Send Orders for Reprints to reprints@benthamscience.net

Current Drug Targets, 2024, *25*, 1041-1065

REVIEW ARTICLE

Advances in Protein-Ligand Binding Affinity Prediction *via* Deep Learning: A Comprehensive Study of Datasets, Data Preprocessing Techniques, and Model Architectures

Gelany Aly Abdelkader[1] and Jeong-Dong Kim[1,2,3,*]

[1]*Department of Computer Science and Electronic Engineering, Sun Moon University, Asan 31460, Republic of Korea;* [2]*Division of Computer Science and Engineering, Sun Moon University, Asan 31460, Republic of Korea;* [3]*Genome-based BioIT Convergence Institute, Sun Moon University, Asan 31460, Korea*

© 2024 The Author(s). Published by Bentham Science Publisher. This is an open access article published under CC BY 4.0 https://creativecommons.org/licenses/by/4.0/legalcode

ARTICLE HISTORY

Received: June 07, 2024
Revised: August 11, 2024
Accepted: August 19, 2024

DOI:
10.2174/0113894501330963240905083020

Abstract: ***Background*:** Drug discovery is a complex and expensive procedure involving several timely and costly phases through which new potential pharmaceutical compounds must pass to get approved. One of these critical steps is the identification and optimization of lead compounds, which has been made more accessible by the introduction of computational methods, including deep learning (DL) techniques. Diverse DL model architectures have been put forward to learn the vast landscape of interaction between proteins and ligands and predict their affinity, helping in the identification of lead compounds.

***Objective*:** This survey fills a gap in previous research by comprehensively analyzing the most commonly used datasets and discussing their quality and limitations. It also offers a comprehensive classification of the most recent DL methods in the context of protein-ligand binding affinity prediction (BAP), providing a fresh perspective on this evolving field.

***Methods*:** We thoroughly examine commonly used datasets for BAP and their inherent characteristics. Our exploration extends to various preprocessing steps and DL techniques, including graph neural networks, convolutional neural networks, and transformers, which are found in the literature. We conducted extensive literature research to ensure that the most recent deep learning approaches for BAP were included by the time of writing this manuscript.

***Results*:** The systematic approach used for the present study highlighted inherent challenges to BAP *via* DL, such as data quality, model interpretability, and explainability, and proposed considerations for future research directions. We present valuable insights to accelerate the development of more effective and reliable DL models for BAP within the research community.

***Conclusion*:** The present study can considerably enhance future research on predicting affinity between protein and ligand molecules, hence further improving the overall drug development process.

Keywords: Deep learning, protein-ligand binding affinity, compound-protein interaction, drug discovery, drug repurposing, DNA sequences.

1. INTRODUCTION

The concept of drug repurposing has emerged as a promising strategy in modern drug discovery, offering a shortcut to identifying new therapeutic uses for existing drugs. Computational methods enable systematic exploration of the vast landscape of molecular interactions, facilitating the repurposing of approved drugs for novel indications [1]. This approach circumvents many of the challenges associated with traditional drug discovery and accelerates the identification of lead compounds with therapeutic potential.

Exploring compound-protein interactions (CPIs) is fundamental to drug discovery and repurposing, as they elucidate the mechanisms of action and therapeutic efficacy of pharmaceutical agents. DL methods play an essential role in elucidating these interactions, covering various aspects such as binding site prediction, molecular docking, binding interaction prediction framed as binary classification, and binding affinity prediction.

With binding site prediction algorithms, one can identify the regions on target proteins where small molecules (ligands) are likely to bind, providing valuable insights into potential drug-target interactions [2, 3]. On the other hand, molecular docking algorithms further facilitate the exploration of CPI by simulating the binding process and predicting the op-

*Address correspondence to this author at the Department of Computer Science and Electronic Engineering, Sun Moon University, Asan 31460, Republic of Korea; E-mail: kjd4u@sunmoon.ac.kr

Current Drug Targets

timal orientation and conformation of ligands within binding sites [4]. BAP algorithms play a crucial role in quantifying the strength of interactions between compounds and target proteins, enabling the prioritization of lead compounds based on their potential efficacy. Additionally, binding interaction prediction, framed as binary classification, allows researchers to distinguish between active and inactive compounds based on their propensity to interact with target proteins [5, 6].

The availability of experimentally measured data on binding interactions has spurred the proliferation of numerous DL methods. The authors across the field propose diverse feature representations and architectures for feature learning, ranging from convolutional neural networks (CNNs) models to recurrent neural networks (RNNs), graph neural networks (GNNs), and even transformers. The diversity reflects the adaptability of DL methodologies in comprehensively capturing the complexities inherent in CPI.

CPI prediction has often been framed as a binary classification. Although this simplification has aided in developing predictive models, it has limitations. Binary classification tends to oversimplify the diverse nature of binding interactions, neglecting the varying degrees of affinity between compounds and proteins. Instead of recognizing a continuum of affinities, where interactions can range from weak to strong, binary classification tends to focus on a rigid categorization of interactions as binding or nonbinding. The approach overlooks the nuanced and varied strengths of molecular interactions inherent in biological systems.

In contrast, framing CPI prediction as a regression task to predict binding affinity offers a more informative perspective. Regression models provide valuable insights into the potency and efficacy of potential drug candidates by quantifying the strength of interactions between compounds and proteins. This nuanced understanding is crucial for prioritizing lead compounds and effectively guiding drug discovery. Therefore, our focus is specifically on studies that employ binding affinity as the target variable for predictive modelling in CPI.

1.1. Motivation

Numerous reviews have explored various aspects of protein-ligand binding affinity prediction, yet a notable gap persists in the examination of DL methodologies. For instance, Zhang *et al.* [7] concentrated solely on graph neural networks for BAP, while Bagherian *et al.* [8] focused their study on machine learning (ML) methods. Similarly, D. Wang *et al.* [9] focused on free energy-based simulations and ML-based scoring functions in their analysis. Although Wang *et al.* [10] delved into DL, they lacked comprehensive discussions on various data representations and commonly used datasets for BAP, with limited exploration of DL-based methods. Additionally, Meli *et al.* [11] predominantly focused on structure-based DL techniques, omitting the most up-to-date models.

Our aim is to bridge the gap by thoroughly examining commonly used datasets, discussing their quality, and offering an in-depth analysis of the most up-to-date DL methods, challenges, and future directions within this context. By addressing these crucial aspects, we aspire to provide a nuanced understanding of the role of DL in BAP, highlighting its potential implications for drug discovery and drug repurposing.

1.2. Contributions

The primary contributions of the study can be outlined as follows:

• In-depth data analysis and quality evaluation for informed dataset selection:

In contrast to previous surveys, we provide a richer data analysis and quality evaluation of the most used datasets in the BAP. By offering an in-depth assessment of dataset characteristics and quality, the paper guides the selection of adequate datasets for their studies.

• Comprehensive analysis of DL methods, with a focus on up-to-date methods:

Given the limited coverage of DL approaches in previous studies, we thoroughly analyzed various DL methods employed in BAP, notably by covering the most up-to-date DL methods from the literature and offering the most detailed classification compared to previous surveys. This comprehensive examination enhances our understanding of the efficacy of DL models in BAP.

• Identification of future directions for research and innovation:

We identify critical areas for further research and innovation by discussing the challenges and future directions in DL for BAP. By highlighting emerging trends and potential avenues for exploration, our survey guides researchers and practitioners toward advancing drug discovery and repurposing initiatives using state-of-the-art DL techniques.

As outlined in Fig. (**1**), the subsequent sections of this article are organized as follows. Section 2 provides a brief yet comprehensive background to foster a clear understanding of the topics at hand. In Section 3, we delve into an extensive data analysis and quality evaluation of the predominant datasets in the BAP. Section 4 discusses common data preprocessing methods, and Section 5 presents a comprehensive analysis of DL methods employed in BAP, detailing various methods and their limitations. In Section 6, we discuss challenges and directions for future works, highlighting critical areas for further exploration and innovation in DL for BAP.

2. BACKGROUND

The drug discovery process unfolds through several distinct stages, each marked by challenges and complexities. It typically begins with target identification and validation, wherein the molecular mechanisms underlying a disease or pathological condition are studied [12]. These potential targets encompass a diverse range of biomolecules, including

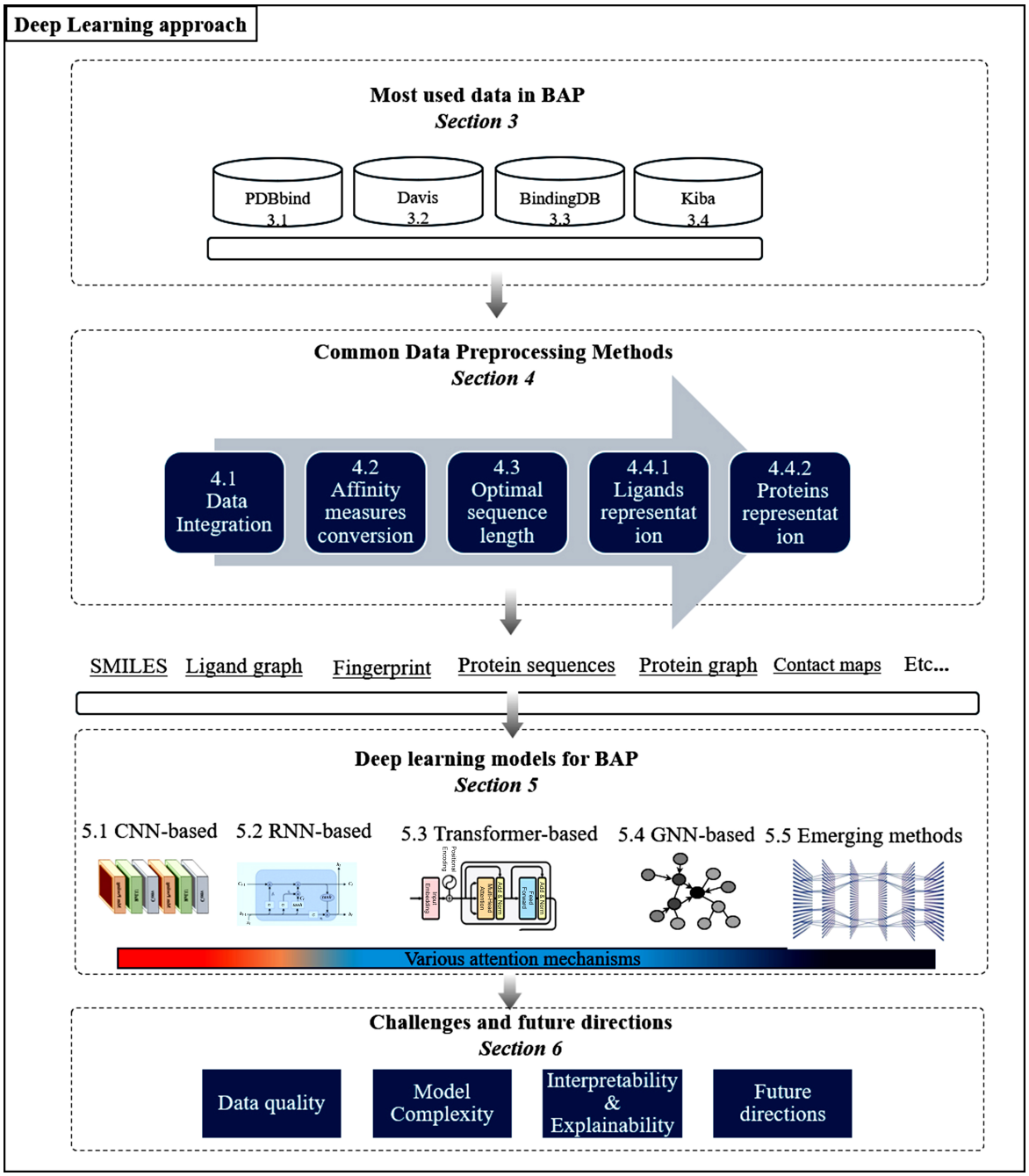

Fig. (1). Proposed survey's structure. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

DNA sequences, RNA molecules, proteins, and metabolites, which hold promise for therapeutic intervention. Once a target has been identified, the search for lead compounds commences, often through high-throughput screening of chemical libraries or rational drug design strategies.

After promising compound identification, a meticulous evaluation process ensues, subjecting each identified compound to rigorous scrutiny concerning its efficacy, safety profile, and pharmacological properties. Following this thorough evaluation, the selected compound advances to preclinical research, undergoing rigorous testing in both *in vitro* and *in vivo* settings.

This phase allows for a comprehensive assessment of the potential impact of compounds on physiological systems, laying the groundwork for subsequent clinical trials and regulatory approval processes. However, as outlined in Fig. (**2**), this traditional approach to drug discovery has some limitations. It is time-consuming, labor-intensive, and prohibitively costly, requiring extensive resources and years of painstaking effort [13, 14].

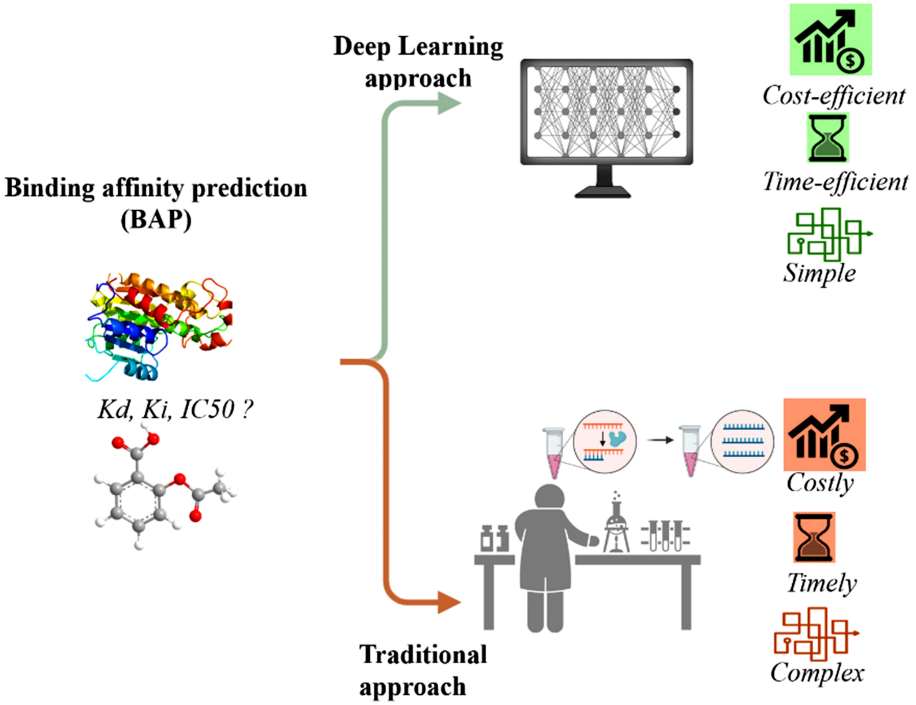

Fig. (2). Deep learning *versus* traditional approach for BAP. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3. MOST USED DATA IN BAP

Recent advancements in high-throughput screening, open data initiatives, and sequencing technologies have enabled large-scale experiments for compound and gene characterization [15-20]. These experiments have led to the identification of novel CPI and contributed to the rapid growth of CPI databases. However, managing and curating these abundant data remains an ongoing challenge in computational biology [21]. For instance, it can be difficult to identify the most relevant and reliable CPI data for training a DL model. Additionally, the quality of CPI data can vary widely, making it challenging to evaluate their reliability.

