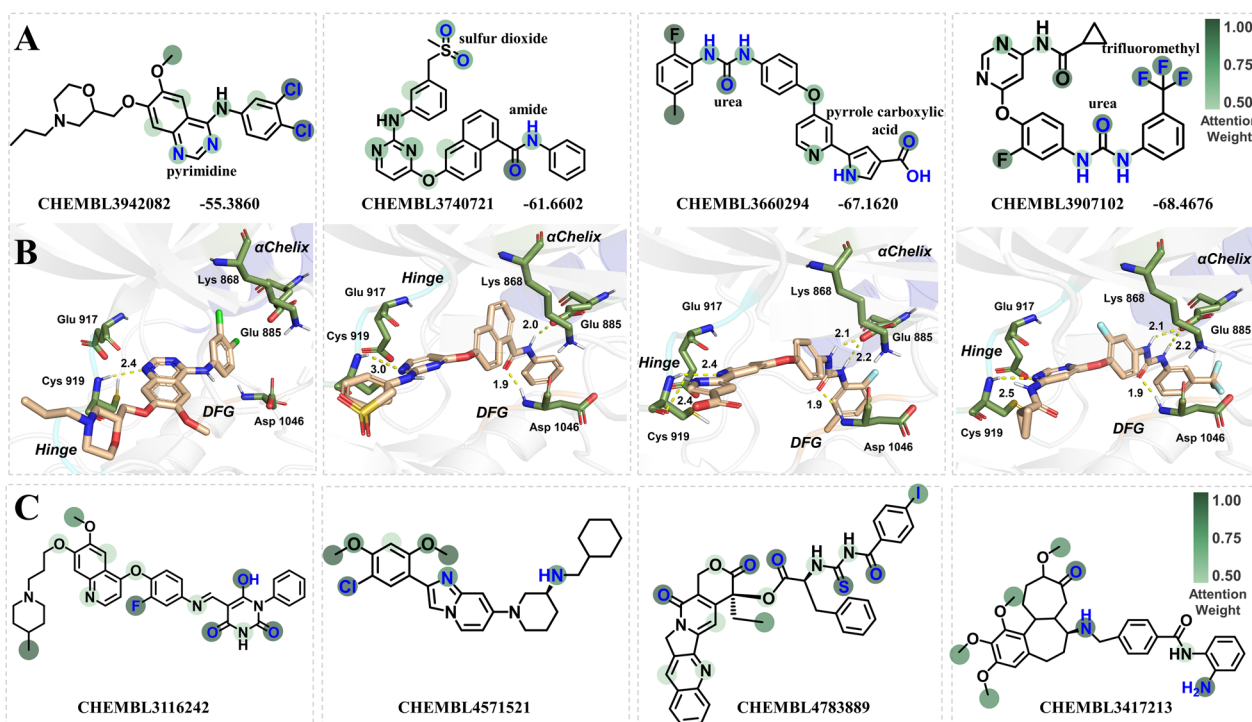
The results of each model run ten times are in Supplementary Tables S1-S8, taken together, VEGFR2-FnGATGCN-08 and A549-FnGATGCN-04 exhibit superior performance in both validation and test sets. VEGFR2-FnGATGCN-08 achieves 98.47% (98.66%), 94.04% (94.10%), 94.97% (94.64%), and 88.32% (88.25%) on the test set (validation set) for AUC, BA, F1 score, and MCC metrics, respectively. A549-FnGATGCN-04 achieves 96.01% (99.01%), 90.65% (95.26%), 87.24% (92.21%), and 82.70% (89.73%) on the test set (validation set) for the same metrics, respectively. Hence, we have chosen these two models for subsequent study.

### Feature visualization

FnGATGCN models offer better interpretability compared to the inherent “black box” nature of traditional machine learning models [46]. They achieve this by intuitively understanding the relationship between molecular structure and prediction results through the output of node weights within the molecular structure. In this section, we selected four molecules from each of the two datasets with prediction accuracies exceeding 98%. We utilized the model to calculate the attention weights of the top 30% of atoms in the molecular graph and mapped them to the

corresponding two-dimensional substructures, thus visualizing some important atoms and functional groups hidden in the data. Furthermore, we conducted molecular docking analyses to investigate the interaction between these features and the VEGFR2 target.

Figure 4A displays the heatmaps of the four VEGFR2 inhibitors, the attention weights highlight the pyrimidine and two chlorine atoms in CHEMBL3942082, the amide and sulfur dioxide structure in CHEMBL3740721, the urea and pyrrole carboxylic acid in CHEMBL3660294, and CHEMBL3907102 similarly contains urea, along with an additional trifluoromethyl group. In addition, the nitrogen atom serving as the core hydrogen bond acceptor focused on by the model effectively constrains the movement of the compound within the protein pocket, these key nodes constitute the fundamental scaffold of VEGFR2 inhibitors. The molecular docking results are depicted in Fig. 4B. We can observe that the atoms or functional groups interacting with the key amino acid residues in the active pocket of the VEGFR2 protein include pyrimidine, amide, urea, and nitrogen atoms, which are essentially consistent with the atoms focused on in the model. These observations suggest that the attention weights of the FnGATGCN model at the atomic level do possess chemical significance. The compounds inhibiting A549 cell proliferation, unlike VEGFR2 inhibitors, possess a more varied structure, as depicted in



**Fig. 4** Feature visualization. **A** Visualization analysis of atomic attention weights for VEGFR2 inhibitors. **B** Molecular docking of VEGFR2 inhibitors, the PDB code for CHEMBL3942082 is 3WZD, and the rest are 3WZE. **C** Visualization analysis of atomic attention weights for A549 cell inhibitors

Fig. 4C. The heatmap displays the crucial atoms of four A549 cell inhibitors, which receive extra attention in the model, possibly correlating with their exhibited anti-A549 cell proliferation activity.

