NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 18 (2026), 1837582
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

GraPNet: Protein-ligand interaction prediction based on graph learning

GraPNet: Graf öğrenmeye dayalı protein-ligand etkileşim tahmini

Turgay Batbat[1, *], Seda Mürütsoy[2]

[1] *Erciyes University, Department of Biomedical Engineering,38039, Kayseri, Türkiye*
[2] *Erciyes University, Graduate School of Natural and Applied Sciences, Department of Biomedical Engineering, 38039, Kayseri, Türkiye*

Abstract

Pharmaceutical research and development are highly time-consuming and costly processes. As computer processing capabilities increase, computer-aided design becomes increasingly important. This study aims to determine protein-ligand interactions that underpin drug discovery and design, using computational methods to reduce the time and cost of laboratory studies. In this study, ligand SMILES representations and protein fingerprint representations were used. Graph Learning, a modern artificial intelligence methodology, effectively simulates biomolecular interactions in similar problems and achieved rapid, effective results in protein-ligand interaction detection, demonstrating 80.96% accuracy on an experimentally validated dataset. This marks a significant milestone in the drug design process and serves as a valuable guide in drug discovery and biotechnology.

Keywords: Protein-ligand interaction, Graph learning, Drug design, Artificial intelligence, FASTA, Bioinformatics

Öz

Farmasötik araştırma ve geliştirme süreçleri oldukça zaman alıcı ve maliyetli çalışmalardır. Bilgisayar işlem kapasitesindeki artışla birlikte bilgisayar destekli tasarım yöntemleri giderek daha fazla önem kazanmaktadır. Bu çalışmada, ilaç keşfi ve tasarımının temelini oluşturan protein-ligand etkileşimlerinin belirlenmesi amaçlanmış, laboratuvar çalışmalarının süre ve maliyetlerini azaltmak için hesaplamalı yöntemlerden yararlanılmıştır. Çalışmada ligandların SMILES temsilleri ve proteinlerin parmak izi temsilleri kullanılmıştır. Modern yapay zekâ yaklaşımlarından biri olan Graf Öğrenme, benzer problemlerde biyomoleküler etkileşimleri modelleyebilmektedir ve bu çalışmada protein-ligand etkileşimlerinin tespitinde hızlı ve başarılı sonuçlar sağlamıştır. Önerilen yöntem, deneysel olarak doğrulanmış bir veri kümesi üzerinde %80,96 doğruluk oranına ulaşmıştır. Elde edilen bulgular, ilaç tasarım sürecinde önemli bir adımı temsil etmekte olup ilaç keşfi ve biyoteknoloji alanlarında değerli bir rehber niteliği taşımaktadır.

Anahtar kelimeler: Protein-ligand etkileşimi, Graf öğrenme, İlaç tasarımı, Yapay zekâ, FASTA, Biyoenformatik

1 Introduction

Drug research and development is a multi-step, control-oriented process that begins with the identification and investigation of a disease, followed by the identification of a suitable target for inhibition, clinical testing, manufacturing approval, and market launch [1]. Clinical testing and laboratory studies constitute the most costly and time-consuming part of the drug discovery process, which can take years and cost millions of dollars [2], [3]. Despite the significant time and money invested, the success rate remains below 15% [4]. In recent years, with the development of computational methods and artificial intelligence techniques, the success rate in drug discovery has increased significantly; the process has accelerated, and the cost has decreased [5].

One of the most widely used methodologies, which is highly suitable for simulation and drug design, is the analytical identification of related proteins. For this process, proteins are extracted from biological samples, such as blood and tissue, using chemical methods and centrifugation. Proteins are identified by mass spectrometry (MS), and their amino acid sequences are subsequently determined by sequencing [6]. These sequences are stored in FASTA format, a standard for analysis and database searches, and are used with bioinformatics software. The format can be used by many software tools to store, compare, and analyze proteins [7].

Accurately identifying and representing the chemical structures of ligands that interact with proteins and elicit specific biological responses plays an influential role in drug development and discovery. The Simplified Molecular Input Line Entry System (SMILES) is used to represent the chemical structures of ligands in a standardized format that includes all necessary information. SMILES notation describes the atoms and bonds of a molecule with ASCII characters (strings), making chemical structures readily usable in information processing and databases. In protein-ligand interaction studies, ligand chemical structures can be

* Sorumlu yazar / Corresponding author, e-posta / e-mail: turgaybatbat@erciyes.edu.tr (T. Batbat)
Geliş / Received: 08.12.2025 Kabul / Accepted: 30.07.2026 Yayımlanma / Published: 24.08.2026
doi: 10.28948/ngumuh.1837582

processed quickly and accurately using SMILES representations, which facilitate binding affinity calculations and related computer-aided analyses [8].

Protein-ligand interaction affinity detection, which is the basis of compound selection for a drug candidate, refers to the tendency of a small-molecule ligand to bind to the protein. Several machine learning methods, such as support vector machines (SVMs), random forests, and k-nearest neighbors (k-NNs), have been applied to predict protein-ligand binding affinity. Applications using the same dataset are discussed in further sections, along with their respective parameters. In this area, two different approaches can be evaluated: binary classification for early stage studies and regression studies to predict affinity scores directly [9]. The proposed method aimed for binary classification to increase understanding and to underscore different scoring strategies. In Li et al.'s study [10], protein-ligand binding affinity was predicted using classical machine learning techniques. High accuracy was achieved through feature engineering and model optimization, and the random forest method was able to keep learning with an increasing dataset size. In the study by Jiménez-Rosés et al. [11], structures were generated using SMILES representations and docking, and machine learning methods were employed to predict the pharmacological effects of ligands. However, the results were only indicative for studies in this field and were not yet suitable for real-life applications.

Deep Learning is a machine learning technique that uses various structures based on artificial neural networks to automate feature extraction and model-building, thereby enabling learning from data and reducing the need for human intervention [12]. Multilayer neural networks can model intricate data patterns and relationships. In Ragoza et al.'s study, a deep learning model using three-dimensional convolutional neural networks was developed to predict protein-ligand binding affinity by identifying binding sites [13]. The model was trained and tested on a large dataset of ligands and proteins, and the results showed that the deep learning model achieved higher accuracy and generalization than traditional machine learning methods. Although this study emphasized the potential of deep learning techniques in interaction analysis and the discovery of new drug candidates, the CNN method had identified weaknesses and opportunities for improvement [13].