Numerous databases and repositories for CPI data are available and can be categorized into protein-centric, compound-centric, and compound-protein-centric databases [22]. Protein-centric databases primarily offer functional and structural information about proteins. The most widely utilized protein-centric databases are the Universal Protein Resource Knowledgebase (UniProtKB) and the Protein Data Bank (PDB). UniProtKB is divided into UniProtKB/Swiss-Prot and UniProtKB/TrEMBL [23]. UniProtKB/Swiss-Prot maintains manually annotated and experimentally verified protein sequences, while UniProtKB/TrEMBL contains computationally analyzed sequences awaiting manual annotation.

As of November 2023, UniProtKB/Swiss-Prot comprises 570 420 meticulously curated sequence entries sourced from 295 467 unique references with a total of 206 321 560 amino acids (UniProtKB/Swiss-Prot release 2023_05). Meanwhile, UniProtKB/TrEMBL featured 251 131 639 sequence entries, comprising 88 223 298 202 amino acids (UniProtKB/TrEMBL release 2023_05). In comparison, the PDB currently contains 211,377 entries corresponding to experimentally determined three-dimensional structures of biological macromolecules, primarily proteins and nucleic acids [24].

On the other hand, compound-centric databases offer comprehensive data about chemical compounds, including their chemical structures, properties, interactions, and bioactivity. Two notable representative compound-centric databases are DrugBank [25] and ChEMBL [26].

Our focus is on compound-protein-centric databases, which provide information on both proteins and compounds and their interactions. The most commonly used compound-protein-centric databases for BAP in the literature are the PDBbind [27] (including PDBbind_core_2013 [28-31] and PDBbind_core_2016 [32]), DAVIS [28], BindingDB [29], and Kiba [30]. The analysis of these data in the following section aims to assist researchers in making well-informed choices when selecting suitable datasets for training and testing DL models designed for BAP tasks.

3.1. PDBbind

The protein molecular weight within the general set spans from 3494 Daltons (Da) to a maximum of 423,549 Da (Figs. **3** and **4**). In contrast, the molecular weights of the compounds range from 57 Da to 3046 Da (Fig. **3**). Outliers for protein molecular weight exceed 75,866 Da, while the upper bound for compound weight is 877 Da.

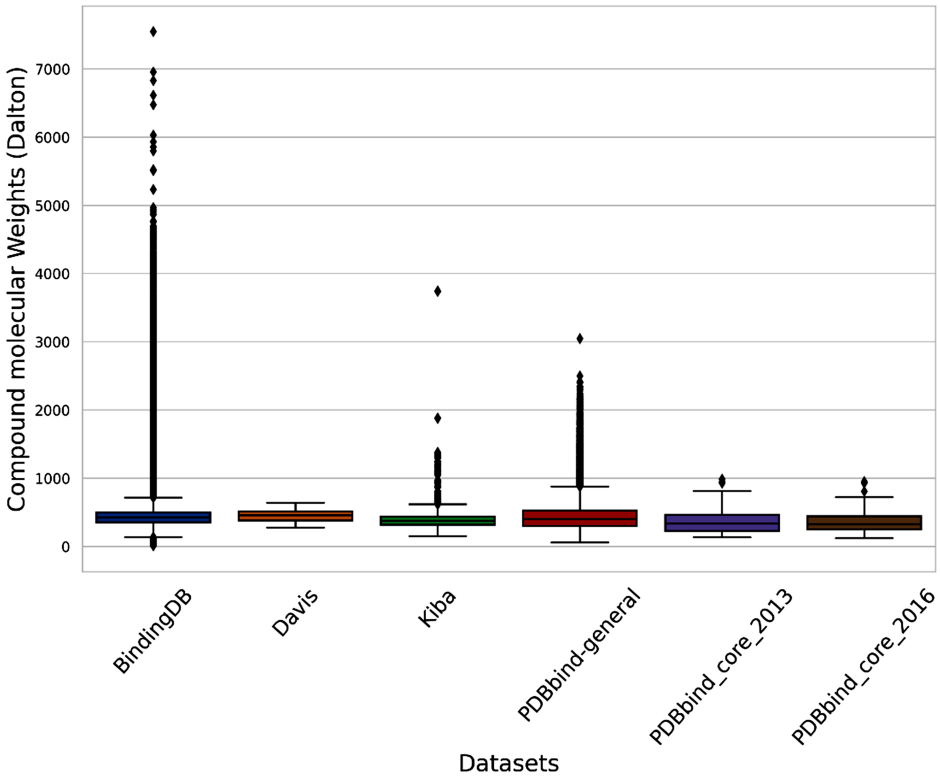

Fig. (3). Box plot of compound molecular weights in commonly used datasets. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

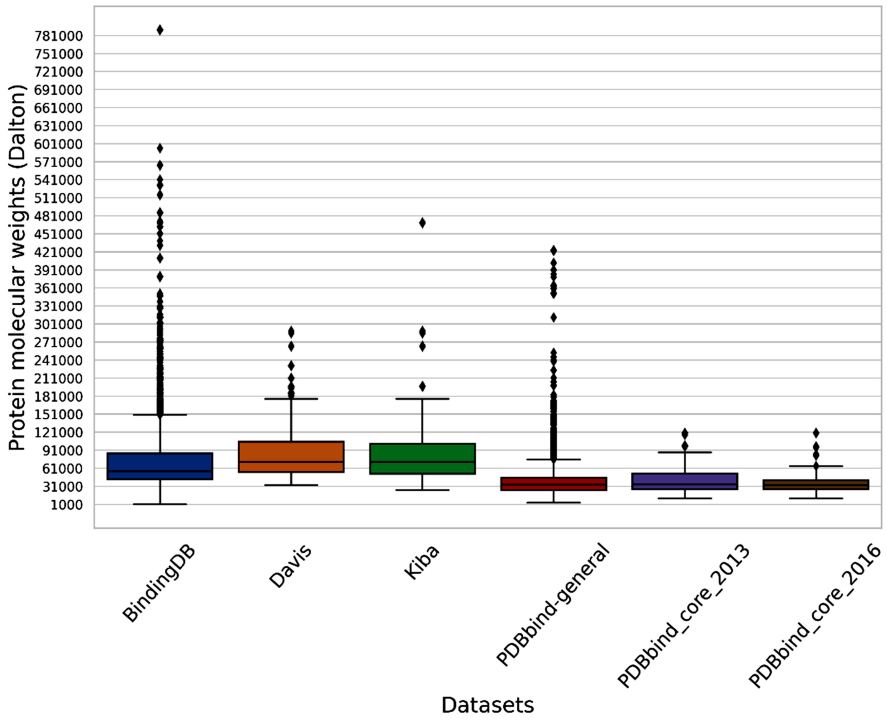

Fig. (4). Box plot of proteins molecular weights in commonly used datasets. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

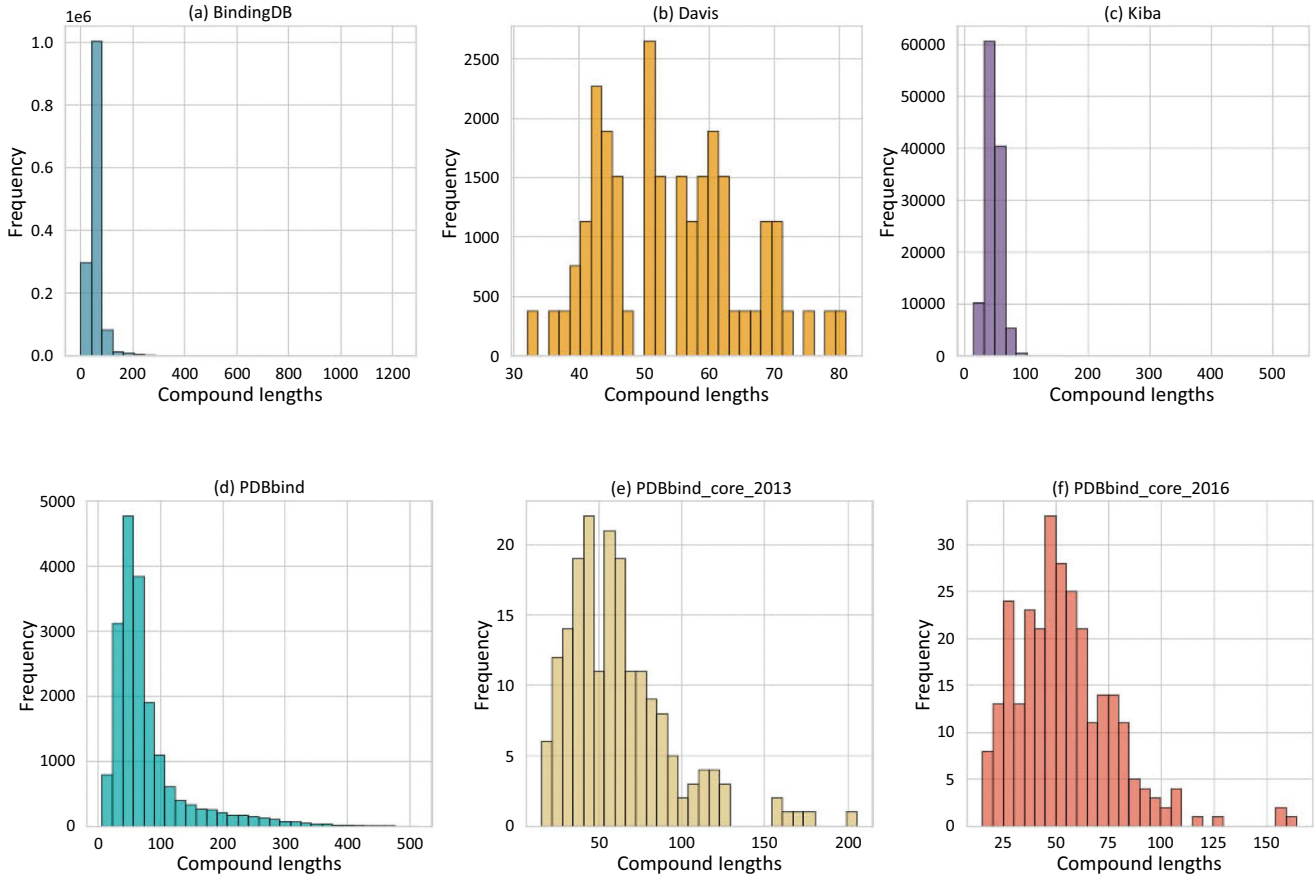

Fig. (5). Histograms of compound lengths in commonly used datasets: (**a**) Binding DB; (**b**) Davis; (**c**) Kiba; (**d**) PDBbind General set ; (**e**) PDBbind_core_2013; (**f**) PDBbind_core_2016. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Most BAP models primarily rely on sequence-level features, and the length distribution of these sequences significantly affects model performance. Identifying the optimal input sequence length during model training is crucial. In the general set, SMILE lengths of compounds vary from 6 to 510 characters (Fig. **5**), while protein sequences range from 31 to 3,833, with outlier thresholds set at 666 for proteins and 147 for compounds (Fig. **6**). When visualizing the compound's polar surface area against molecular weight and the octanol-water partition (LogP), similarities emerge between this dataset and compounds in the BindingDB dataset (Figs. **7a** and **7d**), likely due to overlapping molecules.

3.1.1. PDBbind Core_sets 2016 and 2013

The PDBbind core set is a curated collection of high--quality protein-ligand complexes used to validate and benchmark BAP methods. This essential set is the primary test set for the renowned Comparative Assessment of Scoring Functions (CASF) benchmark. The latest version, referred to here as PDBbind core_2016, is made of 285 proteins–ligand complexes, much more than the older version PDBbind core_2013, which contained 195 complexes.

A box plot analysis reveals that protein molecular weights in both sets range from 10,000 Da to 121,000 Da (Fig. **4**), while compound molecular weights fall between 121 Da and 950 Da (Fig. **3**). The mean compound molecular weight is 351 Da for core_2016 and 365 Da for core_2013.

The length distribution plots for both sets show diverse ranges of compound SMILE lengths. In core_2016, lengths vary from 15 to 164 (Fig. **5f**), while in core_2013, they range from 15 to 205 (Fig. **5e**). Similarly, protein lengths show variability, ranging from 99 to 1045 for core_2016 (Fig. **6f**) and 99 to 1052 for core_2013 (Fig. **6e**).

3.1.2. PDBbind-koff-2020

A more curated set introduced by Liu *et al.*, known as PDBbind-koff-2020 [33], offers a comprehensive collection of 680 protein-ligand complexes, each characterized by experimental dissociation rate constants (koff). PDBbind-koff-2020 encompasses 155 protein types with dissociation rate constants spanning ten orders of magnitude. The protein molecular weights vary from 11,780 Da to 378,000 Da, with 75% falling below 84,660 Da (Fig. **4**). Additionally, the

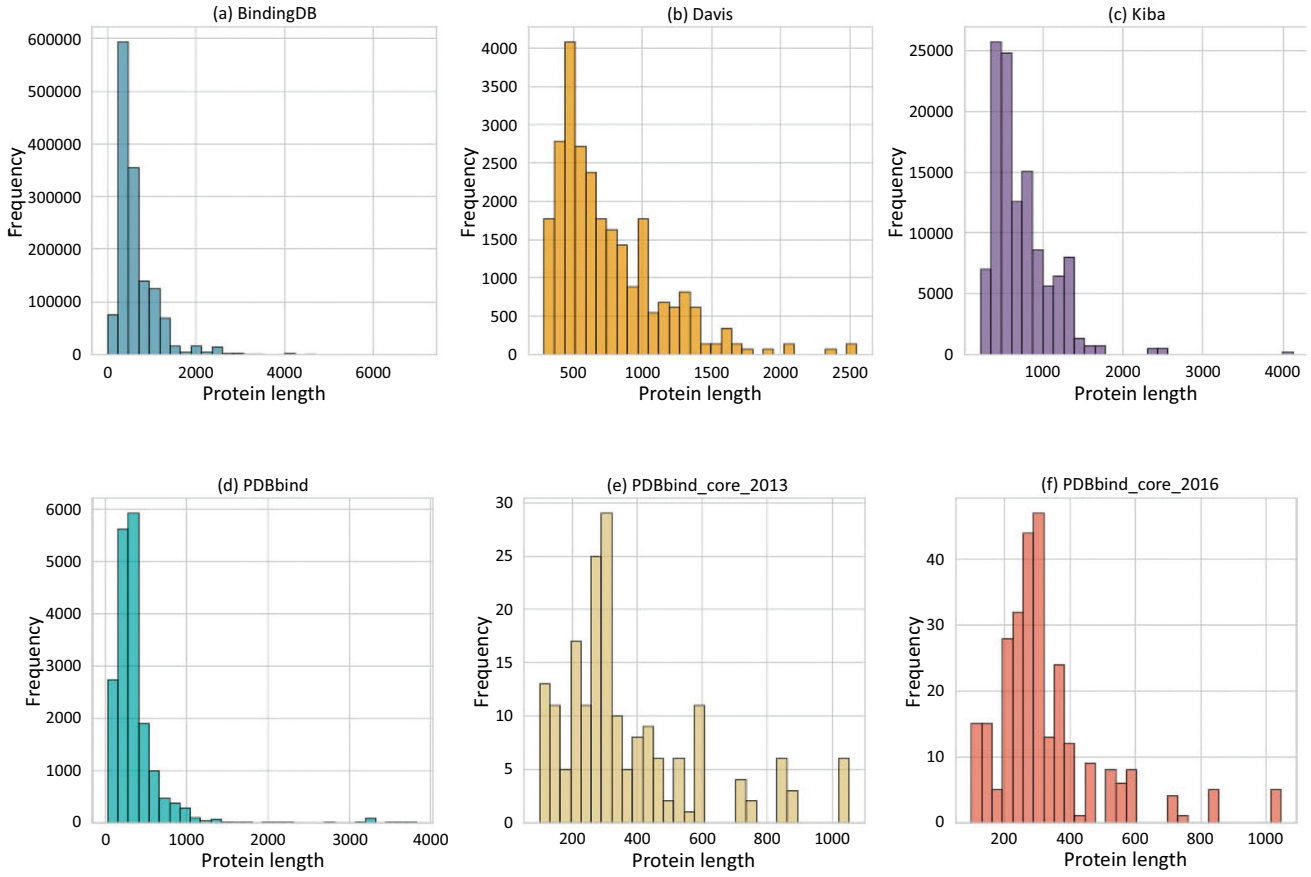

Fig. (6). Histograms of protein lengths in commonly used datasets: (**a**) Binding DB; (**b**) Davis; (**c**) Kiba; (**d**) PDBbind General set; (**e**) PDBbind_core_2013; (**f**) PDBbind_core_2016. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

histogram plot illustrates protein sequence lengths ranging from 99 to 3,390 (Fig. **6g**).