### Screening of potentially active compounds from databases

In this section, we integrated the VEGFR2-FnGATGCN and A549-FnGATGCN models to construct a comprehensive drug screening model targeting NSCLC, which combines both target-based and phenotype-based screening strategies. Subsequently, we screened 3 million molecules from the ZINC database, resulting in a total of 2096 compounds with the potential to simultaneously inhibit VEGFR2 and suppress A549 proliferation ( $\text{EstPGood} > 0.5$ ), and these 2096 compounds were not included in the datasets used for model training.

Furthermore, Drug repurposing is a strategy aimed at identifying alternative applications for approved or investigational drugs beyond their originally intended medical indications. In this study, we also employed the FnGATGCN model for cross-screening and identified 390 potential molecules with VEGFR2 inhibitory activity from the A549 cell activity dataset. Similarly, within the VEGFR2 activity dataset, we discovered 508 potential molecules with anti-A549 proliferation activity. These molecules were clustered into 10 groups using the k-means algorithm and all central molecules adhere to Lipinski's Rule of Five. Detailed results are in Supplementary Tables S9 and S10.

### Docking and Visual Inspection

Molecular docking provides a nuanced understanding at the molecular level of ligand–protein interactions, while also evaluating the stability of resulting complexes [47]. Initially, we retrieved the endogenous ligand Sorafenib from the VEGFR2 crystal structure (PDB ID: 3WZE; resolution 1.90 Å) and redocked it into the receptor's active pocket. Remarkably, the RMSD value of the redocked pose was 0.28 Å, falling below the threshold of 1.90 Å, thus demonstrating the precision of our docking methodology and parameters [48]. Subsequently, we docked the 2096 compounds obtained from the previous step with VEGFR2. The screening process considered several key factors: (1) Ensuring the ligand's chemical structure orientation complements the protein's active pocket. (2) Establishing stable hydrogen bond interactions between the ligand and key amino acid residues within the protein's active pocket, such as Cys 919 and Cys 917 in the hinge region, as well as Asp 1046 and Glu 885 near the DFG sequence. (3) Ensuring that the CDOCKER interaction energy is less than  $-50 \text{ kcal}\cdot\text{mol}^{-1}$ . (4) Encouraging diversity in compound structures. Based on these

criteria, a total of 11 compounds were identified as promising candidates (Table 4).

The molecular docking analysis of potential active molecules and Sorafenib interacting with VEGFR2 is presented in Fig. 5. Concretely, Sorafenib fits well into the VEGFR2 protein pocket, with its core N-methylpicolinamide located in the hinge region forming hydrogen bonds at distances of 2.3 Å and 2.4 Å with amino acid residues, respectively. The 4-phenoxy group adopts a nearly perpendicular conformation to the core through certain bond angles. Meanwhile, the 4-chloro-3-trifluoromethyl phenyl group extends into the rear pocket region, engaging in hydrophobic interactions. The urea structure forms three hydrogen bonds with Glu 885 and Asp 1046 at distances of 2.2 Å, 2.0 Å, and 1.9 Å, respectively.

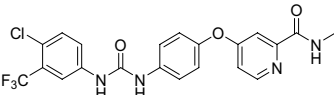
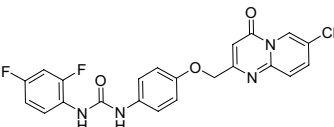
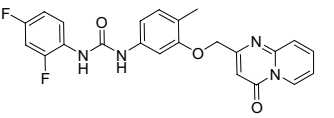
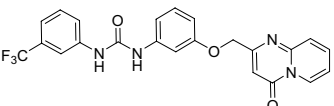
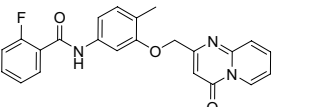
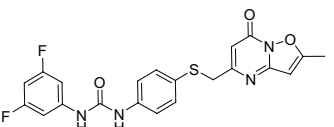
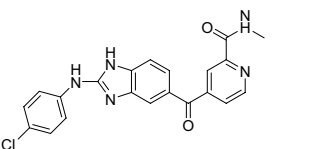
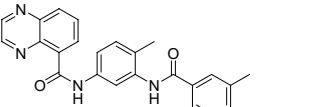
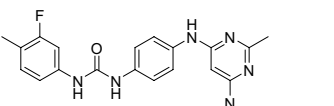
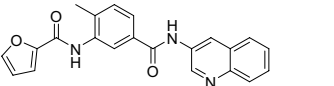
It is notable that the 11 compounds exhibit analogous spatial orientation and functional mechanisms to Sorafenib. Specifically, their structures can be divided into three parts: the core part near the protein hinge region, the part near the DFG sequence and extending the terminal group into the protein's rear pocket, and the linker part connecting the first two. Compound Z1 exemplifies this, as its core docks near the hinge, forming a hydrogen bond with Cys919, while the urea group near the DGF sequence forms three hydrogen bonds with Glu 885 and Asp 1046, and the terminal group inserts into the rear pocket, fostering hydrophobic interactions. Additionally, compounds Z-2, Z-3, Z-4, and Z-5 display similar spatial orientation and interactions as Z-1. Z-6 shares sorafenib's core structure, exhibiting analogous activity near the hinge, with its imidazole and amino groups also forming hydrogen bonds with Glu 885 and Asp 1046. Moreover, Z-7, Z-8, and Z-11 form hydrogen bonds with Glu 885 and Asp 1046 through their amide or urea structures, while Z-9 and Z-10, featuring the same quinoline core, form hydrogen bonds with Cys 919 in the hinge region through their nitrogen atoms.