Geometric deep learning (GDL) is a broad term for new techniques that use deep neural networks on non-Euclidean data, such as graphs and manifolds [14]. GDL encompasses learning and inference methods applied to data represented in a graph structure [15]. In the field of social network analysis, Graph Convolutional Networks (GCN) methods have been used to understand user behavior and community structure by modeling the relationships between nodes [16]. In materials science and molecular engineering, the SchNet model has been successfully applied to the design and simulation of new materials by calculating interatomic forces and energies [17]. These studies demonstrate that GDL has a wide range of applications across various scientific and industrial fields. This study aims to determine protein-ligand interactions, which can naturally be represented as a graph structure. Proteins and ligands are represented as graph nodes, while their interactions are expressed as edges. In this context, graph neural networks (GNNs) offer an ideal approach for analyzing biological networks [18]. GNNs can learn from node features and a topological approach; thus, the proposed method determines whether protein-ligand interactions are active or inactive.

This study presents a new method for rapid, high-accuracy detection of protein-ligand interactions using geometric deep learning in drug discovery and design. This method requires vectors of 1024-2048 bits for this purpose. The second part of the article, the artificial intelligence section, explains the methods and representations used for proteins and ligands. The third section presents the study's findings, and the final section provides conclusions and discussions.

2 Materials and methods

In this study, benchmark datasets were used, including FASTA sequences of proteins obtained from human biological samples and small ligand molecules that inhibit or have therapeutic effects for specific diseases through their interactions with proteins [19]. Proteins are the most abundant biomolecules in living organisms and play a crucial role in performing numerous critical functions within cells [20]. These molecules perform a diverse array of functions, including catalyzing enzymatic reactions, constructing cellular structures, transmitting signals, storing nutrients, contracting muscles, defending the body, and transporting molecules between cells [21]. The functions of proteins are directly determined by their three-dimensional structures, making it crucial to understand these structures for detailed biochemical studies [22]. Researchers typically determine these structures using methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM), and the resulting data are stored in databases like the Protein Data Bank (PDB) [23]. This structural information is vital for elucidating protein function and identifying potential drug targets among their interacting partners (ligands). Consequently, protein structural biology is pivotal in drug discovery and development; bioinformatics tools are employed to analyze this information, facilitating the creation of new therapeutic agents.

Ligands are molecules that regulate biological processes through specific interactions with biomolecules. Typically, ligands (small molecules or ions) bind to protein receptors, thereby influencing cellular signaling, enzymatic activity, and other biological functions [24]. The binding sites and bond strengths with target proteins determine their biological activity, making the study of ligand-protein interactions crucial for drug design. Advances in structural biology and biochemistry have enhanced our understanding of these molecular interactions. In drug discovery, ligands are evaluated as potential therapeutic agents. High-affinity ligands can bind target proteins effectively even at low concentrations, thereby optimizing therapeutic effects. Ligand design and optimization are conducted using high-throughput screening (HTS) techniques and computer-aided drug design (CADD) methods [25]. Systems such as

SMILES represent ligand structural information in a compact, analyzable format, enabling the processing of large datasets and the rapid evaluation of potential ligands [26]. These methodologies expedite the discovery of new drug candidates, offering more effective and safer treatments for diseases.

2.1 Mathematical framework of the proposed graph learning

The primary mathematical innovation of GraPNet lies in its transition from Euclidean processing (1D-CNNs used in DeepDTA) to a non-Euclidean graph convolutional framework. This allows the model to preserve the spatial relationships of chemical motifs. The graph convolution operation in the model is defined as:

$$H^{l+1} = \sigma\left(\tilde{D}^{\left(-\frac{1}{2}\right)}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{l}W^{l}\right) \quad (1)$$

In Equation (1):

- $\tilde{A} = A + I_N$: This represents the adjacency matrix of the protein-ligand graph with added self-loops (I_N), ensuring that the features of a node (atom or residue) are preserved during the update process [27].
- $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$: A diagonal degree matrix used for symmetric normalization, which prevents the numerical instability and gradient explosion common in deep graph networks [27].
- W^l: The learnable weight matrix that enables the model to capture high-order interactions between ligand fingerprints and protein motifs [28].

This approach allows the model to be invariant to the rotation or translation of the molecular structures, a property that sequence-based methods like DeepDTA (0.70 accuracy) cannot inherently possess [29].

2.2 SMILES representation of ligands

With the use of computer technologies, such as machine learning, in predicting chemical properties, not only in vivo and in vitro analyses but also in silico analyses have gained importance in recent years. Some computer-readable file formats for chemical compounds have been defined for in silico analysis, facilitating the screening of new lead compounds, such as those used in drug discovery. Molecular file format (MOL), Structure data file (SDF), Fingerprints, and SMILES are the primary representations used for these purposes [30]. SMILES, which has been increasingly popular over the last decade, was proposed by Weininger [31]. It uses a special grammar and a fixed alphabet of characters to describe the atoms and structure of a chemical compound, providing a unique linear representation. SMILES enables the virtual screening of chemical compounds and the identification of functional substructures, known as chemical motifs.

SMILES notation employs two sets of symbols: atomic symbols and special SMILES symbols. In SMILES, atoms are represented by their atomic symbols, with double and triple bonds indicated by "=" and "#", respectively. Chemical rings are depicted by breaking one bond in the ring and assigning an integer to the atoms at the ends of this bond. Branches are shown using parentheses. The SMILES string for a compound, such as Aspirin (CC(=O)OC1=CC=CC=C1C(=O)O), is converted into a unique representation to avoid ambiguity [32]. This method ensures a single, canonical SMILES representation for each compound [32].

2.3 Obtaining protein in FASTA format

Obtaining protein sequences in FASTA format begins with biochemical laboratory studies and involves a series of bioinformatics steps. A biological sample, the source of the protein to be analyzed, is collected. Sterilization is essential at this stage. It is crucial to ensure the sample is collected without contamination; otherwise, false results may arise. Researchers apply chemical methods and centrifugation processes to the biological sample to disrupt the cells and extract the proteins. Researchers then purify the extracted proteins from other biomolecules. High-resolution sequencing techniques, such as MS, obtain amino acid sequences of purified proteins, which are then converted to FASTA format using bioinformatics software.

The FASTA format consists of a header line (starting with ">") followed by the sequence for each protein, with the header line containing the protein name and identifying information, while sequences are typically divided into lines of 60-80 characters each. This format creates readily usable data for sequence comparison and analysis.