A concluding remark about the PDBbind datasets is the need for careful consideration. Some protein structures provided may contain issues, such as missing coordinates for certain residues or the presence of unusual residues. Moreover, there is significant overlap among the general, refined, and core sets. Due to the high occurrence of low-quality protein-ligand complexes in the general set, it is deemed more suitable to train a DL model using the refined set [34]. Within this refined set, complex structures and binding affinity values have been meticulously curated, thereby ensuring a higher data quality.

3.2. Davis

The Davis dataset, introduced by Davis *et al.* [28], is a collection of small-molecule kinase inhibitors and their interactions across the human protein kinome. It includes 72 known kinase inhibitors tested against a panel of 442 kinase assays, covering more than 80% of human protein kinase domains. It provides information on approximately 30,056 drug-target pairs measured in Kd.

The set contains proteins with molecular weights between 32,990 Da and 290,000 Da, as shown in Fig. (**4**). The compounds have molecular weights varying between 275 Da and 650 Da, as depicted in Fig. (**3**). Protein sequence lengths span between 288 and 2549 (Fig. **6b**), while compound lengths range from 32 to 81 (Fig. **5b**).

A scatter plot of the polar surface area, compound molecular weights, and LogP reveals a close association to the chemical space of the PDBbind core_2013 dataset (Figs. **7b** and **7e**). Both chemical spaces are spread out compared to the clustered space of the Binding DB and PDBbind-general sets.

A notable observation regarding the Davis dataset is the presence of overlapping drug-target pairs with other datasets, namely the Metz [35] and Anastassiadis [36] datasets. The Davis dataset is reported to be sharing 2575 drug-target pairs with the Metz dataset and 4255 with the Anastassiadis dataset [30].

3.3. BindingDB

BindingDB is a rich and diverse repository of experimentally determined protein-ligand binding affinities. The exten-

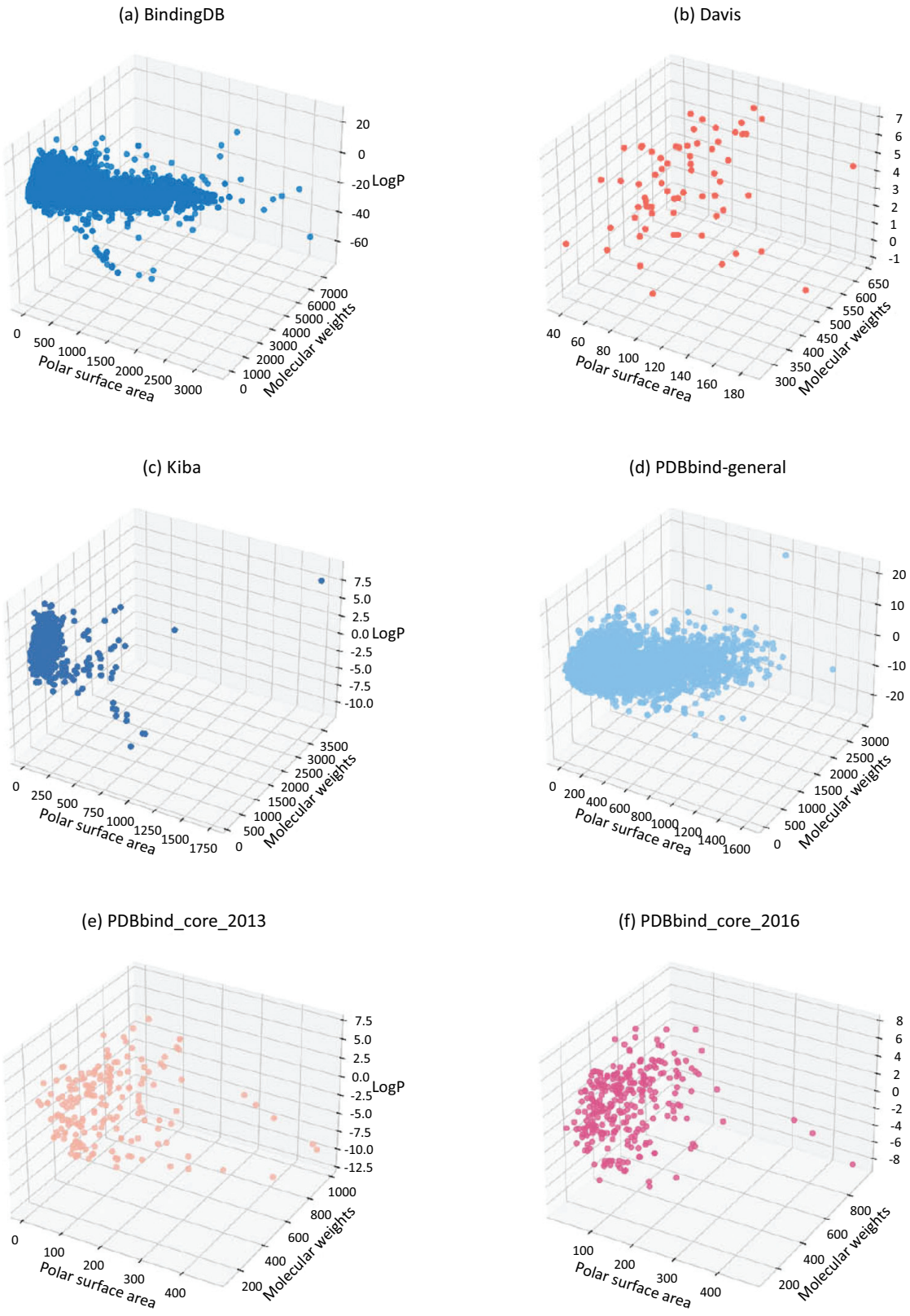

Fig. (7). (**a-f**) 3D Scatter plot illustrating the relationship between polar surface area, molecular weight, and LogP of compounds in commonly used datasets.

sive collection encompasses approximately 2.8 million entries, covering around 1.2 million unique compounds and 9200 distinct proteins. The binding affinity data are expressed as Kd, Ki, or IC_{50}.

A significant subset of this dataset, totaling 1.297 million entries associated with 594,000 compounds and 4500 proteins, has been curated from scientific articles by the BindingDB team. Additionally, BindingDB integrates entries from reputable databases, such as ChEMBL and PubChem, along with information drawn from patents. BindingDB offers a user-friendly web interface, facilitating searches based on protein names, ligand names, binding affinities, and experimental methods.

The molecular weight distribution within a subset of the BindingDB dataset, comprising 1.4 million entries, shows that 75% of compound molecular weights fall below 498 Da (Fig. **3**) and protein molecular weights below 86,000 Da (Fig. **4**). (Figs. **5a** and **6a**) demonstrate that both compound and protein lengths have a positively skewed distribution. Figs. (**7a** and **7d**) indicate a close similarity between the compounds of BindingDB and those of PDBbind in terms of their polar surface area, molecular weights, and LogP.

3.4. Kiba Dataset

KIBA (Kinase Inhibitor Bioactivity) is another commonly used dataset in BAP. It is a collection of bioactivity data related to kinase inhibitors, offering information about their interactions with various kinase targets. KIBA data integrate diverse sources, including large-scale biochemical assays, to provide a more holistic and accurate representation of kinase inhibitors bioactivity.

Containing data from three major assays and complemented by a novel model-based integration approach, this dataset contributes significantly to understanding the selectivity profiles of kinase inhibitors. It enhances the reliability of drug-target interaction classification by effectively addressing data heterogeneity and leveraging a combination of IC_{50}, Ki, and Kd measurements into a single score, known as the KIBA score, which represents the binding affinity.

KIBAs compounds have molecular weights between 150 and 3740 Da (Fig. **3**), while protein molecular weights range between 24,942 and 469,084 Da (Fig. **4**). Length distribution analysis indicates that most compounds have lengths between 14 and 100 (Fig. **5c**), while over 75% of protein lengths are below 915 (Fig. **6c**). Fig. (**7**) shows differences in KIBAs chemical space compared to PDBbind, BindingDB, and Davis.

4. COMMON DATA PREPROCESSING METHODS

4.1. Data Integration

BAP often requires integrating data from various databases to create a comprehensive and diverse dataset. The process involves several steps to ensure the quality, consistency, and reliability of the integrated dataset. The initial step in building a BAP model consists in gathering information on protein-ligand interactions from various datasets, such as those discussed in Section 3. It is essential to carefully select data encompassing a diverse range of protein-ligand interactions (PLIs) to capture different structural and chemical characteristics. This diversity is necessary for building robust and generalizable predictive models.

When performing data integration, it is crucial to establish common identifiers for matching entries across various databases. This involves aligning entries based on shared identifiers, such as protein or compound names, as well as specific identifiers, such as PDB identifiers for proteins in the PDB database, UniProt accession numbers for the UniProt database, or InChIKeys for compounds. Due to varying data acquisition and curation practices, different databases may employ inconsistent naming conventions or data representations.

4.2. Affinity Measures Conversion (IC_{50}, KD, KI)

Converting affinity measures is crucial in building an accurate and generalizable DL model for BAP. The necessity to transform the affinity values arises from the diverse ways in which experimental assays express binding affinity, typically in terms of Kd, IC_{50}, or Ki. These represent different aspects of the binding strength between a protein and a ligand. Using different units across datasets introduces challenges in comparing and integrating data for model training and evaluation. For instance, data from BindingDB comprise a multitude of assays, each adopting its preferred metric for binding affinity.

The variability in unit measurements can result in inconsistencies and inaccuracies within model predictions. Researchers often resort to strategies for converting affinity values to address the issue. One approach, as exemplified by DeepLPI [37], involves selecting a specific unit, such as Kd, and retaining only interactions represented in that unit. The approach ensures a consistent representation of binding affinity throughout the dataset.

In the case of MAPL-FMP [38] working with the PDBbind dataset, they considered instances with activity expressed in terms of either Ki or Kd while excluding those expressed in IC_{50} or EC50 values.

In their work on the PDBbind dataset, Shen *et al.* (2021) used the binding affinity expressed in the log space of the original units (−Log10Kx). Similarly, other models, such as DeepDTA [39], DeepGLSTM [40], and DeepCDA [41], employ a similar approach for datasets such as the Davis dataset. The affinity values were converted into PKD values using Eq. (1), where Kd is the dissociation constant. This conversion ensures that diverse datasets with varying affinity units can be effectively integrated into the model training process.

$$PKd = -\log_{10}\left(\frac{K_d}{1e^9}\right) \quad (1)$$

Table 1. Protein and ligand maximum lengths adopted by various authors.

Models	Datasets	Protein Max Length	Ligand Max Length
Electra-DTA [1]	BindingDB, Kiba, Davis	1000	100
ML_DTI [2]	Davis, Metz, Kiba	1200	100
Multi-PLI [3]	PDBbind, Davis	1200	200
GraphDTA [4]	Davis, Kiba	1000	-
WideDTA [5]	Davis	1000	85
DeepCDA [6]	Kiba, Davis, BindingDB	1000	100
ResBiGAAT [7]	PDBbind, CSAR-HiQ	1000	150
DeepDTA [8]	Davis	1200	85
	Kiba	1000	100
DeepGLSTM [9]	Davis, KIBA, DTC, Metz, ToxCast, STITCH	1000	-
Zhu *et al.* [10]	Davis, Metz, KIBA	1200	-
ColdDTA [11]	Davis KIBA	1200	
PCNN-DTA [12]	KIBA, Davis, BindingDB	1000	100
CAPLA [13]	PDBbind, CSAR-HiQ	1000	150
Zhang *et al.* [14]	Davis, Kiba	1000	100

4.3. Fixing optimal Sequence Length for Proteins and Ligands

In BAP, the length of both protein and ligand sequences can vary significantly, as described in Section 3. The variability poses a challenge for DL models, which typically require fixed-size inputs. A maximum length constraint for protein and ligand sequences is often implemented to address the issue. By implementing maximum length constraints, computational efficiency, generalizability, and performance can be improved. The specific length selection is guided by the datasets characteristics, the complexity of the binding interaction, and the available computational resources.

Determining the optimal length for protein and ligand sequences involves a balance between capturing sufficient information and maintaining computational efficiency. Table **1** lists the maximum length constraints used in various studies. The common choice for protein sequence length is 1000-1200, while the ligand sequence length typically ranges from 85 to 200.

4.4. Proteins and Ligands Representation

4.4.1. Ligands Representation

Diverse data formats and encoding schemes are employed to represent ligand compounds and proteins. Ligand compounds are commonly represented using Simplified Molecular Input Line Entry System (SMILES) notation, fingerprints, and molecular graphs.

SMILES are textual representations of chemical structures providing a concise and human-readable format for representing molecular structures. They encode information about atom types, bond types, and connectivity in a linear string. Widely utilized in literature, SMILES serves as the predominant representation of ligands. This textual representation is subsequently encoded using one-hot or label encoding techniques to facilitate its integration into DL models. The advantages of SMILES include their simplicity, ease of interpretation, and compatibility with various cheminformatics tools. However, SMILES have limitations, such as not capturing 3D spatial information and being sensitive to variations in molecular representations [42-53].