Principal Component Analysis (PCA) discloses that compounds Z1 to Z11 are located within the chemical space of active compounds in both the A549 dataset and the VEGFR2 dataset. This finding suggests that Z1-Z11 have chemical structures similar to those of the active compounds. Moreover, to some extent, this result verifies the reliability of the activity predictions of Z1-Z11 by the FnGATGCN model, since it is based on an understanding of the known chemical space (Figure S1).

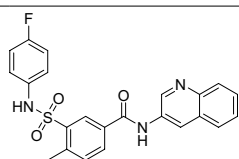
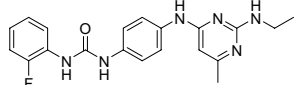
### Molecular dynamics simulation analysis

Molecular dynamics simulation and molecular docking are regarded as two complementary strategies for understanding the interaction between receptors and ligands [49]. We utilized molecular dynamics simulation to evaluate the stability of the 11 complex systems mentioned

**Table 4** Structural, CDOCKER interaction energy, and drug-likeness analysis of potentially active molecules

Compounds	ZINC ID	Structure	CDOCKER interaction energy (kcal·mol <sup>-1</sup> )	RO5
Sorafenib	-		-72.3670	0
Z1	ZINC33068301		-62.2536	0
Z2	ZINC33067816		-61.7714	0
Z3	ZINC8598095		-61.5327	0
Z4	ZINC33067799		-59.2919	0
Z5	ZINC41089729		-58.7570	0
Z6	ZINC85393782		-58.7475	0
Z7	ZINC65460318		-58.4082	0
Z8	ZINC33326674		-57.2331	0
Z9	ZINC23353178		-56.4581	0

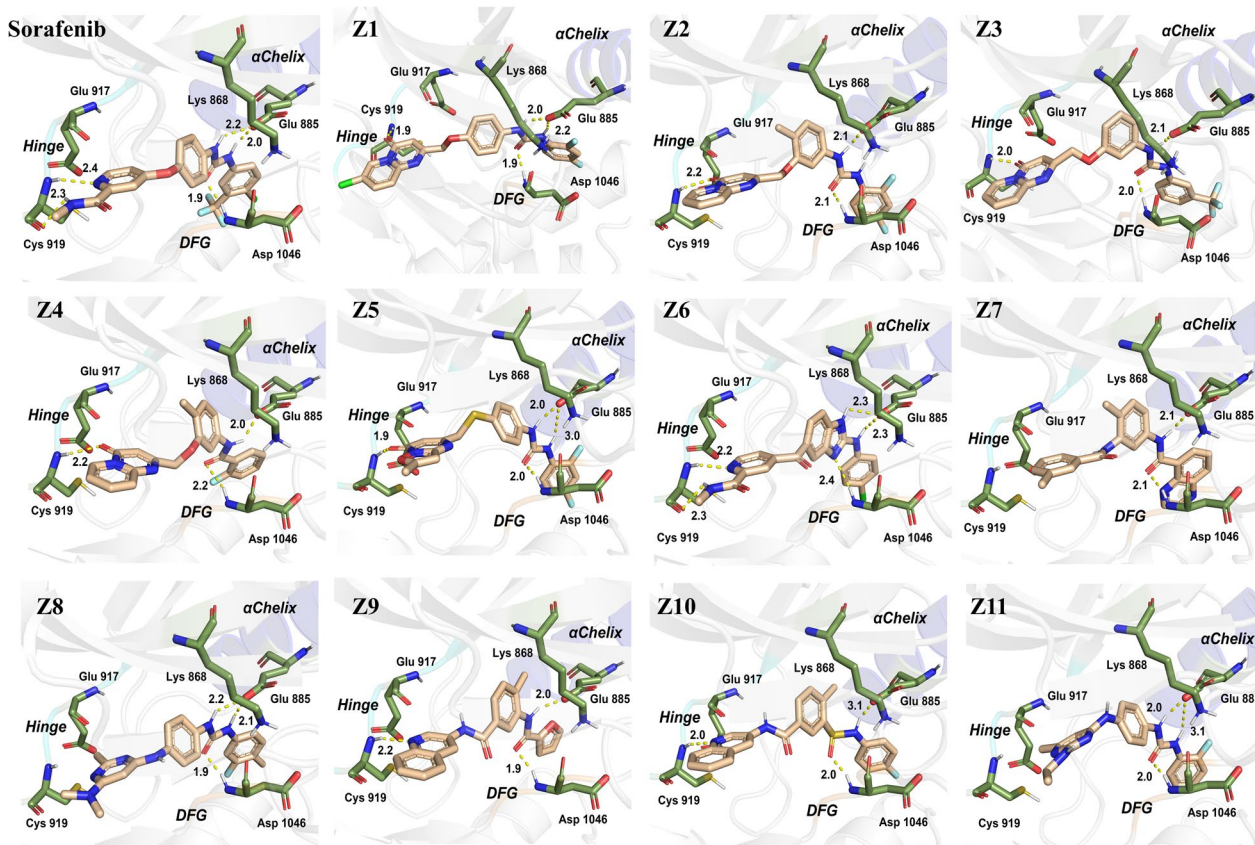
**Table 4** (continued)