2.4 Protein fingerprint conversion from FASTA

In the GraPNet framework, the Biopython library is employed to handle protein sequences in FASTA format and prepare them for feature extraction. The primary goal of this stage is to transform raw amino acid sequences into machine-interpretable embeddings that capture the critical features governing biological activity [19]. Unlike standard OneHot encoding, which often suffers from low expression ability due to the loss of correlation between individual residues, our approach utilizes sequence-based descriptors to enhance the representation of protein features [33].

These descriptors—which we refer to as the protein 'fingerprint'—characterize the protein based on its functional propensities and structural packing patterns [34]. This sequence-level representation is subsequently fused with geometric features derived from the **Triangular Spatial Relationship** algorithm. The TSR component maps the 3D coordinates of atoms into a structural hierarchy, ensuring that the model captures spatial motifs that are otherwise omitted in purely sequence-based pipelines [35].

2.5 Fingerprint conversion from SMILES representation

To generate ligand fingerprints, researchers have applied path-based fingerprints, which identify paths between atoms to represent the structural features of chemical compounds. These types of fingerprints are specifically designed to pre-filter substructure searches. For example, Chemaxon Chemical Fingerprint is a path-based fingerprint that provides a representation of chemical structures. However, this method yields fast results by focusing on specific pathways and substructures. [36].

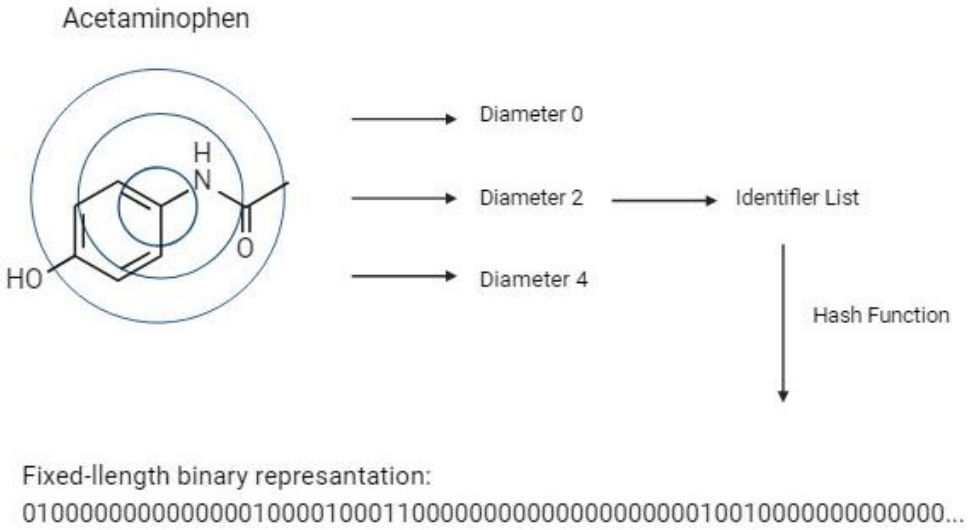

Figure 1. ECFP to fingerprint

Extended-Connectivity Fingerprint (ECFP) generally consists of binary or numeric vectors representing chemical motifs [37] and requires vectors between 1024 and 2048 bits in length as shown in Figure 1 [38]. This process was implemented using the RDKit library.

2.6 Dataset

The DAVIS and KIBA datasets play an essential role in evaluating proposed methods for new and potentially promising drug candidates and identifying their targets. These datasets are used in drug research and development to determine protein-drug interactions and predict affinity, and their features are presented in Table 1 [39]. The DAVIS dataset focuses on protein kinase inhibition, including the binding affinities of these kinases to small molecules. This dataset, published by Davis et al., contains the binding affinities (Kd) of 442 human kinases against 68 small-molecule compounds [38]. The dataset comprehensively maps interactions between kinases and inhibitors, providing the basis for drug development studies of kinase inhibitors. The KIBA dataset is another valuable resource for kinase inhibition. This dataset comprises interactions between various kinases and numerous ligands, which are determined by 'Kinase Inhibitor BioActivity' (KIBA) scores. KIBA scores are derived from biological activity measurements, such as EC50, IC50, Ki, and Kd, collected from various data sources.

Table 1. Specifications of datasets

	Protein	Compounds	Interactions
Davis	442	68	30056
KIBA	229	2111	118254

The KIBA dataset provides information on interactions for 2111 ligands against 229 kinases, which can be used to train machine learning models and enhance drug discovery processes [40]. To adapt binding affinity data for classification, GraPNet employs established thresholds, such as $pK_d \geq 7.0$ for the Davis dataset and a **KIBA score** ≥ 12.1 based on literature.

2.7 Geometric deep learning

GDL is an artificial intelligence method that enables the creation of deep neural network models in geometric areas beyond Euclidean space, such as graphs and manifolds. Unlike classical deep learning methods, it can model irregular data structures and relationships. Examples that exhibit non-Euclidean characteristics and are more consistent with graph structures include social networks created with user features [14], predicting protein structures, and detecting protein binding sites of disease agents [41]. For example, in social network graphs, users representing objects are nodes, and their relationships are edges.

One of the fundamental principles of GDL is to exploit geometric properties such as symmetries and invariances. Symmetries, defined as situations where data remain invariant under certain transformations, enable learning algorithms to be more general and flexible [42]. For example, CNNs used in image recognition applications enable more effective learning by leveraging translational and rotational symmetries in images. Similarly, deep learning models operating on graph data learn complex relationships between nodes and edges by leveraging graph topology. This suggests that applying GDL in bioinformatics yields high accuracy [18].

2.8 Proposed method

While traditional ML often relies on fixed feature vectors, GraPNet represents ligands as molecular graphs where node features (atoms) and edge features (bonds) are processed through Graph Convolutional Network layers. Node features taken as Atomic Number, Atom Degree, Formal Charge, Hybridization, Aromaticity, Hydrogen Count; edge features taken as Bond Type, Conjugation, Ring Membership. To powerfully evaluate node and edge

features, a Graph Convolutional Network (GCN) is generated using ligand properties. GraPNet constructs **independent graphs for each ligand** in the dataset rather than a global heterogeneous graph. The protein is not converted into a graph of residues but is instead processed as **TSR-based fingerprint** [35]. These features are concatenated with the global graph representation of the ligand (after graph pooling) to perform the final classification. GraPNet addresses this by using the Triangular Spatial Relationship algorithm, mapping the 3D coordinates of C_α atoms into a structural hierarchy [35]. This geometric encoding allows the model to learn spatial motifs that are often omitted in the feature extraction pipelines of standard regression models [35].