Fingerprint representations are also common. Many studies have successfully used molecular fingerprints to represent ligands. Fingerprints are binary vectors representing the presence or absence of specific substructures within a molecule. They efficiently capture key structural features and are suitable for similarity searching and ML tasks. For instance, RFDT [54] utilizes the PubChem substructure fingerprint made of 881 bits [55] to represent ligand compounds. Wang *et al.* (2021) investigated a variety of protein-ligand interaction fingerprints (IFPs) for BAP. They found that atom pair count-based and substructure-based IFPs performed well in target-specific and generic scoring tasks, demonstrating their potential for improving BAP accuracy [55]. However, fingerprint representations have limitations, such as the possibility of not capturing all relevant substructures and the variation in encoding schemes between different algorithms.

In addition to SMILES and fingerprints, the molecular graph is another common format for representing ligands. They represent molecules as graphs, where nodes represent

atoms and edges represent bonds. Molecular graphs offer an advantageous format for ligand representation. They capture both structural and topological information, allowing the incorporation of 3D spatial information. This format has been employed in various proposed models, including DeepGLSTM, GraphscoreDTA [56], DGraphDTA [57], and GraphDTA. While suitable for graph-based ML approaches, molecular graphs can be computationally expensive for processing large molecules.

In addition to the three major formats mentioned above, various other data formats are available for representing ligand compounds, each with advantages and disadvantages. For instance, Pafnucy [58] utilized 3D coordinates, explicitly capturing atomic positions in space. While this provides detailed structural information, it demands substantial computational resources.

4.4.2. Proteins Representation

Like ligand representation, various data formats and encoding schemes are used for proteins, including primary amino acid sequences, secondary structures, and graphs. Proteins primary amino acid sequence is a fundamental representation, encoding the linear arrangement of amino acids forming the protein chain. This representation is widely employed due to its simplicity and direct connection to the genetic code. In studies such as ML-DTI, DeepDTAF, Multi PLI, and CAPLA, protein sequences are encoded using methods such as one-hot encoding, where each amino acid is represented as a binary vector. WideDTA, ColdDTA, and PCNN-DTA [50] employed label encoding, assigning each unique amino acid a specific numerical label. While primary sequences offer insights into the linear structure of proteins, they may not capture the spatial relationships crucial for binding interactions.

Secondary structure representations also provide information about local folding patterns in proteins, including alpha-helices, beta-sheets, and loops. These representations are often derived from experimental methods like X-ray crystallography or predicted using algorithms such as DSSP (Dictionary of Secondary Structure of Proteins)) [59]. In MAPL-FMP [38] and CAPLA, protein secondary structures are calculated using the DSSP algorithm and incorporated along with the physicochemical properties of residues for comprehensive protein representation.

For a more detailed and spatially aware representation, proteins can be represented as graphs. In GraphscoreDTA and Lim *et al.* [60], the 3D structures of proteins were used to construct protein graphs, where nodes represent atoms and edges represent interactions. This representation preserves spatial relationships, contributing valuable information for binding affinity prediction. However, working with 3D structures can be computationally demanding, especially for large proteins.

In addition to these representations, contact maps derived from sequence alignment and predicted using methods such as Pconsc4 are utilized in ColdDTA, DGraph-DTA, and GSAML-DTA [61]. These maps capture residue-level interactions, offering a different perspective on protein structure and function.

In summary, the choice of data format for compounds or proteins depends on several factors, such as the desired level of detail and the computational resources available. Understanding the various options and their trade-offs allows one to choose the optimal representation for specific needs.

5. DEEP LEARNING MODELS FOR BAP

Selecting a suitable model becomes crucial after data preprocessing steps and appropriate data representation. The chosen model should effectively learn feature representations and extract valuable information to highlight the interactions between compounds and proteins, consequently enabling accurate binding affinity prediction. DL has emerged as a powerful tool in this domain, demonstrating the ability to learn complex patterns from extensive datasets and provide accurate predictions. Recent publications and trends suggest that DL models for BAP can be broadly classified based on their chosen feature extraction and learning model. Notably, CNNs, RNNs, GNNs, and transformer-based models emerge as the most prevalent choices. This section delves into a detailed exploration of the application of these models to BAP, with a focus on their respective strengths and limitations.

5.1. CNN-based Models

CNNs are inspired by the human visual system [62] and have proven highly effective in image processing [63-66]. They operate through convolutional, pooling, and fully connected layers, learning hierarchical features ranging from simple edges to complex patterns. Initially prominent in image classification and object detection, CNNs have expanded into chemogenomic methods by adapting to different data types. To accommodate various data formats, 1D CNNs, 2D CNNs, and 3D CNNs have evolved with specialized architectures tailored to the nature of the input data.

5.1.1. 1D CNN

1D CNN, or one-dimensional convolution neural network, is a fundamental building block in DL architectures designed for sequential data processing. Tailored for one-dimensional sequences such as time series or text, 1D CNN operates by sliding a filter or kernel along the input sequence, capturing local patterns and relationships within the data. This ability to capture local patterns and dependencies in sequential data makes it valuable in various applications, including speech recognition, sentiment analysis, and computational biology.

1D CNN has been extensively used for BAP because of its suitability for sequential data such as compound SMILES or protein amino acid sequences. In their work, Öztürk *et al.* [39] introduced DeepDTA, the first model to utilize 1D CNN for leveraging raw sequence information, employing SMILES for compounds and primary sequences for proteins. With its simple architecture in the Y shape using sim-

ple compound and protein representations, DeepDTA demonstrated improved predictive capabilities compared to established baselines such as KronRLS [67] and SimBoost [68]. Although DeepDTA's studies represent a significant advancement in utilizing CNNs for molecular sequence representation, they also acknowledge limitations, suggesting exploring alternative approaches, such as long short-term memory (LSTM), for more effective protein sequence representation. Using the same model as DeepDTA, AttentionDTA [69] introduces a novel approach to drug-target interaction prediction by incorporating an attention mechanism to enhance binding affinity predictions. While maintaining the use of the same protein and ligand representations as input, AttentionDTA leverages the power of attention mechanisms to focus on critical subsequences within proteins and drugs selectively. The attention-based enhancements, coupled with the use of 1D-CNN, lead to superior performance compared to DeepDTA.

DeepDTAF [70] presented a comprehensive approach to BAP by integrating diverse input representations and employing a 1D-CNN-based architecture. They integrated protein sequences, pocket sequences, and protein structural properties. For input representation, the method employed label encoding for ligand SMILES and OneHot encoding for amino acid sequences, resulting in rich and informative feature vectors. The proposed model architecture incorporated dilated convolution for proteins and ligands while using traditional convolution for pocket sequences. The model emphasized the importance of both local and global features and highlighted the significance of capturing interactions within binding pocket sequences and entire protein sequences.

Expanding upon the character-based DeepDTA approach, WideDTA [46] introduces a word-based model for representing protein and ligand sequences in BAP. Due to the non-sequential nature of motifs and domains in protein sequences, as well as the potential for overlapping residues in both motifs and maximum common substructures (MCS), they argue for the use of a word-based model over a character-based model. Employing a 1D CNN architecture, the authors integrated ligand SMILES, ligand maximum common substructure (LMCS), protein sequences, and protein motifs and domains as input features into the WideDTA model. Majumdar *et al.* [71] presented a 1D-CNN-based framework for ligand prediction against the S-glycoprotein of SARS-CoV--2. The input representation involves the fusion of protein sequence composition descriptors (PSC) (Lee *et al.*, 2019) and ligand-extended connectivity fingerprints (ECFP4). PSC descriptors capture essential features of protein sequences, comprising amino acid composition (AAC), dipeptide composition (DC), and tripeptide composition (TC). Zhu *et al.* (2023) introduced FingerDTA (X. Zhu *et al.*, 2023), a fingerprint-embedding framework for BAP. Using 1D-CNN and fully connected layers, FingerDTA integrates one-hot encoded protein and ligand sequences with global information represented by fingerprints generated from the whole sequence of drugs and ligands. In PCNN-DTA [50], the authors introduce the pyramid network convolution drug-target binding affinity, utilizing multiple 1D-CNN layers through a feature pyramid network (FPN). PCNN-DTA incorporates a 1D-CNN-based architecture within the FPN, allowing the extraction of hierarchical features from drugs and proteins represented as SMILES and primary sequences, respectively. The bottom-up pathway of the FPN employs residual 1D-CNNs to capture detailed information, while the top-down pathway enhances representation with deconvolution layers. The use of 1D-CNNs at various levels of the FPN enables the retention of both low-level and high-level information, contributing to improved binding affinity prediction accuracy.

While using readily available 1D representations such as SMILES for ligands and primary sequences for proteins offers certain advantages in computational efficiency and data handling, its application in predicting protein-ligand binding affinity comes with inherent limitations due to the complex nature of the interaction. The reliance on string-based representations may lead to the loss of crucial structural information, compromising the models predictive accuracy and diminishing the functional relevance of the learned latent space. Additionally, proteins and ligands are not rigid entities; they exhibit conformational flexibility, adopting different shapes that influence binding [72]. 1D models lack the ability to account for this dynamic behavior, potentially overlooking key conformations crucial for the interaction. These limitations underscore the importance of exploring alternative representations that better capture structural information, thereby enhancing the overall performance and interpretability of BAP models.

5.1.2. 2D CNN

A two-dimensional convolutional neural network, commonly known as 2D CNN, is a DL architecture designed for processing and analyzing structured grid data, such as images. Unlike its one-dimensional counterpart, the 2D CNN operates on two-dimensional input data, allowing it to capture spatial hierarchies and patterns within the input. This type of neural network has proven highly effective in image-related tasks due to its ability to recognize spatial relationships, edges, and hidden features. In BAP and molecular analysis, 2D CNNs represent a significant advancement beyond the constraints of 1D representations. They excel in handling input features such as intermolecular descriptors, structural information, and physicochemical properties, all of which can be effectively represented in two dimensions. This encompasses descriptors such as contact maps and evolutionary information.

OnionNet [73] proposed a 2D CNN for BAP. The authors input representation relied on rotation-free element-pair-specific contacts grouped into distance ranges. This resulted in a dataset with 3840 features transformed into a two-dimensional tensor to leverage the power of 2D CNNs. Similarly, Wang *et al.* (2021) presented OnionNet-2, which used 2D CNN to extract features from residue-atom contacting shells. The number of contacts in multiple distance shells characterized the interactions between protein residues and ligand atoms. Shim *et al.* (2021) introduced Sim-

CNN-DTA [74]. The model applied a 2D CNN to the outer product of column vectors derived from Tanimoto and Smith-Waterman similarity matrices for drugs and targets, respectively.

The workflow involved calculating similarity matrices, computing outer products, and utilizing a 2D CNN to extract deep features and predict binding affinities. In MDeePred [75], a multichannel protein featurization approach for BAP was proposed. The authors integrated diverse protein features, including sequence, structural, evolutionary, and physicochemical properties, into multiple 2D vectors to create comprehensive protein representations. MDeePred adopts a proteochemometric approach, utilizing compound and target protein features at the input level to model their interaction. The proposed method leverages a 2D-CNN architecture for the target protein side, where protein feature matrices are fed as input channels. The compound side employs a feed-forward neural network with circular molecular fingerprints as input. The outputs of the protein-side CNN and the compound-side neural network are combined in a pairwise-input hybrid deep neural network, leading to the binding affinity prediction.

The authors of MPS2IT-DTI [76] encoded molecule and protein sequences into images for BAP. They converted the molecules SMILES and amino acid sequences into matrices of numerical values, creating unique visual signatures for each sequence. The mapping process involved defining k-mers [77, 78], creating and normalizing counting vectors, and reshaping them into image matrices. Subsequently, the deep neural network of MPS2IT-DTI employed a dual 2D-CNN for both molecule and protein inputs to predict the binding affinity.

Although 2D CNNs have demonstrated promising outcomes in various proposed models, they also have certain drawbacks. One significant drawback arises from their inherent design, which is focused on 2D grids, potentially hindering their ability to fully capture the spatial complexity of three-dimensional intermolecular structures between proteins and ligands. The projection onto 2D grids may result in the loss of crucial spatial information critical for understanding molecular interactions. Furthermore, in contrast to 1D CNNs, 2D CNNs are not well suited for modeling sequential dependencies in proteins and ligands, which are essentially composed of amino acids and atoms [79]. These networks might inadequately capture long-range dependencies that may be essential for accurate predictions. Additionally, the performance of 2D CNNs is highly sensitive to the chosen input representations, such as molecular descriptors or features [80]. Inaccurate or insufficiently informative input representations can impede the models ability to discern relevant patterns, affecting its predictive accuracy.

5.1.3. 3D CNN

In contrast to 1D and 2D CNNs, 3D CNNs can leverage the three-dimensional structural information of molecular complexes. Here, intermolecular structures are transformed into 3D grids or voxel representations. Each voxel corresponds to a specific region in three-dimensional space. These grids effectively contain crucial information regarding the spatial arrangement of atoms, their properties, and their interactions.

Employing a 3D grid representation for molecular complexes, Pafnucy [58] utilized a 3D CNN architecture to extract spatial features. The input, structured as a 4D tensor, incorporated cartesian coordinates and atom feature vectors. Pafnucy surpassed classical scoring functions, including the X-Score [81], when evaluated on benchmark datasets like the PDBbind core set 2013 and Astex [82]. Pafnucy underscores the effectiveness of 3D CNNs in directly learning relevant features from the structural information of the complexes.

KDEEP [83] adopted a 3D CNN architecture inspired by SqueezeNet, tailored for 3D convolution tasks. The networks design simplified depth due to constraints related to training sample size and image resolution while maintaining other architectural aspects, including the use of rectified linear units (ReLUs) as activation functions. In KDEEPs input representation, descriptors adapted from prior work are used to represent proteins and ligands. The representation involves a 3D voxel scheme based on the van der Waals radius for each atom type, encompassing properties such as hydrophobicity, hydrogen bonding, aromaticity, and ionizability.

In Sfcnn [84], the authors adopted a concise featurization method, representing the protein-ligand complex with a 3D grid or 4D tensor. Unlike Pafnucy and KDeep, Sfcnn's approach simplifies atom featurization, extracting only basic atomic type information. The 3D CNN architecture was designed to handle the transformed input, and the model was evaluated on various datasets, outperforming other scoring functions. The authors highlighted the interpretability of the model through Grad-CAM, making intermediate layers of the neural network more understandable.

DeepAtom [85] utilized a 3D-CNN architecture to extract binding-related atomic interaction patterns from voxelized complex structures. The input representation in DeepAtom consisted of rasterizing protein-ligand complexes into a 3D grid box centered on the ligand. Each voxel in the grid box contains several input channels representing different raw information of atoms located around the voxel. The authors employed a lightweight 3D-CNN model to hierarchically extract useful atom interaction features supervised by the binding affinity score. The model architecture comprised three building blocks: the atom information integration block, the stacked feature extraction block, and the global affinity regression block. The atom information integration block utilizes a pointwise convolution layer and a 3D max pooling layer to fuse atom information across different channels and increase translational invariance.