Compounds	ZINC ID	Structure	CDOCKER interaction energy (kcal·mol <sup>-1</sup> )	RO5
Z10	ZINC170642714		-53.6049	0
Z11	ZINC64801237		-52.7258	0

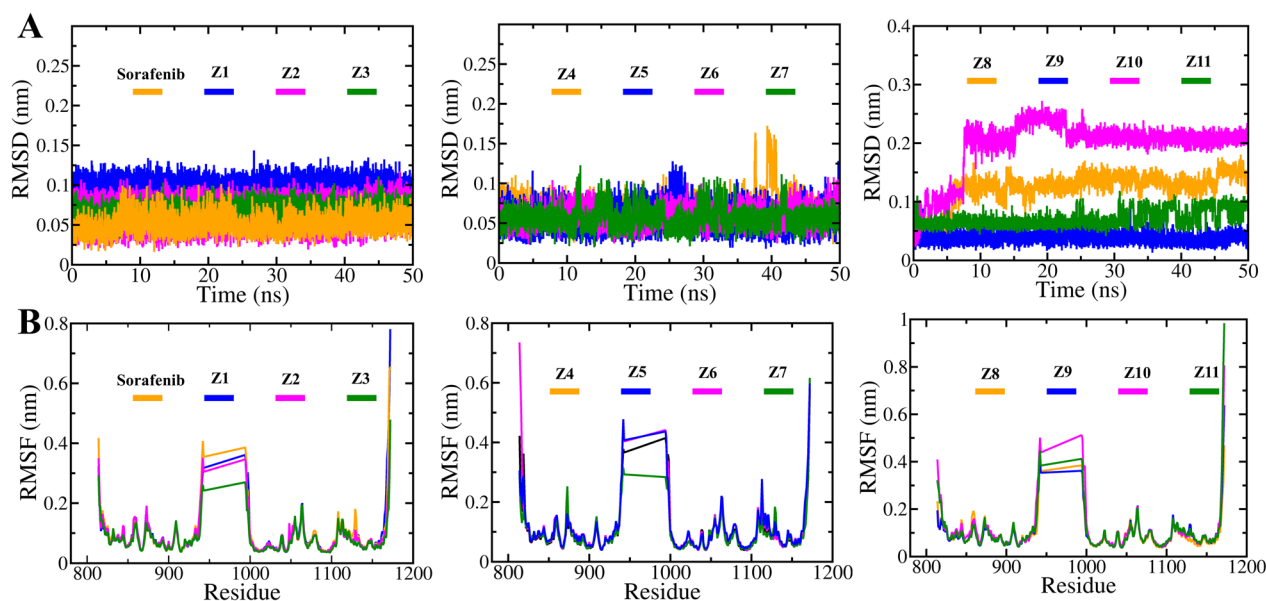
<sup>a</sup> RO5: Besides the Lipinski's rule of five

above, contrasting them with Sorafenib. The RMSD of 12 protein–ligand complexes is depicted in Fig. 6A. Compounds Z1–Z7, Z9, and Z10 display similar RMSD trends to the positive control Sorafenib, with no significant marginal effects observed, and their RMSD values are all less than 0.15 Å. They show stable fluctuations within a certain range over 50 ns, indicating their ability to stably bind in the VEGFR2 protein pocket. At the initial stage, noticeable marginal effects were observed for

Z8 and Z10, but they both converged after 10 ns, with convergent RMSD values less than 0.2 Å and 0.25 Å, respectively. Among them, Z10 underwent the largest conformational change compared to the initial conformation. Upon examining its trajectory, a significant conformational change occurred in the sulfamoyl group of Z10, leading to the disruption of its original hydrogen bonds with Glu 885 and Asp 1046. However, stable



**Fig. 5** Molecular docking analysis of potential active molecules interacting with VEGFR2



**Fig. 6** Simulation trajectory analysis of 12 protein–ligand complexes. **A** RMSD plots of the ligands; **B** RMSF plots of the proteins