Researchers created graph representations from ligand and protein fingerprints using the Python programming language and relevant libraries for data processing and analysis. In these graphs, nodes represent ligands and proteins, while edges indicate their bonding properties. The PyTorch Geometric library and several other software and tools (RDKit, Scikit-learn, NetworkX) were used for the development and training of graph learning models. The model, built as a weighted graph using the TSR criterion, was trained on large datasets and became capable of predicting the presence of protein-ligand interactions.

Figure 2 illustrates the flowchart of the affine prediction system, which uses a combination of features. It shows that the data are converted to numerical formats, starting with SMILES sequences of ligands and FASTA sequences of proteins as input, and then label encoding. The embedding layer completes the first part by applying fingerprint transformations to the generated data. Convolution layers then process both SMILES and FASTA data. These layers extract and highlight specific features from the input data. After each convolution layer, the ReLU activation function enhances feature learning, and fully connected layers aggregate features. Finally, these features are fed into deep learning blocks. These blocks learn the topological relationships and complex interactions between protein and ligand molecules. The model is optimized and trained to predict the binding affinity using these features, yielding classification results. This method combines the strengths of both approaches, leveraging GDL for high-accuracy bioinformatics and integrating CNNs for feature extraction, thereby offering an efficient approach to analyzing protein-ligand interactions. The proposed model does not utilize a single global graph of all drugs and proteins. Instead, it constructs **per-pair representations** where the ligand is modeled as a molecular graph, and the protein is represented via a structured geometric fingerprint fused during the late stages of architecture.

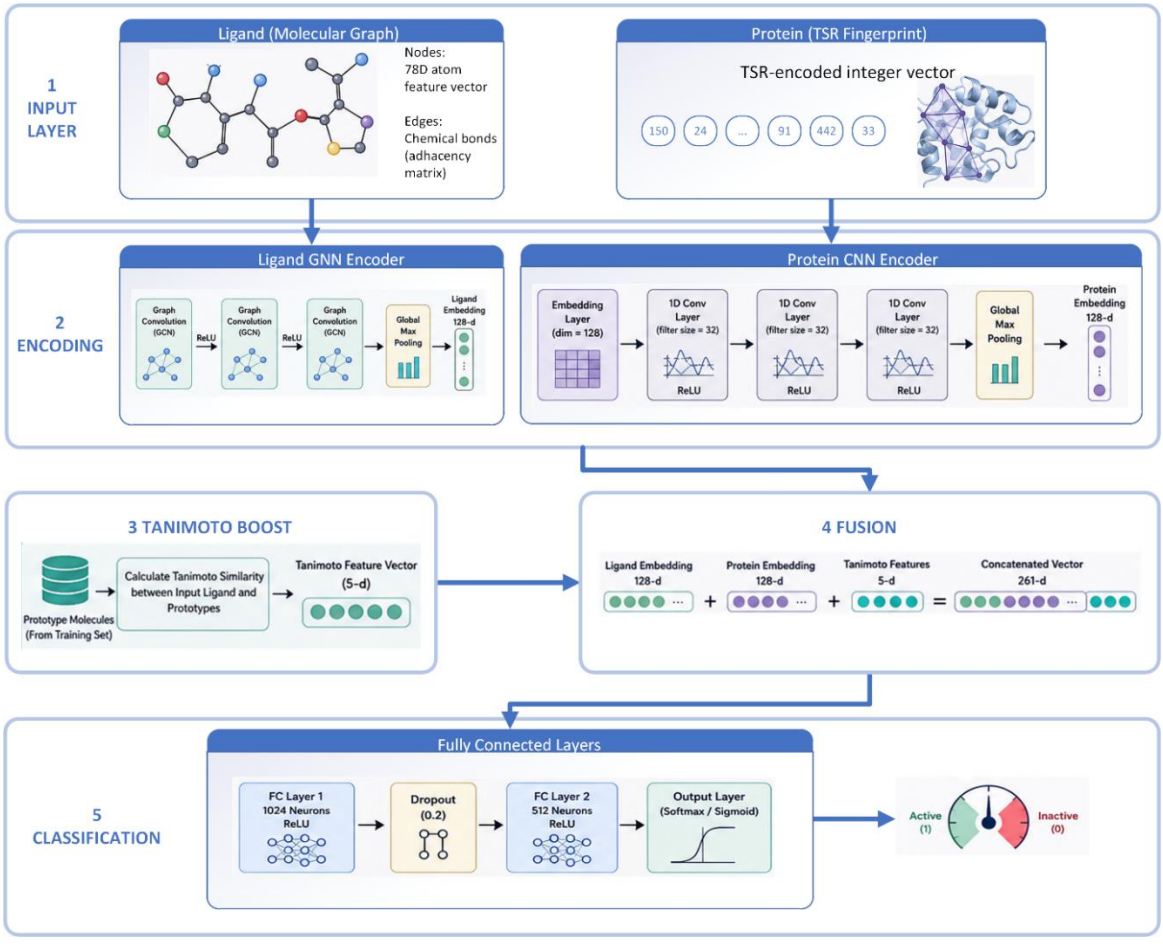

Figure 2. Flowchart of GraPNet

The increasing demand for healthcare services, in tandem with the growing number of diseases, has created a need for cost-effective, time-saving solutions. This approach aims to reduce the time and cost of the clinical phase of drug research and development by simulating it using artificial intelligence technologies enabled by computational sciences.

2.9 Model architecture

GraPNet consists of two primary feature extraction modules that are subsequently fused for classification:

- **Ligand Encoder:** This module processes SMILES-derived graph representations. It utilizes **Graph Convolutional Network** layers to extract topological features from the molecular graph $G = (V, E)$, where atoms are nodes and bonds are edges. The GCN operations follow the propagation rule $H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$, incorporating self-loops to preserve node identity [43].
- **Protein Encoder:** Protein sequences are represented using fingerprints based on the **Triangular Spatial Relationship** criterion [35]. This pathway utilizes an **Embedding layer** followed by **1D Convolutional Neural Network** layers to capture local and global sequence-structure motifs.
- **Tanimoto Booster:** An auxiliary module that calculates 5 additional similarity features by comparing ligand fingerprints against positive and negative prototypes in the training set, enhancing the model's discriminative power.
- **Fusion and Classification:** The outputs from both encoders and the booster are concatenated and passed through **fully connected (dense) layers** with **ReLU activation** functions. A final linear layer outputs logits, which are converted to interaction probabilities using a sigmoid function.