Inspired by the ResNext architecture, AK-Score [86] employed an ensemble of 3D-CNN models, demonstrating improved prediction accuracy compared to previous models. The network architecture comprised 15 stacked layers of an

ensemble-based residual layer (RL) block. Each RL block consisted of three stacks of convolutional layers: batch normalization, rectified linear unit (ReLU) activation, and residual summation, facilitating parallel processing of the input tensors.

Limited generalization due to the scarcity of high-quality and unbiased labelled data poses a significant challenge for 3D CNNs in BAP [87, 88]. The available datasets may not adequately represent the diverse landscape of PLIs. Consequently, models trained on such data may struggle to generalize well on unseen protein-ligand pairs, hindering their predictive accuracy and real-world applicability [89].

Moreover, the memory usage and the computational cost required for 3D CNNs add another layer of complexity to the BAP. The computational demands of these voxel-based approaches escalate rapidly with spatial resolution, leading to significant memory usage and computational costs [90]. As a result, the computational overhead associated with 3D CNNs may limit their scalability and practical utility, particularly in high-throughput screening scenarios where efficiency is paramount. Table **2** presents the summary of CNN-based models for BAP.

5.2. RNN-based Models

A Recurrent Neural Network (RNN) [31] is a neural network architecture designed to process sequential data by maintaining hidden states that capture dependencies between elements in the sequence. Unlike traditional feedforward neural networks, RNNs have feedback loops that allow information to persist over time, making them well-suited for tasks where the order of elements matters, such as natural language processing, time series analysis, and speech recognition. This architecture offers several advantages over CNNs, including temporal dependency modeling, adaptability to variable-length sequences, adaptive feature extraction from sequential data, and computational efficiency. RNNs and their variants have been extensively used in computational biology for tasks ranging from sequence analysis [32] to structure prediction [33-35].

5.2.1. LSTM-based Models

To address the vanishing gradient problem that occurs when training standard RNNs on long sequences, specialized variants such as long short-term memory (LSTM) networks were introduced. As shown in Eq. (2), the LSTM unit updates its cell state. c_t and hidden state h_t based on the current input x_t, the previous hidden state h_{t-1}, the previous cell state c_{t-1}, and the gating mechanisms. This mechanism allows the LSTM to retain information over long sequences and excel in tasks reliant on long-term dependencies.

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1}) \quad (2)$$

In DeepCDA [6], LSTM is utilized as part of the model architecture to predict the binding affinity of compound-protein pairs. The LSTM network is combined with CNN layers to learn representations of both compounds and proteins. The raw protein sequences and compound SMILES strings are initially fed into the model as inputs. Subsequently, CNN is used to extract mid-level features hierarchically before passing them to the LSTM, which encodes the sequences dependencies. Additionally, the model incorporates a two-sided attention mechanism to capture the interaction strength between each protein substructure and compound substructure pair. Meanwhile, the authors of DGDTA [36] modified the GraphDTA model for BAP by introducing a triple-channel model in which LSTM was used to encode the proteins features.

DeepGLSTM [9] integrated a graph convolutional network (GCN) and bidirectional long short-term memory (Bi-LSTM) layers to predict the binding affinity between FDA-approved drugs and viral proteins of SARS-CoV-2. The GCN layers processed drug compounds while the Bi-LSTM layer processed the protein sequences, leveraging its bidirectional nature to capture temporal dependencies effectively.

5.2.2. GRU-based Models

Another variant is the Gated Recurrent Unit (GRU), which simplifies the LSTM architecture by combining the forget and input gates into a single update gate. GRU has been shown to be effective for BAP. To overcome the drawbacks associated with the loss of crucial compound information observed in GCN-based models, Wang *et al.* introduced SSGraphCPI [37], a three-channel DL framework. SSGraphCPI integrates a seq2seq model based on a GRU with an attention mechanism and graph convolutional neural networks (GCNNs). By leveraging GRU-based seq2seq architecture, the model translates input sequences into fixed-dimension vectors termed Thought Vectors, thereby enabling the effective representation of compound and protein features. In GDGRU-DTA [38], the authors proposed to enhance the GraphDTA model by incorporating the gated recurrent unit (GRU) and bidirectional gated recurrent unit (Bi-GRU) for interpreting protein sequences. The integration aimed to capture long-term dependencies in protein sequences more effectively. ResBiGAAT incorporated both physicochemical properties and sequence-level features of proteins and ligands to enhance prediction accuracy. The model employed a multi-layered Residual Bi-GRU coupled with two-sided self-attention mechanisms to capture long-term dependencies within sequences, achieving competitive performance in predicting the binding affinity.

The inherent drawback of the RNN-based models lies in their limited feature representation of proteins and compounds. As a result, they may fail to capture all aspects of PLIs that influence binding affinity. The lack of comprehensive representation can hinder their ability to predict binding affinity in real-world scenarios accurately. Furthermore, these models often suffer from a lack of generalizability, making them less practical for applications in predicting binding affinity.

Table 2. CNN-based methods for BAP.

Models	Ligand Input Representation	Protein Input Representation	Ligand and Protein Feature Learning	Comment
DeepDTA [8]	Lebel encoding of SMILES	Lebel encoding of AA seq	L: 1D CNN P: 1D CNN	No attention mechanism
AttentionDTA [15]	Lebel encoding of SMILES	Lebel encoding of AA seq	L: 1D CNN P: 1D CNN	Additional two-side multi-head attention mechanism
HyperAttentionDTI [91]	Lebel encoding of SMILES	Lebel encoding of AA seq	L: 1D CNN P: 1D CNN	Attention vector for each amino acid-atom pair
CAPLA [13]	one-hot encoding of SMILES	one-hot encoding of AA seq + SSE + PP	L: 1D CNN P: 1D CNN	Added protein-binding pocket AA seq the SSE
DeepDTAF [17]	Lebel encoding of SMILES	one-hot encoding of AA seq + SSE + PP	L: dilated convolution P: dilated convolution	Integration of protein-binding pocket and structural properties
WideDTA [5]	Lebel encoding of SMILES + LMCS	Lebel encoding of AA seq + Protein motifs and domains	L: 1D CNN P: 1D CNN	Combined 4 sequential features
[18]	ECFP4	PSC	L: 1D CNN P: 1D CNN	The 8420-dimensional feature vector for each protein sequence
DeepLPI [19]	Mol2Vec	ProSE	L: 1D CNN P: 1D CNN	CNN combined with LSTM
DeepCDA [6]	SMILES	Protein n-gram	L: 1D CNN P: 1D CNN	CNN combined with LSTM
FingerDTA [20]	one-hot encoding of SMILES + ECFP	one-hot encoding of AA seq + Target fingerprint	L: FC + 1D CNN P: FC + 1D CNN	Used 3 dense convolution blocks of 1D CNN
PCNN-DTA [12]	Lebel encoding of SMILES	Lebel encoding of AA seq	L: Residual 1D CNN P: Residual 1D CNN	Integration of double-sided MultiHead attention into a pyramid network
OnionNet [21]	RFEPSC	RFEPSC	L: 2D CNN P: 2D CNN	Contacts are grouped into different distance ranges.
OnionNet-2	Atom residues of complexes	Atom residues of complexes	L: 2D CNN P: 2D CNN	168 residue-atom combinations per shell
SimCNN-DTA [22]	drug similarity matrix	protein similarity matrix	L:2D CNN P:2D CNN	Similarity-based on the Tanimoto coefficient and normalized Smith-Waterman score
MDeePred [23]	ECFP4	structural + evolutionary + PP	L: FFN P: 2D CNN	1024-dimensional ECFP4 fingerprint generated from ligand SMILES
MPS2IT-DTI [24]	Molecule image	Protein image	L: 2D CNN P: 2D CNN	Treated drugs and protein sequences as 2D images
Pafnucy [25]	3D atomic coordinate of 19 channels		L: 3D CNN P: 3D CNN	The complex was cropped to a 20-Å cubic box centered on the ligands geometric midpoint.
KDEEP [26]	3D grid of 16 channels		L: 3D CNN P: 3D CNN	Van der Waals radius was employed for each atom type.
Sfcnn [27]	3D grid of 28 channels		L: 3D CNN P: 3D CNN	Atoms in protein-ligand complexes were grouped into 28 categories
DeepAtom [28]	3D grid of 24 channels		L: 3D CNN P: 3D CNN	11 Arpeggio atom types were used
AK-Score [29]	3D grid of 16 channels		L: 3D CNN P: 3D CNN	Van der Waals radius was employed for each atom type.
SE-OnionNet [30]	3D grid of 64 channels		L: 3D CNN P: 3D CNN	Eight element types were chosen to characterize ligand-protein contacts

Abbreviations: L: ligand; P: Proteins; ECFP: Extended-connectivity fingerprints; AA seq: Amino acid sequence; SSE: Secondary structure; PCS: protein sequence composition; RFEPSC: rotation-free element-pair-specific contacts; LMCS: ligand max common substructure; PP: physicochemical properties; ProSE: protein sequence embedding.

5.3. Transformer-based models

Transformers, including variants like BERT and GPT, are renowned for their effectiveness in sequence-to-sequence tasks, such as machine translation and text generation [39]. Transformers recently gained traction in biomedical and pharmaceutical research, including drug discovery. They have been successfully applied in various tasks such as molecular property prediction [40-42], protein structure prediction, sequence analysis [43-45], and genome analysis [46-48].

Hu *et al.* [49] combined a transformer and a graph attention network (GAT) to reduce data sparseness and computational costs while achieving better accuracy. The Transformer encoder processed the protein sequence, while the GAT handled the protein contact map and drug embeddings. This approach enabled the model to effectively capture the structural information of proteins and drugs, leading to improved performance compared to traditional methods. TranDTA [50] investigated BAP through diverse protein input feature techniques, including UniRep [51], ProtBert, and ProtAlbert [52]. Compound drugs were represented using molecular fingerprints and RoBERTa-encoded SMILES sequences [53]. The authors found that using ProtAlbert for protein representation and molecular fingerprint representation for drugs produced optimal results. DTITR [54] employed two parallel Transformer-Encoders for proteins and compounds, incorporating self-attention and cross-attention layers to predict binding affinity effectively. In PLAPT [55], the authors leveraged transfer learning from pretrained transformers such as ProtBERT [52, 56], and ChemBERTa [57], utilizing protein primary sequences and SMILES notation for ligands. The approach enabled them to achieve high accuracy while minimizing computational resources, highlighting the importance of utilizing pretrained models in computational drug discovery.

While transformers offer significant advantages over CNN models in capturing complex patterns, there are some constraints inherent to their architecture and functionality. Transformer-based models typically require large amounts of labeled data for effective training. However, obtaining labeled data for compound-protein interactions can be challenging and may limit the models ability to generalize well to unseen interactions or compounds. Furthermore, transformers may struggle to generalize effectively across diverse compound-protein interaction datasets due to variations in compound structures, protein conformations, and binding affinities. Differences across datasets can pose significant challenges for transformers in learning robust representations that generalize across different domains. Additionally, transformers, especially large-scale models like ProtBert and ProtAlbert, are computationally intensive and require substantial resources for training and inference. The complexity can pose challenges in terms of computational cost and scalability, particularly for large datasets or real-time applications.

5.4. Graph-based Models

Graph neural networks (GNNs) [58] have emerged as powerful tools in drug discovery and computational biology, particularly in binding affinity prediction. Traditional methods often rely on sequence-based or structure-based approaches, which may overlook molecular interactions and fail to capture the complex relationships between molecules. GNNs offer a unique advantage by enabling the representation of molecular structures as graphs, where atoms or residues serve as nodes and chemical bonds or interactions as edges. The graph-based representation allows GNNs to capture the spatial and functional relationships between atoms or residues more effectively, thus enhancing the accuracy of binding affinity prediction.

In terms of input representation, some authors represent only drug molecules as a graph while maintaining the primary sequences of the target proteins. Examples include GraphDTA [4], DeepGLSTM [9], GDGRU-DTA [38], DeepGS [59], and EmbedDTI [60]. In these approaches, the drug molecules are represented as graphs to capture their structural features, while the primary sequences of the proteins are processed separately to extract relevant information. Conversely, some authors represent drug molecules and the target proteins as graphs. Examples include Dgraph-DTA [61], X-DPI [62], and WGNN-DTA [63]. These methods construct graphs for both the drug molecule and the protein, allowing for the integration of structural and interaction information from both entities.

Various GNNs have been proposed to address the complexities inherent in BAP. Among these models is the Graph Attention Networks (GAT), which assign varying weights to neighboring nodes, enabling focused attention on important features during message passing. DGDTA [36] introduced a dynamic graph attention network combined with Bi-LSTM to enhance binding affinity prediction (BAP). The model integrated dynamic graph attention by representing drug compounds as graphs and proteins as 1D sequences to capture important features from drug graphs. Additionally, Bi-LSTM processed protein sequences to extract contextual information. By integrating 3D structural information of proteins with 2D graph representations of ligands, PSG-BAR [64] employed a residual graph attention network to predict binding affinity effectively. Their approach enhances the models predictive capabilities by leveraging both the protein structures and molecular graph features of ligands, offering a comprehensive understanding of the complex PLIs.

Another prominent graph architecture used in BAP is the graph convolutional network (GCN) [65], which applies convolutional operations on graph-structured data to aggregate information from neighboring nodes. GCNs excel at capturing local structural features within molecular graphs, providing insights into the complex relationships between atoms or residues. To reduce the computational cost associated with utilizing 3D voxelized grid cubes, Son *et al.* proposed a graph convolutional neural network termed GraphBAR [66]. Their model employs a graph convolutional network converting binding complexes into graphs with the binding site de-

termined by the distance between ligands and proteins. APMNet [67] introduced a cascade graph convolutional neural network model for binding affinity prediction. The architecture incorporates graph convolutional layers, including ARMA [68] and MPNN, to capture molecular graph features effectively and improve predictive performance. In GCAT [69], the attention-enhanced graph cross-convolution network was employed as the model architecture for exploring the binding affinity between drugs and proteins based on the atomic arrangement in three-dimensional space. GCAT consisted of a cross-convolution, which simulated interactions between the protein and the drug through an aggregate-update mechanism and self-attention pooling, utilized for generating graph-level representations.

In addition, graph isomorphism networks (GIN) [70] have also proven to be efficient in BAP by learning graph representations invariant to node ordering. GINs employ message-passing schemes to iteratively update node embeddings, enabling robust modeling of complex molecular structures. Their ability to capture variations in molecular configurations makes them well-suited for predicting binding affinities and elucidating the underlying mechanisms of molecular interactions. GIN was proven effective in the GraphDTA [4] study, demonstrating the best performance across the Kiba and Davis datasets compared to GCN and GAT. Table **3** presents a comprehensive list of various GNNs introduced to predict binding affinity between proteins and ligands efficiently.

Table 3. Graph-based models for BAP.