**Table 5** Complex combined with free energy analysis ( $\text{kJ}\cdot\text{mol}^{-1}$ )

Complex	Contribution						
	$\Delta E_{\text{vdw}}$	$\Delta E_{\text{ele}}$	$\Delta G_{\text{polar}}$	$\Delta G_{\text{nonpolar}}$	$\Delta H$	$-\Delta S$	$\Delta G_{\text{bind}}$
Sorafenib	−247.664	−108.609	195.857	−30.97	−191.386	22.888	−168.498
Z1	−234.286	−85.119	200.748	−30.983	−149.64	9.19	−140.45
Z2	−271.283	−76.399	216.318	−30.644	−162.008	15.318	−146.69
Z3	−263.889	−94.331	211.432	−31.446	−178.234	25.824	−152.41
Z4	−241.392	−70.387	183.865	−28.857	−156.771	19.538	−137.233
Z5	−30.665	−288.366	202.333	−30.665	−147.363	16.457	−130.906
Z6	−247.358	−114.651	213.991	−28.466	−176.484	14.408	−162.076
Z7	−246.843	−57.449	193.834	−30.262	−140.72	10.313	−130.407
Z8	−224.756	−80.035	174.499	−31.575	−161.867	32.862	−129.005
Z9	−231.982	−47.99	174.409	−28.257	−133.82	10.035	−123.785
Z10	−250.256	−52.985	192.322	−29.924	−140.843	31.289	−109.554
Z11	−221.546	−75.332	177.818	−29.672	−148.732	35.324	−113.408

hydrogen bonds could still be formed with Cys 919 in the hinge region.

RMSF is employed to assess the flexibility of protein residues. Figure 6B illustrates the fluctuation of amino acid residues during the Molecular dynamics simulation of 12 systems. Overall, all compounds display highly analogous variation trends to the Sorafenib. Notably, a deletion is observed in the amino acid sequence of the VEGFR2 protein used, spanning positions 944–993. This deletion is located within the kinase insert domain (KID), situated in the kinase C-terminal lobe and linking helices

$\alpha$ D and  $\alpha$ E (933–1000), a region distant from the catalytic site. Studies indicate that while the majority of the KID is dispensable for catalysis, a few residues are necessary to form a bridge between  $\alpha$ D and  $\alpha$ E, ensuring the integrity of the kinase structure [50]. Consequently, the partial deletion in the protein sequence is not anticipated to unduly impact the outcomes. Within our systems, Pro 937–Glu 943 and Asp 994–Phe 999 in the KID region exhibit increased flexibility, suggesting their non-critical role in ligand–protein binding. Conversely, regions 915–935 and 1015–1050 demonstrate considerable

**Table 6** In vitro VEGFR2 inhibitory activity and antiproliferative activity of Z1-Z6

Complex	IC <sub>50</sub> (μM)				
	VEGFR2	A549 <sup>a</sup>	NCI-H23 <sup>a</sup>	NCI-H460 <sup>a</sup>	GES-1 <sup>a</sup>
Z1	7.05	6.71 ± 0.56	5.04 ± 0.82	6.43 ± 1.08	35.44 ± 3.34
Z2	4.29	8.56 ± 0.91	9.26 ± 0.77	7.45 ± 0.43	26.56 ± 2.03
Z3	0.88	4.23 ± 0.45	4.78 ± 0.62	5.24 ± 1.13	19.78 ± 1.93
Z4	18.92	12.46 ± 1.06	9.47 ± 1.23	11.24 ± 1.60	32.95 ± 2.54
Z5	8.25	18.59 ± 1.28	16.55 ± 3.13	14.56 ± 0.52	17.56 ± 2.11
Z6	2.68	9.19 ± 0.66	8.45 ± 1.31	9.56 ± 0.98	28.34 ± 2.88
Sorafenib	0.034	5.58 ± 0.88	4.25 ± 0.42	3.81 ± 0.71	11.09 ± 1.08

<sup>a</sup> IC<sub>50</sub> values are presented as the means ± SD of triplicate experiments

stability, implying their significance in ligand–protein interactions. These regions encompass the protein's hinge region, DFG motif, and represent stable hydrogen interaction zones. The above results support the stability and reliability of the docking results.

#### Binding free energy analysis

To assess the binding affinity of the 11 potential VEGFR2 inhibitors, we utilized the MM-PBSA method by calculating the binding free energy ( $\Delta G_{\text{bind}}$ ) of the system (Table 5). The  $\Delta G_{\text{bind}}$  values for all compounds interacting with the target protein were found to be below  $-100 \text{ kJ}\cdot\text{mol}^{-1}$ , indicating their ability to stably bind within the active pocket of VEGFR2. Notably, compound Z6 displayed the lowest binding free energy of  $-162.076 \text{ kJ}\cdot\text{mol}^{-1}$ , closely resembling that of the positive control sorafenib at  $-168.498 \text{ kJ}\cdot\text{mol}^{-1}$ .

#### Biological evaluation

Upon retrieval, there were no reports on the biological activities of the screened compounds Z1-Z11 against VEGFR2, A549, and other NSCLC cell lines. To further validate the reliability of the model and screening method, we assessed the biological activity of the top six compounds (Z1-Z6) (Figure S2) with the highest predicted binding affinities. As shown in Table 6, all six compounds exhibited effective inhibitory activity against VEGFR2, with IC<sub>50</sub> values ranging from 0.88 to 18.92 μM. Additionally, these compounds demonstrated notable anti-proliferative activity against A549 cells, with IC<sub>50</sub> values ranging from 4.23 to 18.58 μM. Notably, compound Z3 demonstrates strong inhibitory activity against both VEGFR2 (IC<sub>50</sub> = 0.88 μM) and A549 cells (4.23 ± 0.45 μM). Although its activity against VEGFR2 is not as potent as Sorafenib (IC<sub>50</sub> = 0.034 μM), it remains a promising candidate. This result demonstrates that through screening, the hit compounds have

been successfully targeted to VEGFR2 and A549, further validating the reliability of the FnGATGCN model.