2.10 Hyperparameters and training configuration

GraPNet The following parameters as seen in Table 2 were utilized to achieve the reported 80.14% accuracy.

Table 2. Hyperparameters for proposed method

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	64
Epoch Count	20
Hidden Dimensions	128
Protein Max Length	1000 residues
Ligand Fingerprint	ECFP (1024–2048 bits)
Loss Function	BCEWithLogitsLoss
Class Balancing	Loss-weighting (pos_weight)
Cross-Validation	5-fold Stratified

To maintain parity with benchmark models like **DeepDTA** [44] and **DeepConv-DTI** [42], GraPNet utilizes a **5-fold stratified cross-validation** protocol. Baseline results were generated using the original authors' published code where available (e.g., DeepDTA and DeepConv-DTI GitHub repositories), ensuring that the hyperparameter optimization and split execution followed the original developers' specifications [42]. By applying these identical binarization rules to the benchmark datasets, the performance metrics (Accuracy, F1-score) are directly comparable to GraPNet's results.

One of the key contributions of this study is the incorporation of Tanimoto-based enhancement mechanisms into the proposed framework. In this approach, compatibility scores generated between proteins and two representative ligand candidates, namely active and inactive compounds, were integrated into the model as additional discriminative features. By enriching the learned representations with similarity-driven statistical information, the proposed enhancement aimed to improve the predictive performance of the model, particularly for borderline classification cases. To evaluate the effectiveness of this strategy, experiments were conducted under two different configurations: with and without the Tanimoto enhancement module. The corresponding results are presented in Table 3. According to the obtained findings, the proposed method achieved clear performance improvements in challenging classification scenarios when the Tanimoto-based enhancement was incorporated.

Table 3. Ablation study results

		Accuracy	f1-Score
With Boost	Tanimoto	0.80	0.80
Without Boost	Tanimoto	0.78	0.79

Another important aspect of the study concerns the effects of batch size and epoch number on computational efficiency and model performance. Experimental evaluations were performed using batch sizes of 16, 32, and 64. Although larger batch sizes provided slight improvements in performance metrics for this specific application, they also significantly increased the computational burden. Considering the balance between predictive performance and computational cost, the highest batch size was preferred, as it did not lead to prohibitively high processing requirements for the proposed method. Similarly, epoch values of 20, 50, and 100 were investigated. The experimental results demonstrated that increasing the number of epochs did not produce any meaningful improvement in predictive scores, while substantially extending the training time. Therefore, the lowest epoch value, namely 20, was selected as the optimal configuration in terms of computational efficiency and model stability. Additionally, all scores were gathered randomly subsampled dataset for class balancing. Without class balancing average accuracy scores gathered as 88.52% but f-scores 62.99% as expected.

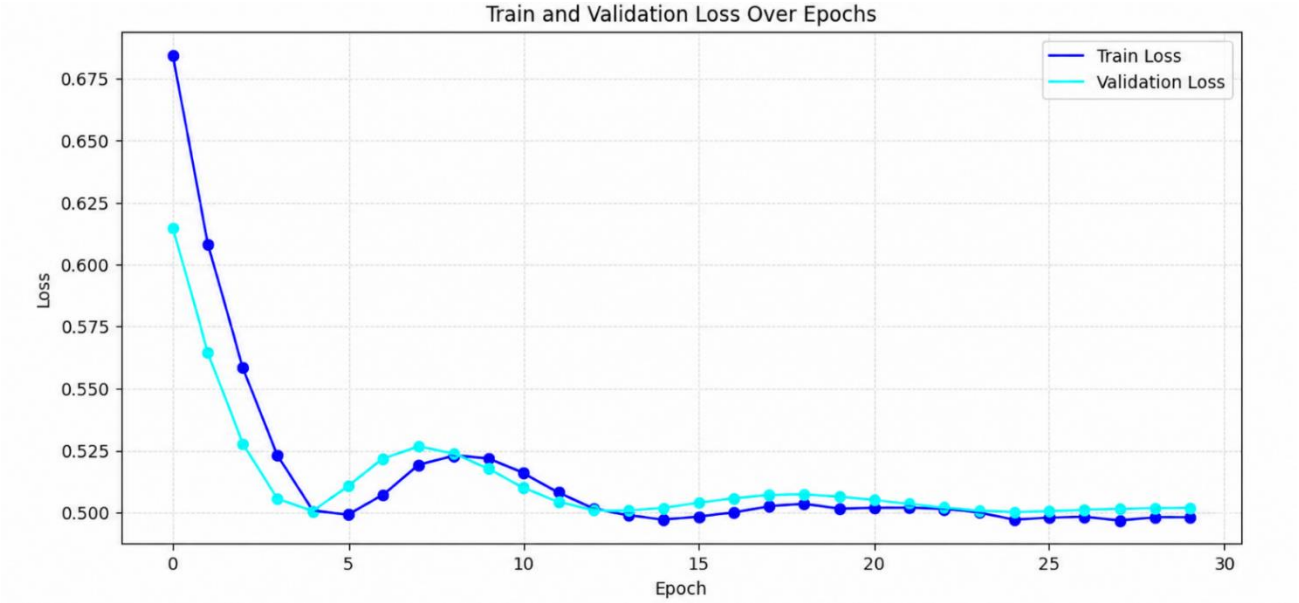

Figure 3. Train and validation loss over epoch

3 Results and discussion

In this study, we employed a graph structure to extract features from fingerprint information generated based on protein and ligand sequences and to evaluate their interactions on Intel i9-13900HX, 32GB memory and NVIDIA GeForce RTX 4060 Laptop GPU with 8GB dedicated GPU memory environment. The results showed the effectiveness of this method in predicting protein-ligand affinity. It achieved high accuracy, enabling the correct identification of active pairs of molecules with high affinity. Analyses using TSR show superior performance and increased model accuracy. This method played a crucial role in predicting protein-ligand interactions by evaluating the structural similarities [45]. The **TSR algorithm** specifically constructs geometric triangles using C_α atoms as vertices, mapping these to unique integer keys that represent the protein's structural hierarchy [35]. For ligand representation, the **RDKit library** was used to generate canonical SMILES and path-based fingerprints. Training and validation were conducted using the **PyTorch Geometric** and **Scikit-learn** frameworks.