Models	Ligand Input Representation	Protein Input Representation	Ligand Feature Learning	Protein Feature Learning
GraphDTA [4]	Molecular graph	one-hot encoding of AA seq	GCN\GAT\GIN\GATGCN	CNN
DeepGLSTM [9],	Molecular graph	Lebel encoding of AA seq	Multiblock GCN	Bi-LSTM
GDGRU-DTA [38],	Molecular graph	Lebel encoding of AA seq	GNN	GRU/BiGRU
EmbedDTI [60]	Atom graph and substructure graph	AA seq	GCN with attention	1D CNN
DeepGS [59]	Molecular graph + Smi2Vec	Prot2Vec	GAT + BiGRU	CNN
Dgraph-DTA [61]	Molecular graph	Protein graph	GCN	GCN
X-DPI [62]	Molecular graph + Mol2vec embedding	Protein graph + TAPE embedding	GCN	GCN
WGNN-DTA [63]	Molecular graph	Weighted protein graph	GCN/GAT	GCN/GAT
DGDTA [36]	Molecular graph	Lebel encoding of AA seq	Dynamic GAT + GCN	Bi-LSTM + CNN
PSG-BAR [64]	Molecular graph	Protein graph	Residual GAT	Residual GAT
GraphBAR [66]	Binding complexes graph	Binding complexes graph	Multiblock GCN	Multiblock GCN
APMNet [67]	Binding complexes graph	Binding complexes graph	GCN	GCN
LGN [71]	Molecular graph + IFP	Complexes graph	GIN	GNN
PLANET [72]	Molecular graph	Protein graph	GNN	EGCL
GraphscoreDTA [73]	Molecular graph	Protein graph + Interaction Graph	GNN-GRU	GNN + GNN-GRU
GraphATT-DTA [74]	Molecular graph	Lebel encoding of AA seq	GNN(GAT/GIN/GCN/MPNN/DMPNN)	1D CNN
AttentionMGT-DTA [75]	Molecular graph	Protein graph	Graph transformer	Graph transformer
GLCN-DTA [76]	Molecular graph	Protein graph	GLCN	GLCN
IEDGEDTA [77]	Molecular graph	Protein graph	Edge-GCN	1D-GCN
SAG-DTA [78]	Molecular graph	Lebel encoding of AA seq	GCN	1D CNN

Abbreviations: AA seq: Amino acid sequence; GNN — graph neural network; MLP — multilayer perceptron; GCN: graph convolutional network; GAT: graph attention network; GIN: graph isomorphism network; CNN: convolutional neural network; FC: fully connected layer; GRU: gated recurrent unit; EGCL: Equivariant graph convolutional layer, MPNN: message passing neural network, IFP: interaction fingerprint; GLCN: Graph learning convolutional network.

5.5. Emerging Deep Learning Methods for BAP

In addition to GNNs and transformers, several other emerging DL techniques are being explored for BAP. Some of these include Transfer learning and Reinforcement learning.

Transfer learning techniques leverage pre-trained models or knowledge from related tasks to improve CPI prediction performance, particularly in limited labelled data contexts. By transferring knowledge from tasks with abundant data to tasks with sparse data, transfer learning enhances model generalizability and prediction accuracy. [92-140] explored deep transfer learning to predict drug-target interactions for understudied proteins, showing superior performance for datasets with fewer than 100 compounds compared to training from scratch. In [141], A3C, a novel reinforcement learning technique, notably improved protein-ligand docking predictions for single and multi-atom ligands compared to a basic model. Extension of this approach to binding affinity prediction could yield improved performance.

Another emerging method is the capsule networks, which offer an alternative approach to traditional CNNs by representing hierarchical structures within data. Capsule networks have shown promise in capturing spatial relationships and hierarchical features in molecular structures, making them suitable for CPI prediction [142]. introduced CapB-M-DTI, combining capsule networks with pre-trained BERT for target protein sequence extraction and MPNN for compound graph feature extraction, showing good performance and applicability in virtual screening, including for COVID-19 treatment. Despite not directly treating binding affinity as a regression task, these emerging methods show potential for application in binding affinity prediction.

Furthermore, Meta-learning and Few-shot learning methods are gaining attention in the context of CPI. The authors of MetaDTA proposed Meta-learning, a technique focused on developing models that can quickly adapt to new tasks or domains based on previous experiences. MetaDTAs approach leverages Attentive Neural Processes (ANPs) to model binding affinities for each target protein as a regression function of compounds, demonstrating superior performance even with limited data availability [143-151].

On the other hand, ZeroBind authors proposed Few-shot learning, where models are trained to generalize from a small number of labeled examples [152]. ZeroBinds protein-specific zero-shot predictor utilizes subgraph matching for drug-target interactions, achieving remarkable performance, especially for unseen proteins and drugs, and showcasing adaptability even with limited prior information. Integrating these approaches into the landscape of BAP methods could lead to further improvements in predictive accuracy and generalizability, particularly for underrepresented proteins.

6. CHALLENGES AND FUTURE DIRECTIONS FOR BAP

Despite significant advancements, several challenges persist, including data quality and availability, model complexity, interpretability and explainability.

6.1. Data Quality

While the quantity of available datasets has increased, ensuring their quality remains a persistent challenge. Achieving accurate predictions in BAP heavily relies on the quality and quantity of available datasets [143]. The challenge lies in obtaining comprehensive datasets encompassing a diverse range of molecular interactions. The need for more high-quality data, particularly for specific protein-ligand pairs or underrepresented classes, hampers the ability to train models that generalize well across different independent test sets.

6.2. Model Complexity

Deep learning models, while powerful in their predictive capabilities, often exhibit high levels of complexity. For instance, when using a 3D voxel grid representation of protein-ligand complexes to train a 3D CNN, the number of trainable parameters scales with factors like grid size, number of channels per voxel, and the complexity of the convolutional layers. The dimensions of the voxel grid determine the input size, with larger and higher-resolution grids resulting in more input features and, subsequently, more trainable parameters.

Similarly, in GNN, the size of the molecular graph, determined by the number of atoms and interactions, influences the model complexity. The chosen GNN architecture, including the number of layers, message-passing functions, and hidden feature dimensions, also contributes to the total trainable parameters. Simpler GNN architectures with fewer layers and lower dimensions can enhance computational efficiency. Notably, research suggests that effective GNN models can achieve good performance with a relatively small number of parameters [144], making them potentially more scalable for practical applications compared to 3D CNNs with larger parameter sizes associated with voxel grids.

6.3. Interpretability and Explainability

Many DL models used for BAP are complex and act as black boxes. Understanding how they arrive at a specific prediction can make the model trustworthy and help to identify potential biases. PLIs involve complex 3D structures and numerous physicochemical properties. Capturing these complexities often leads to high-dimensional data representations that are challenging to interpret directly. Attention mechanisms can highlight specific regions of the protein or ligand that the model focuses on for prediction. This can provide clues about the interatomic interactions driving binding affinity. However, it comes at the cost of making the model more computationally expensive and adding more learnable parameters.

Some emerging techniques, such as Grad-CAM (Gradient-weighted Class Activation Mapping) [145], can help visualize the areas of a protein structure most influential for

the models prediction [84, 86]. The visual heatmap can offer insights into the spatial patterns the model considers important. Similarly, the LIME [146] and SHAP [147] techniques can be applied to understand how individual features contribute to specific predictions or the overall model behavior. Even if the feature that contributed to the prediction is known, it is still necessary to understand how it contributed. Hence, it is necessary to make the models more explainable. This underscores the ongoing efforts in the scientific community to develop explainable models tailored to the complexity of BAP.

6.4. Integration of Computational and Experimental Approaches

Recent research efforts, particularly in the context of COVID-19, underscore the practical impact of computational methods in drug discovery. For example, at the Experimental Drug Development Centre (EDDC), computational molecular modeling was employed to screen a large library of FDA-approved drugs against the SARS-CoV-2 main protease (3CLpro) [148]. This approach led to the identification of 47 promising candidates based on their computed binding affinities, which were then experimentally validated. Several drugs, including boceprevir and ivermectin, demonstrated significant inhibitory effects on 3CLpro, highlighting how computational predictions can effectively guide experimental drug development. This underscores that while computational methods are invaluable in narrowing down potential drugs, they are most effective when combined with experimental approaches to confirm therapeutic efficacy and safety.

6.5. Shifts in BAP Modeling

There has been a noticeable shift in the paradigm for BAP, with graph neural networks emerging as the trending architecture. This transition signifies a move from the conventional use of convolutional neural networks toward adopting graph neural networks. The unique ability of GNNs to effectively model complex relationships within molecular structures has positioned them as a preferred choice in BAP, reflecting the continuous evolution and exploration of innovative GNN architectures in computational biology. Future directions could focus on further enhancing the capabilities and applications of GNNs for BAP. One potential avenue for exploration involves refining GNN architectures to be more explainable and interpretable. This would address the challenge of understanding how GNNs arrive at specific predictions, thus increasing trust in the models and facilitating their adoption in practical settings.

In addition to interpretability, it is essential to ensure that GNNs remain computationally efficient in terms of time and resources without compromising model performance. Optimizing GNN architectures and training procedures can help mitigate computational burdens while maintaining predictive accuracy, enabling widespread adoption across various computational biology applications.

Future directions should focus on refining data splitting strategies and model assessment techniques to address the challenges of imbalanced affinity value distributions and dataset representativeness. By considering protein and compound similarity and affinity value distribution during dataset splitting, researchers can ensure the creation of representative sets of protein-ligand pairs and prevent the overestimation of DL models. This comprehensive approach enhances model robustness, generalizability, and practical applicability in BAP tasks. Additionally, Generative models have the potential to revolutionize BAP prediction by facilitating de novo ligand design and data augmentation. By generating novel molecular structures with optimized binding properties, thus expanding the chemical space available for exploration, these models can improve the generalizability and optimize the BAP algorithm. NeuralPLexer [149] is a valuable example, which predicts 3D structures of protein-ligand complexes, which is particularly beneficial for proteins with limited experimental data. Moreover, the integration of protein structural information beyond the binding pocket into BAP models holds significant potential. Exploring the influence of allosteric sites and distant protein regions on ligand binding can provide valuable insights into drug design.

The reliability and consistency of the experimental measurements in the datasets should be ensured. More precisely, there is a need to address multiple binding affinity reports for the same protein-ligand pair. Sometimes, the same protein-ligand complex may have reported different binding affinity values. The problem can arise from various factors, including experimental conditions, measurement techniques, or data curation practices. A strategy for handling multiple affinity reports must be devised to address these inconsistencies. Options include selecting a representative value, considering a range, or employing consensus methods. The rationale behind the chosen approach should be clearly documented to maintain transparency.

Duplicate entries should be appropriately addressed. Duplicate entries can introduce biases and affect the models generalizability by over-representing certain instances. Identifying and removing duplicate entries from the integrated dataset is essential to maintaining data integrity and preventing biases. This can be achieved by utilizing common identifiers, such as PDB IDs, UniProt IDs, or compound IDs, to match and eliminate redundant records.

Finally, recent studies have highlighted the efficacy of Voronoi entropy as ligand molecular descriptors. The innovative approach, as demonstrated by Sergey *et al.* [150], emphasizes the importance of integrating advanced molecular descriptors into predictive models to enhance accuracy and insight. Future research could explore the synergies between GNNs and Voronoi entropy, leveraging the unique capabilities of both approaches to improve binding affinity prediction models.

CONCLUSION

In conclusion, this survey provides a comprehensive overview of various preprocessing steps and deep learning models for predicting the interaction strength between proteins and ligands. Additionally, our analysis of commonly used datasets offers insights into the diverse properties of these data, such as molecular weights, length of sequences, and the octanol-water partition of proteins and ligands. The state-of-the-art approach for BAP currently relies on GNN. However, despite their effectiveness, GNNs still require further efforts for interpretability and explainability. Moving forward, exploring additional avenues, such as generative AI and incorporating new molecular descriptors, is essential to enhance the predictive performance of DL models for BAP. Our work serves as a valuable starting point for both new researchers entering the field of protein-ligand binding affinity and experienced researchers. It offers guidance on dataset exploration, best practices, and effective model building while highlighting and addressing the challenges encountered.

AUTHORS' CONTRIBUTIONS

It is hereby acknowledged that all authors have accepted responsibility for the manuscripts content and consented to its submission. They have meticulously reviewed all results and unanimously approved the final version of the manuscript.

LIST OF ABBREVIATIONS

BAP	=	Binding Affinity Prediction
BIGRU	=	Bidirectional Gated Recurrent Unit
CNN	=	Convolutional Neural Network
CPI	=	Compound-Protein Interaction
DL	=	Deep Learning
GNN	=	Graph Neural Network
GRU	=	Gated Recurrent Unit
IC_{50}	=	Half Maximal Inhibitory Concentration
KD	=	Equilibrium Dissociation Constant
KI	=	Inhibition Constant
LOGP	=	Octanol-Water Partition
LSTM	=	Long Short-Term Memory
ML	=	Machine Learning
PLA	=	Protein-Ligand Binding Affinity
PLI	=	Protein-Ligand Interaction
RESBIGAAT	=	Residual Bidirectional Gated Recurrent Unit with Attention
RNN	=	Recurrent Neural Network
SMILES	=	Simplified Molecular Input Line Entry System

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & Communications Technology Promotion).

CONFLICT OF INTEREST

The authors declared no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

During the preparation of this work, the author(s) used Grammarly and its AI assistance functionality solely to improve readability and correct grammatical errors.