We further evaluated the in vitro antiproliferative activities of Z1-Z6 in two NSCLC cell lines (NCI-H23 and NCI-H460) and a normal cell line (GES-1). The results showed that the compounds also had good antiproliferative activities against NCI-H23 and NCI-H460 with IC<sub>50</sub> values ranging from 4.78 to 16.55 μM. Compared with their antiproliferative activities against NSCLC cell lines, except for Z5, the compounds had lower toxicity to GES-1. Although the toxicity of Z1-Z3 and Z6 to GES-1 may be lower than that of positive drugs, it is still not ideal. This may be related to the hit compounds acting on multiple targets. Considering the weak VEGFR2 inhibitory activity of Z4 and the high toxicity of Z5, Z1-Z4 and Z6 may be compounds with greater potential for drug development, and they require further extensive in vitro/in vivo experimental evaluations.

#### Conclusion

Active molecules screened against VEGFR2 may not necessarily be effective against NSCLC phenotype cell line A549. Therefore, the development of anti-NSCLC drugs targeting VEGFR2 still faces challenges. This study integrated the biological activity data of VEGFR2 and A549 to construct a novel deep learning framework called FnGATGCN. This model integrates GAT-GCN with molecular fingerprint-based feature extraction methods and demonstrates outstanding accuracy and robustness in anti-NSCLC drug screening. The model performance evaluation indicates that the AUC, BA, F1, and MCC values of FnGATGCN are superior to those of other published graph neural network models. Additionally, the feature visualization results demonstrate that the FnGATGCN model can allocate different weights to atoms based on active molecule characteristics and the attention weights do have

chemical significance. This further underscores the model's strong interpretability.

Subsequently, the model was employed to screen 2475 compounds from 3 million small molecules, all exhibiting activity against both targets with a predicted probability exceeding 50%. After considering the molecular docking results comprehensively, 11 potential active compounds were selected. They can form stable hydrogen bonds with key amino acid residues Asp1046, Glu885, and Cys919 within the active pocket of VEGFR2. Molecular dynamics results show that the simulation trajectories of these 11 complexes are essentially consistent with those of the positive control drug sorafenib, with relatively small binding free energy ( $\Delta G_{\text{bind}}$ ) values, indicating high affinity binding to VEGFR2 and stable interaction with amino acid residues. Furthermore, based on the results of molecular dynamics simulations, biological activity tests were conducted on the top 6 compounds, and Z1-Z3 and Z6 exhibited good biological activities. However, since selectivity tests have not been performed yet, the observed effects may also be related to interactions with off-target proteins. Further *in vivo/in vitro* experiments are needed to clarify the mechanism of action and biological activity of the hit compounds to discover potential clinical candidate compounds. Despite some shortcomings, this model is a promising tool for screening anti-NSCLC drugs, enhancing screening efficiency and success rates. It can be further refined and its application can be extended to other targets. This approach will hopefully become a powerful means of accelerating the discovery of clinical candidate drugs.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-024-05893-2>.

Additional file 1: Table S1. Performance evaluation of the models in the VEGFR2 (AUC%). Table S2. Performance evaluation of the models in the VEGFR2 (BA%). Table S3. Performance evaluation of the models in the VEGFR2 (F1%). Table S4. Performance evaluation of the models in the VEGFR2 (MCC%). Table S5. Performance evaluation of the models in the A549 (AUC%). Table S6. Performance evaluation of the models in the A549 (BA%). Table S7. Performance evaluation of the models in the A549 (F1%). Table S8. Performance evaluation of the models in the A549 (MCC%). Table S9. The potential VEGFR2 inhibitors from A549 cell growth inhibitors. Table S10. The potential A549 cell growth inhibitors from VEGFR2 inhibitors. Table S11. The SMILES of Z1-Z11 and Sorafenib. Figure S1. The chemical space of compounds. (A) Z1-Z11 and active compounds from VEGFR2 datasets. (B) Z1-Z11 and active compounds from A549 datasets. Figure S2. The structural analysis spectra of Z1-Z6 provided by the supplier.

Additional file 2: The .mdp files used for GROMACS.

## Acknowledgements

We gratefully acknowledge the invaluable contributions of all individuals involved in this research.

## Author contributions

Z. W.: conceptualization, methodology, data curation, and writing-original draft. L. S.: methodology, software, validation, and data curation. X. Y.: supervision and software. J. H.: validation, and data curation. F. Y.: software, Formal analysis, and visualization. Y. C.: supervision, methodology, resources and writing-review & editing. All authors have read and agreed to the submission of the manuscript.

## Funding

This work was supported by the hospital fund of The First Affiliated Hospital of Xi'an Jiaotong University under the grant number 2024-QN-44 and PT002191.

## Availability of data and materials

Compounds exhibiting inhibitory activity against both VEGFR2 or A549 cell lines were sourced from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>). Approximately 15 million compounds were retrieved from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>). The dataset used for virtual screening was sourced from ZINC database (<https://zinc.docking.org/>). The source codes of FnGATGCN are available at <https://github.com/zixiaowang1995/FnGATGCN>.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

All authors of this article agree to publish.