Figure 3 illustrates the changes in the model's training and validation losses across both datasets during training. The training and validation losses are initially high but decrease over time; this indicates that the model is adapting to the training data, and the learning process is proceeding efficiently. In the first few epochs, the training loss decreases faster than the validation loss. This indicates that the model initially learns quickly, after which the losses stabilize. After the 20th epoch, the losses converge, indicating that the model performs well on both training and validation data.

This convergence indicates that the model generalizes well during training and does not overfit. The training and validation losses are similar, suggesting that the model performs consistently across both datasets, which contributes to its high overall validity. These results suggest that even if the model is trained for more epochs, it is unlikely to lead to overfitting, and the current training period is optimal for model performance.

In general, the methods used in this study, which were run 10 times and achieved an average accuracy of 80.14%, yielded successful results in detecting protein-ligand interactions and offered potential approaches for drug discovery.

The regression approach of Gök et al. is highly valuable for lead optimization, where distinguishing between subtle affinity differences is necessary [46]. Conversely, GraPNet's classification model is optimized for the rapid prioritization of active compounds from vast chemical libraries.

Table 4. Specifications of datasets

Model	Drug Representation	Protein Representation	Core Innovation
GraphDTA [47]	GNN	1D-CNN	Graph-based drug modeling [47].
DGraphDTA [43]	GNN	GNN	Unified GNN for both entities [43].
MGraphDTA [48]	Multiscale GNN	3-branch CNN	27-layer multiscale blocks [49].
Proposed Method	Graph Learning	TSR Criterion	

In the dNNDR article, which aimed to predict protein-drug interactions, three different models were developed with distinct approaches: dNNDRa, dNNDRb, and dNNDRc. The integration of methods such as graph learning has improved the accuracy of protein-ligand interaction detection and enhanced drug discovery [19]. These approaches have broad applications in modern biotechnology and bioinformatics, providing a solid foundation for future research as seen in Table 4.

GraphDTI uses protein sequence descriptors and drug Morgan fingerprints as inputs to the training model [45]. Deep convolutional neural networks, such as FRnet-DTI,

DeepConv-DTI, and DeepDTA, demonstrate effectiveness in modeling complex relationships in biological systems [50], [42], [44]. All comparisons are presented in Table 5.

Table 5. Accuracy and F1-score of the models compared.

Method	Accuracy	F1-Score
Sequence-based	0.69	0.64
Ctd	0.19	0.06
Deepdta	0.70	0.68
Deepconv-dtı	0.72	0.64
GraPNet	0.80	0.71
Dnndrb	0.76	0.66
Dnndra	0.47	0.51

3.1 Comparative advantages

The integration of GNNs with the TSR criterion provides several technical advantages over existing models:

- **Structural Fidelity:** Unlike **GraphDTA**, which primarily treats proteins as 1D sequences [33], the proposed method utilizes the **TSR criterion** to map 3D geometric motifs into a structural hierarchy [35]. This allows the model to capture spatial relationships between residues that are not evident in primary sequences [35].
- **Scalable Complexity:** Models like **MGraphDTA** involve significant computational overhead due to very deep GNNs [51]. In contrast, the proposed use of the **Tanimoto Booster** and graph learning aims to achieve high discriminative power by combining topological features with statistical similarity, potentially offering a more efficient alternative to hyper-deep architectures [52].
- **Global vs. Local Learning:** While GNN models excel at learning local chemical motifs, they can struggle to capture global graph patterns effectively [51]. The proposed hybrid approach combines local graph convolutions for ligands with hierarchical geometric fingerprints for proteins to bridge this information gap [35].

4 Conclusion

The DAVIS and KIBA datasets, which contain experimentally validated protein-ligand interactions, were used to train and test the models. The data were partitioned into training and test sets, and graph-based learning models were trained on them. Model performance was assessed using metrics such as accuracy, sensitivity, and specificity. Furthermore, the role and effectiveness of the models in detecting target-site recognition interactions were analyzed. To ensure a fair comparison, studies using similar datasets were included. Successful results were obtained from reference datasets. For evaluation, the dataset was split into two classes: active and inactive. In this context, a graph-learning-based protein-ligand interaction classification method was developed, achieving 80.96% accuracy and an F1-score of 80.97%. These results show the potential of graph learning approaches for detecting protein-ligand interactions. This accuracy rate demonstrates competitive performance when compared to traditional and modern methods used in similar studies.

The findings of this study demonstrate that integrating fingerprint transformation, TSR, and graph learning techniques provides significant advantages for understanding and classifying protein–ligand interactions. GDL is an effective tool for revealing connections and relationships within data, especially when biological data and big data can be naturally expressed as graph structures [53, 54]. It is predicted that using tertiary structures instead of fingerprint transformations as a dataset will yield better results. In future studies, this method can guide similar studies on other types of biomolecular interactions.

CRediT authorship contribution statement

Turgay Batbat: Conceptualization, Writing–review and editing, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation; **Seda Mürütsoy:** Writing–original draft, Visualization, Formal analysis, Methodology, Validation, Software.

Acknowledgement

This work was conducted within the scope of the master's thesis (Thesis No. 895936).

Conflict of Interest

The authors declare that there is no conflict of interest

Similarity Ratio (iThenticate): 14%

Declaration of generative artificial intelligence (AI) and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] N. Singh, P. Vayer, S. Tanwar, J. Poyet, K. Tsaioun and B. O. Villoutreix, "Drug discovery and development: introduction to the general public and patient groups," *Frontiers in Drug Discovery,* vol. 3, 2023. https://doi.org/10.3389/fddsv.2023.1201419

[2] J. A. DiMasi, R. W. Hansen and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics,* vol. 22, no. 2, pp. 151-185, 2003. https://doi.org/10.1016/s0167-6296(02)00126-1

[3] A. B. Deore, J. R. Dhumane, R. Wagh and R. Sonawane, "The Stages of Drug Discovery and Development Process," *Asian Journal of Pharmaceutical Research and Development,* vol. 7, no. 6, pp. 62-67, 2019. https://doi.org/10.22270/ajprd.v7i6.616

[4] D. R. Flower, "Drug Discovery: Today and Tomorrow," *Bioinformation,* vol. 16, no. 1, pp. 1-3, 2020. https://doi.org/10.6026/97320630016001

[5] A. C. Pushkaran and A. A. Arabi, "From understanding disease s to drug design: can artificial intelligence bridge the gap?," *Artificial Intelligence Review,* vol. 57, no. 4, 2024. https://doi.org/10.1007/s10462-024-10714-5