REFERENCES

[1] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004; 3(8): 673-83. http://dx.doi.org/10.1038/nrd1468 PMID: 15286734

[2] Mylonas SK, Axenopoulos A, Daras P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. Bioinformatics 2021; 37(12): 1681-90. http://dx.doi.org/10.1093/bioinformatics/btab009 PMID: 33471069

[3] Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 2010; 38(2): W469-73. http://dx.doi.org/10.1093/nar/gkq406

[4] McNutt AT, Francoeur P, Aggarwal R, *et al.* GNINA 1.0: molecular docking with deep learning. J Cheminform 2021; 13(1): 43. http://dx.doi.org/10.1186/s13321-021-00522-2 PMID: 34108002

[5] Chen L, Tan X, Wang D, *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics 2020; 36(16): 4406-14. http://dx.doi.org/10.1093/bioinformatics/btaa524 PMID: 32428219

[6] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions *via* deep learning with convolution on protein sequences. PLOS Comput Biol 2019; 15(6): e1007129. http://dx.doi.org/10.1371/journal.pcbi.1007129 PMID: 31199797

[7] 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. http://dx.doi.org/10.1016/j.compbiomed.2023.107136 PMID: 37329615

[8] Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform 2021; 22(1): 247-69. http://dx.doi.org/10.1093/bib/bbz157 PMID: 31950972

[9] Wang DD, Zhu M, Yan H. Computationally predicting binding affinity in protein–ligand complexes: free energy-based simulations and machine learning-based scoring functions. Brief Bioinform 2021; 22(3): bbaa107. http://dx.doi.org/10.1093/bib/bbaa107 PMID: 32591817

[10] Wang Y, Jiao Q, Wang J, Cai X, Zhao W, Cui X. Prediction of protein-ligand binding affinity with deep learning. Comput Struct Biotechnol J 2023; 21: 5796-806. http://dx.doi.org/10.1016/j.csbj.2023.11.009

[11] Meli R, Morris GM, Biggin PC. Scoring Functions for Protein-Li-

gand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. Front Bioinform 2022; 2: 885983. http://dx.doi.org/10.3389/fbinf.2022.885983 PMID: 36187180

[12] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol 2011; 162(6): 1239-49. http://dx.doi.org/10.1111/j.1476-5381.2010.01127.x PMID: 21091654

[13] Mullard A. New drugs cost US$2.6 billion to develop. Nat Rev Drug Discov 2014; 13(12): 877-7. http://dx.doi.org/10.1038/nrd4507 PMID: 25435204

[14] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ 2016; 47: 20-33. http://dx.doi.org/10.1016/j.jhealeco.2016.01.012 PMID: 26928437

[15] Pushpakom S, Iorio F, Eyers PA, *et al.* Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2019; 18(1): 41-58. http://dx.doi.org/10.1038/nrd.2018.168 PMID: 30310233

[16] Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 2011; 12(4): 303-11. http://dx.doi.org/10.1093/bib/bbr013 PMID: 21690101

[17] Acharya C, Coop A, Polli JE, Mackerell AD Jr. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Computeraided Drug Des 2011; 7(1): 10-22. http://dx.doi.org/10.2174/157340911793743547 PMID: 20807187

[18] Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins 1990; 8(3): 195-202. http://dx.doi.org/10.1002/prot.340080302 PMID: 2281083

[19] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen J Mol Biol 1997; 267(3): 727-48. http://dx.doi.org/10.1006/jmbi.1996.0897

[20] Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 2016; 99(3): 285-97. http://dx.doi.org/10.1002/cpt.318 PMID: 26659699

[21] Huang H, Shukla H, Wu C, Saxena S. Challenges and solutions in proteomics. Curr Genomics 2007; 8(1): 21-8. http://dx.doi.org/10.2174/138920207780076910 PMID: 18645629

[22] Lim S, Lu Y, Cho CY, *et al.* A review on compound-protein interaction prediction methods: Data, format, representation and model. Comput Struct Biotechnol J 2021; 19: 1541-56. http://dx.doi.org/10.1016/j.csbj.2021.03.004 PMID: 33841755

[23] Bateman A, Martin M-J, Orchard S, *et al.* UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021; 49(D1): D480-9. http://dx.doi.org/10.1093/nar/gkaa1100 PMID: 33237286

[24] Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. Nucleic Acids Res 2000; 28(1): 235-42. http://dx.doi.org/10.1093/nar/28.1.235 PMID: 10592235

[25] Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. Nucleic Acids Res 2006; 34: D668-72. http://dx.doi.org/10.1093/nar/gkj067

[26] Gaulton A, Hersey A, Nowotka M, *et al.* The ChEMBL database in 2017. Nucleic Acids Res 2017; 45(D1): D945-54. http://dx.doi.org/10.1093/nar/gkw1074 PMID: 27899562

[27] Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 2004; 47(12): 2977-80. http://dx.doi.org/10.1021/jm030580l PMID: 15163179

[28] Davis MI, Hunt JP, Herrgard S, *et al.* Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 2011; 29(11): 1046-51. http://dx.doi.org/10.1038/nbt.1990 PMID: 22037378

[29] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 2007; 35(1): D198-201. http://dx.doi.org/10.1093/nar/gkl999

[30] Tang J, Szwajda A, Shakyawar S, *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 2014; 54(3): 735-43. http://dx.doi.org/10.1021/ci400709d PMID: 24521231

[31] Li Y, Han L, Liu Z, Wang R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. J Chem Inf Model 2014; 54(6): 1717-36. http://dx.doi.org/10.1021/ci500081m PMID: 24708446

[32] Su M, Yang Q, Du Y, *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model 2019; 59(2): 895-913. http://dx.doi.org/10.1021/acs.jcim.8b00545 PMID: 30481020

[33] Liu H, Su M, Lin HX, Wang R, Li Y. Public Data Set of Protein–Ligand Dissociation Kinetic Constants for Quantitative Structure–Kinetics Relationship Studies. ACS Omega 2022; 7(22): 18985-96. http://dx.doi.org/10.1021/acsomega.2c02156 PMID: 35694511

[34] Liu Z, Li Y, Han L, *et al.* PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 2015; 31(3): 405-12. http://dx.doi.org/10.1093/bioinformatics/btu626 PMID: 25301850

[35] Metz JT, Johnson EF, Soni NB, Merta PJ, Kifle L, Hajduk PJ. Navigating the kinome. Nat Chem Biol 2011; 7(4): 200-2. http://dx.doi.org/10.1038/nchembio.530 PMID: 21336281

[36] Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol 2011; 29(11): 1039-45. http://dx.doi.org/10.1038/nbt.2017 PMID: 22037377

[37] Wei B, Zhang Y, Gong X. DeepLPI: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. Sci Rep 2022; 12(1): 18200. http://dx.doi.org/10.1038/s41598-022-23014-1 PMID: 36307509

[38] Kundu I, Paul G, Banerjee R. A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties. RSC Advances 2018; 8(22): 12127-37. http://dx.doi.org/10.1039/C8RA00003D PMID: 35539386

[39] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug– target binding affinity prediction. Bioinformatics 2018; 34(17): i821-9. http://dx.doi.org/10.1093/bioinformatics/bty593 PMID: 30423097

[40] Mukherjee S, Ghosh M, Basuchowdhuri P. Deep Graph Convolutional Network and LSTM based approach for predicting drug-target binding affinity 2022. http://dx.doi.org/10.1137/1.9781611977172.82

[41] Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics 2020; 36(17): 4633-42. http://dx.doi.org/10.1093/bioinformatics/btaa544 PMID: 32462178

[42] Wang J, Wen N, Wang C, Zhao L, Cheng L. ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. J Cheminform 2022; 14(1): 14. http://dx.doi.org/10.1186/s13321-022-00591-x PMID: 35292100

[43] Yang Z, Zhong W, Zhao L, Chen CYC. ML-DTI: Mutual Learning Mechanism for Interpretable Drug–Target Interaction Prediction. J Phys Chem Lett 2021; 12(17): 4247-61. http://dx.doi.org/10.1021/acs.jpclett.1c00867 PMID: 33904745

[44] Hu F, Jiang J, Wang D, Zhu M, Yin P. Multi-PLI: interpretable multi-task deep learning model for unifying protein–ligand interaction datasets. J Cheminform 2021; 13(1): 30. http://dx.doi.org/10.1186/s13321-021-00510-6 PMID: 33858485

[45] Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 2021; 37(8): 1140-7. http://dx.doi.org/10.1093/bioinformatics/btaa921 PMID: 33119053

[46] Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drug--target binding affinity 2019. http://dx.doi.org/10.48550/ARXIV.1902.04166

[47] Aly Abdelkader G, Ngnamsie Njimbouom S, Oh TJ, Kim JD. ResBiGAAT: Residual Bi-GRU with attention for protein-ligand binding affinity prediction. Comput Biol Chem 2023; 107: 107969.

http://dx.doi.org/10.1016/j.compbiolchem.2023.107969 PMID: 37866117

[48] Zhu Z, Yao Z, Zheng X, *et al.* Drug–target affinity prediction method based on multi-scale information interaction and graph optimization. Comput Biol Med 2023; 167: 107621. http://dx.doi.org/10.1016/j.compbiomed.2023.107621 PMID: 37907030

[49] Fang K, Zhang Y, Du S, He J. ColdDTA: Utilizing data augmentation and attention-based feature fusion for drug-target binding affinity prediction. Comput Biol Med 2023; 164: 107372. http://dx.doi.org/10.1016/j.compbiomed.2023.107372 PMID: 37597410

[50] Chen Y, Zhu Y, Zhang Z, Wang J, Wang C. Prediction of drug protein interactions based on variable scale characteristic pyramid convolution network. Methods 2023; 211: 42-7. http://dx.doi.org/10.1016/j.ymeth.2023.02.007 PMID: 36804213

[51] Jin Z, Wu T, Chen T, *et al.* CAPLA: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. Bioinformatics 2023; 39(2): btad049. http://dx.doi.org/10.1093/bioinformatics/btad049 PMID: 36688724

[52] Zhang L, Ouyang C, Liu Y, Liao Y, Gao Z. Multimodal contrastive representation learning for drug-target binding affinity prediction. Methods 2023; 220: 126-33. http://dx.doi.org/10.1016/j.ymeth.2023.11.005 PMID: 37952703

[53] Mokaya M, Imrie F, Van Hoorn WP, Kalisz A, Bradley AR, Deane CM. "Testing the Limits of SMILES-based *De Novo* Molecular Generation with Curriculum and Deep Reinforcement Learning," Bioinformatics 2022. http://dx.doi.org/10.1101/2022.07.15.500218

[54] Wang L, You ZH, Chen X, Yan X, Liu G, Zhang W. RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions Using Drug Structure and Protein Sequence Information. Curr Protein Pept Sci 2018; 19(5): 445-54. http://dx.doi.org/10.2174/1389203718666161114111656 PMID: 27842479

[55] Fernández-de Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform 2017; 9(1): 9. http://dx.doi.org/10.1186/s13321-017-0195-1 PMID: 28224019

[56] Wang K, Zhou R, Tang J, Li M. GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics 2023; 39(6): btad340. http://dx.doi.org/10.1093/bioinformatics/btad340 PMID: 37225408

[57] Jiang M, Li Z, Zhang S, *et al.* Drug–target affinity prediction using graph neural network and contact maps. RSC Advances 2020; 10(35): 20701-12. http://dx.doi.org/10.1039/D0RA02297G PMID: 35517730

[58] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics 2018; 34(21): 3666-74. http://dx.doi.org/10.1093/bioinformatics/bty374 PMID: 29757353

[59] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22(12): 2577-637. http://dx.doi.org/10.1002/bip.360221211 PMID: 6667333

[60] Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY. Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. J Chem Inf Model 2019; 59(9): 3981-8. http://dx.doi.org/10.1021/acs.jcim.9b00387 PMID: 31443612

[61] Liao J, Chen H, Wei L, Wei L. GSAML-DTA: An interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. Comput Biol Med 2022; 150: 106145. http://dx.doi.org/10.1016/j.compbiomed.2022.106145 PMID: 37859276

[62] Fu H, Niu Z, Zhang C, Ma J, Chen J. Visual Cortex Inspired CNN Model for Feature Construction in Text Analysis. Front Comput Neurosci 2016; 10: 64. http://dx.doi.org/10.3389/fncom.2016.00064 PMID: 27471460

[63] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition 2015. http://dx.doi.org/10.48550/ARXIV.1512.03385

[64] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017; 60(6): 84-90. http://dx.doi.org/10.1145/3065386

[65] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2014. http://dx.doi.org/10.48550/ARXIV.1409.1556

[66] Szegedy C, *et al.* Going Deeper with Convolutions 2014. http://dx.doi.org/10.48550/ARXIV.1409.4842

[67] Pahikkala T, Airola A, Pietilä S, *et al.* Toward more realistic drug-target interaction predictions. Brief Bioinform 2015; 16(2): 325-37. http://dx.doi.org/10.1093/bib/bbu010 PMID: 24723570

[68] He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform 2017; 9(1): 24. http://dx.doi.org/10.1186/s13321-017-0209-z PMID: 29086119

[69] Zhao Q, Xiao F, Yang M, Li Y, Wang J. AttentionDTA: prediction of drug–target binding affinity using attention model 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). San Diego, CA, USA. 2019; pp. In: IEEE; 64-9. http://dx.doi.org/10.1109/BIBM47256.2019.8983125

[70] Wang K, Zhou R, Li Y, Li M. DeepDTAF: a deep learning method to predict protein–ligand binding affinity. Brief Bioinform 2021; 22(5): bbab072. http://dx.doi.org/10.1093/bib/bbab072 PMID: 33834190

[71] Majumdar S, Nandi SK, Ghosal S, *et al.* Deep Learning-Based Potential Ligand Prediction Framework for COVID-19 with Drug-Target Interaction Model. Cognit Comput 2021; (Feb): 1-13. http://dx.doi.org/10.1007/s12559-021-09840-x PMID: 33552306

[72] Du X, Li Y, Xia YL, *et al.* Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci 2016; 17(2): 144. http://dx.doi.org/10.3390/ijms17020144 PMID: 26821017

[73] Zheng L, Fan J, Mu Y. OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. ACS Omega 2019; 4(14): 15956-65. http://dx.doi.org/10.1021/acsomega.9b01997 PMID: 31592466

[74] Shim J, Hong ZY, Sohn I, Hwang C. Prediction of drug–target binding affinity using similarity-based convolutional neural network. Sci Rep 2021; 11(1): 4416. http://dx.doi.org/10.1038/s41598-021-83679-y PMID: 33627791

[75] Rifaioglu AS, Cetin Atalay R, Cansen Kahraman D, Doğan T, Martin M, Atalay V. MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. Bioinformatics 2021; 37(5): 693-704. http://dx.doi.org/10.1093/bioinformatics/btaa858 PMID: 33067636

[76] de Souza JG, Fernandes MAC, de Melo Barbosa R. A Novel Deep Neural Network Technique for Drug–Target Interaction. Pharmaceutics 2022; 14(3): 625. http://dx.doi.org/10.3390/pharmaceutics14030625 PMID: 35336000

[77] Melsted P, Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics 2011; 12(1): 333. http://dx.doi.org/10.1186/1471-2105-12-333 PMID: 21831268

[78] Rizk G, Lavenier D, Chikhi R. DSK: *k* -mer counting with very low memory usage. Bioinformatics 2013; 29(5): 652-3. http://dx.doi.org/10.1093/bioinformatics/btt020 PMID: 23325618

[79] Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: A survey. Mech Syst Signal Process 2021; 151: 107398. http://dx.doi.org/10.1016/j.ymssp.2020.107398

[80] Alzubaidi L, Zhang J, Humaidi AJ, *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021; 8(1): 53.