### Competing interests

The authors declare no competing interests.

## Author details

<sup>1</sup>Department of Pharmacy, Honghui Hospital, Xi'an Jiaotong University, Xi'an 710054, China. <sup>2</sup>Department of Pharmacy, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China. <sup>3</sup>State Key Laboratory of Natural Medicines, Jiangsu Key Laboratory of Drug Discovery for Metabolic Diseases, Center of Drug Discovery, China Pharmaceutical University, Nanjing 210009, China.

Received: 11 August 2024 Accepted: 14 November 2024

Published online: 03 December 2024

## References

- Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature*. 2018;553(7689):446–54.
- Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(1):17–48.
- Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. *Nat Rev Clin Oncol*. 2023;20(9):624–39.
- Xie C, Zhou X, Liang C, Li X, Ge M, Chen Y, et al. Apatinib triggers autophagic and apoptotic cell death via VEGFR2/STAT3/PD-L1 and ROS/Nrf2/p62 signaling in lung cancer. *J Exp Clin Cancer Res*. 2021;40(1):266.
- Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. *Mol Cancer*. 2023;22(1):40.
- Arbour KC, Riely GJ. Systemic therapy for locally advanced and metastatic non-small cell lung cancer: a review. *JAMA*. 2019;322(8):764–74.
- Wang M, Herbst RS, Boshoff C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat Med*. 2021;27(8):1345–56.
- Skribek M, Rounis K, Tsakonas G, Ekman S. Complications following novel therapies for non-small cell lung cancer. *J Intern Med*. 2022;291(6):732–54.
- Xie J, Liu J, Liu H, Liang S, Lin M, et al. The antitumor effect of tanshinone IIA on anti-proliferation and decreasing VEGF/VEGFR2 expression on the human non-small cell lung cancer A549 cell line. *Acta Pharm Sin B*. 2015;5(6):554–63.