[6] T. C. Walther and M. Mann, *Mass spectrometry–based proteomics in cell biology,* vol. 190, Rockefeller University Press, 2010, pp. 491-500. https://doi.org/10.1083/jcb.201004052

[7] W. R. Pearson, "Finding Protein and Nucleotide Similarities with FASTA," *Current Protocols in Bioinformatics,* vol. 53, no. 1, 2016. https://doi.org/10.1002/0471250953.bi0309s53

[8] R. Özçelik, H. Öztürk, A. Özgür and E. Özkırımlı, "ChemBoost: A Chemical Language Based Approach for Protein – Ligand Binding Affinity Prediction," *Molecular Informatics,* vol. 40, no. 5, 2020. https://doi.org/10.1002/minf.202000212

[9] G. A. Abdelkader and J-D. Kim, "Advances in Protein-Ligand Binding Affinity Prediction via Deep Learning: A Comprehensive Study of Datasets, Data Preprocessing Techniques, and Model Architectures," *Current Drug Targets,* 2024. https://doi.org/10.2174/0113894501330963240905083020

[10] H. Li, K. Leung, M.-H. Wong and P. J. Ballester, "Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study," *BMC Bioinformatics,* vol. 15, no. 1, pp. 291-291, 2014. https://doi.org/10.1186/1471-2105-15-291

[11] M. Jiménez-Rosés, B. A. Morgan, M. J. Sigstad, T. D. Z. Tran, R. Srivastava, A. Bunsuz, L. Borrega-Román, P. Hompluem, S. A. Cullum, C. R. Harwood, E. J. Koers, D. A. Sykes, I. B. Styles and D. B. Veprintsev, "Combined docking and machine learning identify key molecular determinants of ligand pharmacological activity on β2 adrenoceptor," *Pharmacology Research & Perspectives,* vol. 10, no. 5, 2022. https://doi.org/10.1002/prp2.994

[12] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Q. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal Of Big Data,* vol. 8, no. 1, pp. 53-53, 2021. https://doi.org/10.1186/s40537-021-00444-8

[13] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri and D. R. Koes, "Protein–Ligand Scoring with Convolutional Neural Networks," *J. Chem. Inf. Model.,* vol. 57, no. 4, pp. 942–957, 2017. https://doi.org/10.1021/acs.jcim.6b00740

[14] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and P. Vandergheynst, "Geometric Deep Learning: Going beyond Euclidean data," *IEEE Signal Processing Magazine,* vol. 34, no. 4, pp. 18-42, 2017. https://doi.org/10.1109/msp.2017.2693418

[15] C. Shen, J. Luo and K. Xia, "Molecular geometric deep learning," *Cell Reports Methods,* vol. 3, no. 11, pp. 100621-100621, 2023. https://doi.org/10.1016/j.crmeth.2023.100621

[16] T. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv (Cornell University),* 2016. https://doi.org/10.48550/arxiv.1609.02907

[17] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. Müller and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Communications,* vol. 8, no. 1, pp. 13890-13890, 2017. https://doi.org/10.1038/ncomms13890

[18] T. Gaudelet, B. Day, A. R. Jamasb, J. Soman, C. Regep, G. Liu, J. B. R. Hayter, R. Vickers, C. S. Roberts, J. Tang, D. Roblin, T. L. Blundell, M. M. Bronstein and J. P. Taylor-King, "Utilizing graph machine learning within drug discovery and development," 2021. https://doi.org/10.1093/bib/bbab159

[19] Y. Kalakoti, S. Yadav and D. Sundar, "Deep Neural Network-Assisted Drug Recommendation Systems for Identifying Potential Drug–Target Interactions," *ACS Omega,* vol. 7, no. 14, pp. 12138–12146,2022. https://doi.org/10.1021/acsomega.2c00424

[20] A. Dhakal, C. McKay, J. J. Tanner and J. Cheng, "Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions," *Briefings in Bioinformatics,* vol. 23, no. 1, 2021. https://doi.org/10.1093/bib/bbab476

[21] G. Sudhakararao, K. A. Priyadarsini, G. Kiran, P. Karunakar and K. Chegu, "Physiological Role of Proteins and their Functions in Human Body," *International Journal of Pharma Research and Health Sciences,* vol. 7, no. 1, pp. 2874-78, 2019. https://doi.org/10.21276/ijprhs.2019.01.02

[22] T. Batbat and C. Öztürk, "Ayrık Yapay Arı Kolonisi Algoritması ile Protein Yapısı Tahmini", *Bilişim Teknolojileri Dergisi*, vol. 9, no 3, pp. 263, Eyl. 2016. https://doi.org/10.17671/btd.97757

[23] M. Graille, S. Sacquin-Mora and A. Taly, "Best Practices of Using AI-Based Models in Crystallography and Their Impact in Structural Biology," *J. Chem. Inf. Model.,* vol. 63, no. 12, pp. 3637–3646, 2023. https://doi.org/10.1021/acs.jcim.3c00381

[24] H. Gohlke and G. Klebe, "Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors," *Angewandte Chemie International Edition,* vol. 41, no. 15, pp. 2644-2676, 2002.

https://doi.org/10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O

[25] H. S. Lee and W. Im, "Stalis: A Computational Method for Template-Based Ab Initio Ligand Design," *Journal of Computational Chemistry,* vol. 40, no. 17, pp. 1622-1632, 2019. https://doi.org/10.1002/jcc.25813

[26] A. Drefahl, "CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures," *Journal of Cheminformatics,* vol. 3, no. 1, pp. 1-1, 2011. https://doi.org/10.1186/1758-2946-3-1

[27] Y. Liu, L. Xing, L. Zhang, H. Cai and M. Guo, "GEFormerDTA: drug target affinity prediction based on transformer graph for early fusion," *Scientific Reports,* vol. 14, no. 1, 2024. https://doi.org/10.1038/s41598-024-57879-1

[28] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan and M. Grohe, "Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 4602–4609, Jul. 2019, 2019. https://doi.org/10.1609/aaai.v33i01.33014602

[29] D. M. Chen, G. Duan, D. Miao, X. Zheng and Y. Zhu, "InvarNet: Molecular property prediction via rotation invariant graph neural networks," *Machine Learning with Applications,* vol. 18, pp. 100587-100587, 2024. https://doi.org/10.1016/j.mlwa.2024.100587

[30] D. Bajusz, A. Rácz and K. Héberger, "Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching," %1 içinde *Elsevier eBooks*, Elsevier BV, 2017, pp. 329-378. https://doi.org/10.1016/b978-0-12-409547-2.12345-5