http://dx.doi.org/10.1186/s40537-021-00444-8 PMID: 33816053
[81] Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des 2002; 16(1): 11-26. http://dx.doi.org/10.1023/A:1016357811882 PMID: 12197663
[82] Hartshorn MJ, Verdonk ML, Chessari G, *et al.* Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 2007; 50(4): 726-41. http://dx.doi.org/10.1021/jm061277y PMID: 17300160
[83] Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. K_{DEEP} : Protein–Ligand Absolute Binding Affinity Prediction *via* 3D-Convolutional Neural Networks. J Chem Inf Model 2018; 58(2): 287-96. http://dx.doi.org/10.1021/acs.jcim.7b00650 PMID: 29309725
[84] Wang Y, Wei Z, Xi L. Sfcnn: a novel scoring function based on 3D convolutional neural network for accurate and stable protein–ligand affinity prediction. BMC Bioinformatics 2022; 23(1): 222. http://dx.doi.org/10.1186/s12859-022-04762-3 PMID: 35676617
[85] Li Y, Rezaei MA, Li C, Li X. DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). San Diego, CA, USA. 2019; pp. In: IEEE; 303-10. http://dx.doi.org/10.1109/BIBM47256.2019.8982964
[86] Kwon Y, Shin WH, Ko J, Lee J. AK-Score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. Int J Mol Sci 2020; 21(22): 8424. http://dx.doi.org/10.3390/ijms21228424 PMID: 33182567
[87] Volkov M, Turk JA, Drizard N, *et al.* On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. J Med Chem 2022; 65(11): 7946-58. http://dx.doi.org/10.1021/acs.jmedchem.2c00487 PMID: 35608179
[88] Yang J, Shen C, Huang N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. Front Pharmacol 2020; 11: 69. http://dx.doi.org/10.3389/fphar.2020.00069 PMID: 32161539
[89] Chen L, Cruz A, Ramsey S, *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One 2019; 14(8): e0220113. http://dx.doi.org/10.1371/journal.pone.0220113 PMID: 31430292
[90] Liu Q, Wang PS, Zhu C, *et al.* OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein-ligand affinity prediction. J Mol Graph Model 2021; 105: 107865. http://dx.doi.org/10.1016/j.jmgm.2021.107865 PMID: 33640787
[91] Zhao Q, Zhao H, Zheng K, Wang J. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 2022; 38(3): 655-62. http://dx.doi.org/10.1093/bioinformatics/btab715 PMID: 34664614
[92] Zhu X, Liu J, Zhang J, Yang Z, Yang F, Zhang X. FingerDTA: A Fingerprint-Embedding Framework for Drug-Target Binding Affinity Prediction. Big Data Mining and Analytics 2023; 6(1): 1-10. http://dx.doi.org/10.26599/BDMA.2022.9020005
[93] Wang S, Liu D, Ding M, *et al.* SE-OnionNet: A Convolution Neural Network for Protein–Ligand Binding Affinity Prediction. Front Genet 2021; 11: 607824. http://dx.doi.org/10.3389/fgene.2020.607824 PMID: 33737946
[94] Schmidt RM. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview 2019. http://dx.doi.org/10.48550/ARXIV.1912.05911
[95] Shen Z, Zhang Q, Han K, Huang DS. A Deep Learning Model for RNA-Protein Binding Preference Prediction Based on Hierarchical LSTM and Attention Network IEEE/ACM Transactions on Computational Biology and Bioinformatics. Washington, DC, United States.: IEEE Computer Society Press 2020; pp. 1-1. http://dx.doi.org/10.1109/TCBB.2020.3007544
[96] Jisna VA, Jayaraj PB. Protein Structure Prediction: Conventional and Deep Learning Perspectives. Protein J 2021; 40(4): 522-44. http://dx.doi.org/10.1007/s10930-021-10003-y PMID: 34050498
[97] Panda B, Majhi B. A novel improved prediction of protein structural class using deep recurrent neural network. Evol Intell 2021; 14(2): 253-60. http://dx.doi.org/10.1007/s12065-018-0171-3
[98] Wang L, Zhong X, Wang S, Zhang H, Liu Y. A novel end-to-end method to predict RNA secondary structure profile based on bidirectional LSTM and residual neural network. BMC Bioinformatics 2021; 22(1): 169. http://dx.doi.org/10.1186/s12859-021-04102-x PMID: 33789581
[99] Zhai H, Hou H, Luo J, Liu X, Wu Z, Wang J. DGDTA: dynamic graph attention network for predicting drug–target binding affinity. BMC Bioinformatics 2023; 24(1): 367. http://dx.doi.org/10.1186/s12859-023-05497-5 PMID: 37777712
[100] Wang X, Liu J, Zhang C, Wang S. SSGraphCPI: A Novel Model for Predicting Compound-Protein Interactions Based on Deep Learning. Int J Mol Sci 2022; 23(7): 3780. http://dx.doi.org/10.3390/ijms23073780 PMID: 35409140
[101] Zhijian L, Shaohua J, Yigao L, Min G. GDGRU-DTA: Predicting Drug-Target Binding Affinity Based on GNN and Double GRU 2022. http://dx.doi.org/10.5121/csit.2022.120703
[102] Vaswani A, *et al.* Attention Is All You Need 2017. http://dx.doi.org/10.48550/ARXIV.1706.03762
[103] Li H, Zhao D, Zeng J. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. In: ACM; 14 August; Washington DC USA: 2022; pp. 857-67. http://dx.doi.org/10.1145/3534678.3539426
[104] Wang S, Guo Y, Wang Y, Sun H, Huang J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Niagara Falls NY USA. 2019; pp. In: ACM; 04 September; 429-36. http://dx.doi.org/10.1145/3307339.3342186
[105] Zhang X, Guo H, Zhang F, *et al.* HNetGO: protein function prediction *via* heterogeneous network transformer. Brief Bioinform 2023; 24(6): bbab556. http://dx.doi.org/10.1093/bib/bbab556 PMID: 37861172
[106] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 2021; 37(15): 2112-20. http://dx.doi.org/10.1093/bioinformatics/btab083 PMID: 33538820
[107] Le NQK, Ho QT, Nguyen TTD, Ou YY. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. Brief Bioinform 2021; 22(5): bbab005. http://dx.doi.org/10.1093/bib/bbab005 PMID: 33539511
[108] Zhang L, Qin X, Liu M, Liu G, Ren Y. BERT-m7G: A Transformer Architecture Based on BERT and Stacking Ensemble to Identify RNA N7-Methylguanosine Sites from Sequence Information. Comput Math Methods Med 2021; 2021: 1-10. http://dx.doi.org/10.1155/2021/7764764 PMID: 34484416
[109] Baid G, Cook DE, Shafin K, *et al.* DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat Biotechnol 2022; (Sep): http://dx.doi.org/10.1038/s41587-022-01435-7 PMID: 36050551
[110] Clauwaert J, Menschaert G, Waegeman W. Explainability in transformer models for functional genomics. Brief Bioinform 2021; 22(5): bbab060. http://dx.doi.org/10.1093/bib/bbab060 PMID: 33834200
[111] Raad J, Bugnon LA, Milone DH, Stegmayer G. miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs. Bioinformatics 2022; 38(5): 1191-7. http://dx.doi.org/10.1093/bioinformatics/btab823 PMID: 34875006
[112] Hu F, Hu Y, Zhang J, Wang D, Yin P. Structure Enhanced Protein-Drug Interaction Prediction using Transformer and Graph Embedding 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). In: IEEE; Seoul, Korea. 2020; pp. 16-19 December; 1010-4. http://dx.doi.org/10.1109/BIBM49941.2020.9313456
[113] Saadat M, Behjati A, Zare-Mirakabad F, Gharaghani S. “Drug-Tar-

get Binding Affinity Prediction Using Transformers," Bioinformatics 2021.
http://dx.doi.org/10.1101/2021.09.30.462610
[114] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019; 16(12): 1315-22.
http://dx.doi.org/10.1038/s41592-019-0598-1 PMID: 31636460
[115] Elnaggar A, *et al.* CodeTrans: Towards Cracking the Language of Silicon's Code Through Self-Supervised Deep Learning and High Performance Computing 2021.
http://dx.doi.org/10.48550/ARXIV.2104.02443
[116] Liu Y, *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach 2019.
http://dx.doi.org/10.48550/ARXIV.1907.11692
[117] Monteiro NRC, Oliveira JL, Arrais JP. DTITR: End-to-end drug–target binding affinity prediction with transformers. Comput Biol Med 2022; 147: 105772.
http://dx.doi.org/10.1016/j.compbiomed.2022.105772 PMID: 35777085
[118] Rose T, Monti N, Anand N, Shen T. "PLAPT: Protein-Ligand Binding Affinity Prediction Using Pretrained Transformers," Bioinformatics 2024.
http://dx.doi.org/10.1101/2024.02.08.575577
[119] Elnaggar A, *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing 2020.
http://dx.doi.org/10.48550/ARXIV.2007.06225
[120] Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction 2020.
http://dx.doi.org/10.48550/ARXIV.2010.09885
[121] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains 2005 IEEE International Joint Conference on Neural Networks. 31 July 2005; Montreal, Que., Canada. 2005; pp. In: IEEE; 729-34.
http://dx.doi.org/10.1109/IJCNN.2005.1555942
[122] Lin X. DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction 2020.
http://dx.doi.org/10.48550/ARXIV.2003.13902
[123] Jin Y, Lu J, Shi R, Yang Y. EmbedDTI: Enhancing the Molecular Representations *via* Sequence Embedding and Graph Convolutional Network for the Prediction of Drug-Target Interaction. Biomolecules 2021; 11(12): 1783.
http://dx.doi.org/10.3390/biom11121783 PMID: 34944427
[124] Wang P, *et al.* "X-DPI: A structure-aware multi-modal deep learning model for drug-protein interactions prediction," Bioinformatics 2021.
http://dx.doi.org/10.1101/2021.06.17.448780
[125] Jiang M, Wang S, Zhang S, Zhou W, Zhang Y, Li Z. Sequence-based drug-target affinity prediction using weighted graph neural networks. BMC Genomics 2022; 23(1): 449.
http://dx.doi.org/10.1186/s12864-022-08648-9 PMID: 35715739
[126] Pandey M, Radaeva M, Mslati H, *et al.* Ligand Binding Prediction Using Protein Structure Graphs and Residual Graph Attention Networks. Molecules 2022; 27(16): 5114.
http://dx.doi.org/10.3390/molecules27165114 PMID: 36014351
[127] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks 2016.
http://dx.doi.org/10.48550/ARXIV.1609.02907
[128] Son J, Kim D. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. PLoS One 2021; 16(4): e0249404.
http://dx.doi.org/10.1371/journal.pone.0249404 PMID: 33831016
[129] Shen H, Zhang Y, Zheng C, Wang B, Chen P. A Cascade Graph Convolutional Network for Predicting Protein–Ligand Binding Affinity. Int J Mol Sci 2021; 22(8): 4023.
http://dx.doi.org/10.3390/ijms22084023 PMID: 33919681
[130] Bianchi FM, Grattarola D, Livi L, Alippi C. Graph Neural Networks with Convolutional ARMA Filters. IEEE Trans Pattern Anal Mach Intell 2022; 44(7): 3496-507.
http://dx.doi.org/10.1109/TPAMI.2021.3054830 PMID: 33497331
[131] Feng X, Qu J, Wang T, Wang B, Lyu X, Tang Z. Attention-enhanced Graph Cross-convolution for Protein-Ligand Binding Affinity Prediction 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Houston, TX, USA. 2021; pp. In: IEEE; 1299-302.
http://dx.doi.org/10.1109/BIBM52615.2021.9669341
[132] Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? 2018.
http://dx.doi.org/10.48550/ARXIV.1810.00826
[133] Guo J. Improving structure-based protein-ligand affinity prediction by graph representation learning and ensemble learning. PLoS One 2024; 19(1): e0296676.
http://dx.doi.org/10.1371/journal.pone.0296676 PMID: 38232063
[134] Zhang X, *et al.* "PLANET: A Multi-Objective Graph Neural Network Model for Protein–Ligand Binding Affinity Prediction," Bioinformatics 2023.
http://dx.doi.org/10.1101/2023.02.01.526585
[135] Bae H, Nam H. GraphATT-DTA: Attention-Based Novel Representation of Interaction to Predict Drug-Target Binding Affinity. Biomedicines 2022; 11(1): 67.
http://dx.doi.org/10.3390/biomedicines11010067 PMID: 36672575
[136] Wu H, Liu J, Jiang T, *et al.* AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Neural Netw 2024; 169: 623-36.
http://dx.doi.org/10.1016/j.neunet.2023.11.018 PMID: 37976593
[137] Qi H, Yu T, Yu W, Liu C. Drug–target affinity prediction with extended graph learning-convolutional networks. BMC Bioinformatics 2024; 25(1): 75.
http://dx.doi.org/10.1186/s12859-024-05698-6 PMID: 38365583
[138] Suviriyapaisal N, Wichadakul D. "iEdgeDTA: integrated edge information and 1D graph convolutional neural networks for binding affinity prediction," Chemistry 2023.
http://dx.doi.org/10.26434/chemrxiv-2023-qs2w0
[139] Zhang S, Jiang M, Wang S, Wang X, Wei Z, Li Z. SAG-DTA: Prediction of Drug–Target Affinity Using Self-Attention Graph Network. Int J Mol Sci 2021; 22(16): 8993.
http://dx.doi.org/10.3390/ijms22168993 PMID: 34445696
[140] Dalkıran A, Atakan A, Rifaioğlu AS, *et al.* Transfer learning for drug–target interaction prediction Bioinformatics 2023; 39(1): i103-10.
http://dx.doi.org/10.1093/bioinformatics/btad234
[141] Wang C, Chen Y, Zhang Y, *et al.* A reinforcement learning approach for protein–ligand binding pose prediction. BMC Bioinformatics 2022; 23(1): 368.
http://dx.doi.org/10.1186/s12859-022-04912-7 PMID: 36076158
[142] Huang Y, Huang HY, Chen Y, *et al.* A Robust Drug–Target Interaction Prediction Framework with Capsule Network and Transfer Learning. Int J Mol Sci 2023; 24(18): 14061.
http://dx.doi.org/10.3390/ijms241814061 PMID: 37762364
[143] Fan FJ, Shi Y. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. Bioorg Med Chem 2022; 72: 117003.
http://dx.doi.org/10.1016/j.bmc.2022.117003 PMID: 36103795
[144] Zhang S, Jin Y, Liu T, *et al.* SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction. ACS Omega 2023; 8(25): 22496-507.
http://dx.doi.org/10.1021/acsomega.3c00085 PMID: 37396234
[145] Selvaraju RR, Cogswell M, Das A, *et al.* Grad-CAM: Visual Explanations from Deep Networks *via* Gradient-Based Localization 2017 IEEE International Conference on Computer Vision (ICCV). Venice. 2017; pp. In: IEEE; 22-29 October 2017; 618-26.
http://dx.doi.org/10.1109/ICCV.2017.74
[146] Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You?': Explaining the Predictions of Any Classifier Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. 2016; pp. In: ACM; 13 August 2016; 1135-44.
http://dx.doi.org/10.1145/2939672.2939778
[147] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions.Advances in Neural Information Processing Systems 30. Curran Associates, Inc. 2017; I: pp. 4765-74.https://proceedings.neurips.cc/paper_files/paper/2017/-file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[148] Mody V, Ho J, Wills S, *et al.* Identification of 3-chymotrypsin

like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents. Commun Biol 2021; 4(1): 93. http://dx.doi.org/10.1038/s42003-020-01577-x PMID: 33473151

[149] Qiao Z, Nie W, Vahdat A, Miller TF III, Anandkumar A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell 2024; 6(2): 195-208. http://dx.doi.org/10.1038/s42256-024-00792-z

[150] Shityakov S, Aglikov AS, Skorb EV, Nosonovsky M. Voronoi Entropy as a Ligand Molecular Descriptor of Protein–Ligand Interactions. ACS Omega 2023; 8(48): 46190-6. http://dx.doi.org/10.1021/acsomega.3c07328 PMID: 38075811

[151] Lee E, Yoo J, Lee H, *et al.* MetaDTA: meta-learning-based drug-target binding affinity prediction. MLDD workshop, ICLR 2022.

[152] Wang Y, Xia Y, Yan J, Yuan Y, Shen HB, Pan X. ZeroBind: a protein-specific zero-shot predictor with subgraph matching for drug-target interactions. Nat Commun 2023; 14(1): 7861. http://dx.doi.org/10.1038/s41467-023-43597-1 PMID: 38030641