10. Abd El-Lateef HM, Elbastawesy MAI, Abdelghani Ibrahim TM, Khalaf MM, Gouda M, et al. Design, synthesis, docking study, and antiproliferative evaluation of novel schiff base-benzimidazole hybrids with VEGFR-2 inhibitory activity. *Molecules*. 2023;28(2):481.
11. Tsai CY, Wu JCC, Wu CJ, Chan SHH. Protective role of VEGF/VEGFR2 signaling against fatal fatality associated with hepatic encephalopathy via sustaining mitochondrial bioenergetics functions. *J Biomed Sci*. 2022;29(1):47.
12. Chatterjee S, Heukamp LC, Siobal M, Schöttle J, Wiecekorek C, Peifer M, et al. Tumor VEGF:VEGFR2 autocrine feed-forward loop triggers angiogenesis in lung cancer. *J Clin Invest*. 2013;123(4):1732–40.
13. Fontanella C, Ongaro E, Bolzonello S, Guardascione M, Fasola G, Aprile G. Clinical advances in the development of novel VEGFR2 inhibitors. *Ann Transl Med*. 2014;2(12):123.
14. Watanabe H, Ichihara E, Kayatani H, Makimoto G, Ninomiya K, Nishii K, et al. VEGFR2 blockade augments the effects of tyrosine kinase inhibitors by inhibiting angiogenesis and oncogenic signaling in oncogene-driven non-small-cell lung cancers. *Cancer Sci*. 2021;112(5):1853–64.
15. Nakagawa K, Garon EB, Seto T, Nishio M, Ponce Aix S, Paz-Ares L, et al. Ramucirumab plus erlotinib in patients with untreated, EGFR-mutated, advanced non-small-cell lung cancer (RELAY): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol*. 2019;20(12):1655–69.
16. Xu Y, Wang J, Wang X, Zhou X, Tang J, Jie X, et al. Targeting ADRB2 enhances sensitivity of non-small cell lung cancer to VEGFR2 tyrosine kinase inhibitors. *Cell Death Discov*. 2022;8(1):36.
17. Sadri A. Is target-based drug discovery efficient? discovery and “off-target” mechanisms of all drugs. *J Med Chem*. 2023;66(18):12651–77.
18. Ye Z, Chen F, Zeng J, Gao J, Zhang MQ. ScaffoldComb: a phenotype-based framework for drug combination virtual screening in large-scale chemical datasets. *Adv Sci*. 2021;8(24):e2102092.
19. Szabo M, Svensson Akusjärvi S, Saxena A, Liu J, Chandrasekar G, Kitambi SS. Cell and small animal models for phenotypic drug discovery. *Drug Des Devel Ther*. 2017;11:1957–67.
20. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem*. 2020;63(16):8749–60.
21. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, et al. PubChem's BioAssay database. *Nucleic Acids Res*. 2012;40(Database issue):D400–12.
22. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44(D1):D1075–9.
23. Gao J, Shen Z, Xie Y, Lu J, Lu Y, Chen S, et al. TransFoxMol: predicting molecular property with focused attention. *Brief Bioinform*. 2023. <https://doi.org/10.1093/bib/bba306>.
24. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans Neural Netw Learn Syst*. 2021;32(1):4–24.
25. Choo HY, Wee J, Shen C, Xia K. Fingerprint-enhanced graph attention network (FinGAT) model for antibiotic discovery. *J Chem Inf Model*. 2023;63(10):2928–35.
26. Zhang Y, Hu Y, Han N, Yang A, Liu X, Cai H. A survey of drug-target interaction and affinity prediction methods via graph neural networks. *Comput Biol Med*. 2023;163: 107136.
27. Liu X, Luo Y, Li P, Song S, Peng J. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput Biol*. 2021;17(8): e1009284.
28. Zhong Y, Zheng H, Chen X, Zhao Y, Gao T, Dong H, et al. DDI-GCN: drug-drug interaction prediction via explainable graph convolutional networks. *Artif Intell Med*. 2023;144: 102640.
29. Li Y, Liu J, Jiang Y, Liu Y, Lei B. Virtual adversarial training-based deep feature aggregation network from dynamic effective connectivity for MCI identification. *IEEE Trans Med Imaging*. 2022;41(1):237–51.
30. Jiang B, Wang B, Tang J, Luo B. GeCNs: graph elastic convolutional networks for data representation. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(9):4935–47.
31. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks, *ArXiv* (2017) abs/1710.10903. <https://doi.org/10.48550/arXiv.1710.10903>.
32. Qiu M, Liang X, Deng S, Li Y, Ke Y, Wang P, et al. A unified GCNN model for predicting CYP450 inhibitors by using graph convolutional neural networks with attention mechanism. *Comput Biol Med*. 2022;150: 106177.
33. Yu Z, Huang F, Zhao X, Xiao W, Zhang W. Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform*. 2021. <https://doi.org/10.1093/bib/bba243>.
34. Sun J, Wen M, Wang H, Ruan Y, Yang Q, Kang X, et al. Prediction of drug-likeness using graph convolutional attention network. *Bioinformatics*. 2022;38(23):5262–9.
35. Wang S, Chen W, Han P, Li X, Song T. RGN: residue-based graph attention and convolutional network for protein-protein interaction site prediction. *J Chem Inf Model*. 2022;62(23):5961–74.
36. Van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model*. 2022;62(23):5938–51.
37. Wu S, Fang Z, Tan J, Li M, Wang C, Guo Q, et al. DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *Gigascience*. 2021. <https://doi.org/10.1093/gigascience/giab056>.
38. Chen Z, Wang X, Huang J, Lu J, Zheng J. Deep attention and graphical neural network for multiple sclerosis lesion segmentation from MR imaging sequences. *IEEE J Biomed Health Inform*. 2022;26(3):1196–207.
39. Jiang J, Wang T, Wang B, Ma L, Guan Y. Gated tree-based graph attention network (GTGAT) for medical knowledge graph reasoning. *Artif Intell Med*. 2022;130: 102329.
40. Gan J, Hu R, Mo Y, Kang Z, Peng L, Zhu Y, et al. Multigraph Fusion for Dynamic Graph Convolutional Network. *IEEE Trans Neural Netw Learn Syst*. 2022.
41. Al-Balas QA, Amawi HA, Hassan MA, Qandil AM, Almaaytah AM, Mhaidat NM. Virtual lead identification of farnesyltransferase inhibitors based on ligand and structure-based pharmacophore techniques. *Pharmaceuticals*. 2013;6(6):700–15.
42. Lu T, Chen F. Multiwfn: a multifunctional wavefunction analyzer. *J Comput Chem*. 2012;33(5):580–92.
43. Neese F. Software update: the ORCA program system—version 5.0. *WIREs Comput Mol Sci*. 2022;12(5):e1606.
44. Lu T. Sobtop, Version [1.0(dev3.1)]. <http://sobereva.com/soft/Sobtop>. Accessed 10 Oct 2022.
45. Wang Z, Sun L, Xu Y, Liang P, Xu K, Huang J. Discovery of novel JAK1 inhibitors through combining machine learning, structure-based pharmacophore modeling and bio-evaluation. *J Transl Med*. 2023;21(1):579.
46. Wu Y, Li K, Li M, Pu X, Guo Y. Attention mechanism-based graph neural network model for effective activity prediction of SARS-CoV-2 main protease inhibitors: application to drug repurposing as potential COVID-19 therapy. *J Chem Inf Model*. 2023;63(22):7011–31.
47. Sun L, Wang Z, Yang Z, Liu X, Dong H. Virtual screening and structure-activity relationship study of novel BTK inhibitors in traditional Chinese Medicine for the treatment of rheumatoid arthritis. *J Biomol Struct Dyn*. 2023;41(24):15219–33.
48. Okamoto K, Ikemori-Kawada M, Jestel A, von König K, Funahashi Y, Matsushima T, et al. Distinct binding mode of multikinase inhibitor lenvatinib revealed by biochemical characterization. *ACS Med Chem Lett*. 2015;6(1):89–94.
49. Zhao Y, Yang H, Wu F, Luo X, Sun Q, Feng W, et al. Exploration of N-arylsulfonyl-indole-2-carboxamide derivatives as novel fructose-1,6-bisphosphatase inhibitors by molecular simulation. *Int J Mol Sci*. 2022;23(18):10259.
50. McTigue MA, Wickersham JA, Pinko C, Showalter RE, Parast CV, Tempczyk-Russell A, et al. Crystal structure of the kinase domain of human vascular endothelial growth factor receptor 2: a key enzyme in angiogenesis. *Structure*. 1999;7(3):319–30.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.