[31] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences,* vol. 28, no. 1, pp. 31-36, 1988. https://doi.org/10.1021/ci00057a005

[32] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzábal and A. Valencia, *Information Retrieval and Text Mining Technologies for Chemistry,* vol. 117, American Chemical Society, 2017, pp. 7673-7761. https://doi.org/10.1021/acs.chemrev.6b00851

[33] X. Wang, Y. Liu, L. Fan, H. Li, P. Gao and D. Wei, "Dipeptide Frequency of Word Frequency and Graph Convolutional Networks for DTA Prediction," *Frontiers in Bioengineering and Biotechnology,* vol. 8, 2020. https://doi.org/10.3389/fbioe.2020.00267

[34] D. Bandyopadhyay, J. Huan, J. Liu, J. F. Prins, J. Snoeyink, W. Wang and A. Tropsha, "Structure-based function inference using protein family-specific fingerprints," *Protein Science,* vol. 15, no. 6, pp. 1537-1543, 2006. https://doi.org/10.1110/ps.062189906

[35] S. Kondra, F. Chen, Y. Chen, Y. Chen, C. J. Collette and W. Xu, "A study of a hierarchical structure of proteins and ligand binding sites of receptors using the triangular spatial relationship-based structure comparison method and development of a size-filtering feature designed for comparing different sizes of protein structures," *Proteins Structure Function and Bioinformatics,* vol. 90, no. 1, pp. 239-257, 2021. https://doi.org/10.1002/prot.26215

[36] A. Gimeno, M. J. Ojeda, S. Tomas-Hernández, A. Cereto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas and S. García-Vallvé, "The Light and Dark Sides of Virtual Screening: What Is There to Know?," *Int. J. Mol. Sci.*, vol. 20, no. 6, 2019. https://doi.org/10.3390/ijms20061375

[37] T. Pahikkala, A. Airola, S. Pietilä, S. K. Shakyawar, A. Szwajda, J. Tang and T. Aittokallio, "Toward more realistic drug-target interaction predictions," *Briefings in Bioinformatics,* vol. 16, no. 2, pp. 325-337, 2014. https://doi.org/10.1093/bib/bbu010

[38] M. I. Davis, J. P. Hunt, S. Herrgård, P. Ciceri, L. Wodicka, G. Pallares, M. Höcker, D. K. Treiber and P. P. Zarrinkar, "Comprehensive analysis of kinase inhibitor selectivity," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, 2011. https://doi.org/10.1038/nbt.1990

[39] Z. Chu, F. Huang, H. Fu, Q. Yuan, X. Zhou, S. Liu and W. Zhang, "Hierarchical graph representation learning for the prediction of drug-target binding affinity," *Information Sciences,* vol. 613, pp. 507-523, 2022. https://doi.org/10.1016/j.ins.2022.09.043

[40] J. Tang, A. Szwajda, S. K. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg and T. Aittokallio, "Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis," *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, 2014. https://doi.org/10.1021/ci400709d

[41] D. Bajusz, A. Rácz and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics,* vol. 7, no. 1, pp. 20-20, 2015. https://doi.org/10.1186/s13321-015-0069-3

[42] I. Lee, J. Keum and H. Nam, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences," *PLoS Comput. Biol.*, vol. 15, no. 6, pp. 1–21, 2019. https://doi.org/10.1371/journal.pcbi.1007129

[43] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan and Z. Wei, "Drug–target affinity prediction using graph neural network and contact maps," *RSC Advances,* vol. 10, no. 35, pp. 20701-20712, 2020. https://doi.org/10.1039/d0ra02297g

[44] H. Öztürk, A. Özgür and E. Özkırımlı, "DeepDTA: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018. https://doi.org/10.1093/bioinformatics/bty593

[45] G. Liu, M. Singha, L. Pu, P. Neupane, J. Feinstein, H. Wu, J. Ramanujam and M. Bryliński, "GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data," *Journal of Cheminformatics,* vol. 13, no. 1, pp. 58-58, 2021. https://doi.org/10.1186/s13321-021-00540-0

[46] M. Gök, E. Cengiz and M. M. Mijwil, "Unveiling Protein-Ligand Interactions: Regression Methods for Binding Affinity Prediction," *International Journal on Engineering Applications (IREA),* vol. 12, no. 6, pp. 508-508, 2024. https://doi.org/10.15866/irea.v12i6.25407

[47] T. Nguyen , H. Le, T. P. Quinn, T. Nguyen, T. D. Le and S. Venkatesh, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2021. https://doi.org/10.1093/bioinformatics/btaa921

[48] Z. Yang, W. Zhong, L. Zhao and C. Y. Chen, "MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction," *Chemical Science,* vol. 13, no. 3, pp. 816-833, 2022. https://doi.org/10.1039/d1sc05180f

[49] Q. Lv, G. Chen, H. He, Z. Yang, L. Zhao, H. Chen and C. Y. Chen, "TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining," *Chemical Science,* vol. 14, no. 39, pp. 10684-10701, 2023. https://doi.org/10.1039/d3sc02139d

[50] F. Rayhan, S. Ahmed, Z. Mousavian, D. M. Farid and S. Shatabda, "FRnet-DTI: Deep convolutional neural network for drug-target interaction prediction," *Heliyon,* vol. 6, no. 3, 2020. https://doi.org/10.1016/j.heliyon.2020.e03444

[51] M. Kalemati, M. Z. Emani and S. Koohi, "BiComp-DTA: Drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach," *PLoS Computational Biology,* vol. 19, no. 3, 2023. https://doi.org/10.1371/journal.pcbi.1011036

[52] M. A. Thafar, M. Alshahrani, S. Albaradei, T. Gojobori, M. Essack and X. Gao, "Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning," *Scientific Reports,* 2022. https://doi.org/10.1038/s41598-022-08787-9

[53] H.-C. Yi, Z. You, D.-S. Huang and C. K. Kwoh, "Graph representation learning in bioinformatics: trends, methods and applications," *Briefings in Bioinformatics,* vol. 23, no. 1, 2021. https://doi.org/10.1093/bib/bbab340

[54] X. Zhang, L. Li, L. Liu and M. Tang, "Graph Neural Networks and Their Current Applications in Bioinformatics," *Frontiers in Genetics,* vol. 12, 2021. https://doi.org/10.3389/fgene.2021.690049