Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7749988

FEATURE EXTRACTION OF DRUG–TARGET INTERACTIONS USING MODIFIED TRANSFORMER BINDING PHASE

K. Nandhini
Research Scholar, Department of Computer Science, Vels Institute of Science,Technology & Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India
kgnandhugopi@gmail.com

Dr. G. Thailambal
Associate Professor, Department of Computer Science, Vels Institute of Science,Technology & Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India, thaila.research@gmail.com

Abstract— Exploration of drug–target interactions (DTIs) requirement a financial, human and materialistic resources in conducting biomedical experimentations. In order to reduce the cost and time to meet the present needs Artificial Intelligence (AI) is introduced that helps in predicting the DTIs. With available target and drug data in conventional databases enables the machine or deep learning model a mainstream technology for DTIs. In this paper, we develop anImproved Frequent Subsequence Mining(IFSM) based transformer binding phase (TBP) for the pre-processing and extraction of features for drug-target interaction. At the initial phase, we use IFSM to extract the meaningful frequent subsequence from the input datasets that forms an intuitive pattern expression. The study aims at extraction of the instances by initially transforming the datasets to thesub-structures using Improved Frequent Subsequence Mining (IFSM).The TBP enables the extraction of semantic relations between the sub-structures extracted from previous IFSM phase, where it is of an unlabelled biomedical data. The simulation is conducted to test the efficacy of the model is tested on state-of-art DTI feature extraction models to test the efficacy of accurate contextual structural binding generation. The efficacy of the model is tested in terms of accuracy, precision, recall and f-measure.
Keywords: Drug–Target Interactions, Improved Frequent Subsequence Mining, Transformer Binding Phase

INTRODUCTION

A change in the behaviour or function of an organ is caused when medicine attaches to a specific place on that body part, known as a binding site. An FDA-approved drug or medicine is defined as any chemical molecule that changes the physiological state of the body when consumed, administered, or absorbed by the body [1].]. A biological target is a term that refers to any component of a living creature to which a medication adheres in order to cause a physiological change. Proteins and nucleic acids, for example, are examples of potential targets. Genetically engineered nuclear receptors, G-protein receptors, potassium channels, and enzymes are among the most commonly studied biological targets in scientific study. An important step in the process of identifying novel medications for biological targets is the used for predicting the DTIs [2].

During the drug chemical reaction, the drug chemical component creates temporary bonds with the target molecule. Following that, the connected drug interacts with the target, resulting in either a negative or positive modifications [3]. It is the goal of illness treatment to prevent the target from performing its intended function, which is accomplished through the use of drugs. Inhibiting the activity of these enzymes, which are known as substrates, is one method of lowering their activity. The interaction between a medicine and its target might take place in two ways [4].

The term competitive inhibitor refers to medications that attach itself with an active target site that prevents the reaction from occurring. The target recognition as a substrate is prevented by modifying the target form and structure [5]. As a result, there are no adverse reactions. The target reactions blocking helps in correcting the correct metabolic imbalances as well as to kill infections in order to treat illness.

Various approaches in traditional and reverse pharmacology, as well as wet lab investigations, can be utilised to infer the interactions between a drug and its intended target. Laboratory tests, on the other hand, are both expensive and time-consuming to conduct in order to predict the interactions between drugs. [6]. As a result, in-silico prediction of drug-target protein interactions is quite desirable. By effectively forecasting likely interactions with surprising precision, computer algorithms can reduce the amount of search space that needs to be investigated in laboratory tests [7].

Because of a variety of factors, it has become increasingly important to predict drug target interactions in the present environment. In recent time, numerous compounds are discovered and synthesised. The pharmaceutical effects and target profiles are still a mystery to researchers. The remedy for many diseases, such as Parkinson disease, lichen planus, and others, is still elusive, and new diseases are being discovered on a yearly basis [9]. Consequently, scientists have accumulated a massive quantity of data on a diverse variety of substances, including their qualities, features, and responses, as well as the proteins that they target. As a result, researchers must devise efficient models of manipulating and evaluating the intricate and high-dimensional datasets in order to succeed. This necessitates the development of more precise and sophisticated computational techniques for the prediction of DTIs [10].

Prediction of DTIs has a wide variety of applications. This strategy has been demonstrated to be effective in drug discovery, repositioning, and the prediction of side-effects [11] - [15]. Drug discovery is considered as the process of finding novel drugs with a potential to interact with a specific target. It is possible to predict in silico drug target interactions, which can aid in the discovery of drugs that tends to bind with the target. A novel medicine discovery is a time-consuming and expensive procedure that takes years to complete.

Following all of this, medications are subjected to a series of clinical trials before being approved for sale on the open market. The new molecular entity (NME) identification cost is found to be around $1.8 billion for each NME discovered. It also takes roughly a decade for newly developed pharmaceutical compounds to reach the market and be available for human use. As a result, the process of drug-discovery is time-consuming and difficult. As a result, many chemical compounds that are known in prior is not used as pharmaceuticals at this time because the interactions of these chemicals with proteins are not totally understood. Despite the fact that the PubChem database contains numerous compounds, where the majority of

interaction profiles are unknown. It is possible that new medications will be identified by in-silico analysis of these interaction profiles, which will aid in the narrowing of the search space for new drugs as well as the development of new drugs.

When it comes to interaction prediction, the in-silico strategy to adopt is dependent on a number of distinct aspects. The stage of drug development is the most important factor to consider [16]. For example, during the initial stages of research, the primary focus would be on the disease under investigation. Genes associated with disease could be discovered, or genes associated with infection and health could be distinguished. After that comes the selection of a lead chemical compound that assist in disease therapy optimisation, among other ways. Research into quantitative structure-activity relationships (QSARs) is then used to determine the pharmacological features of the lead chemical [17]. Similarly, the type of data that is available has an impact on the method that is employed. Access to medication and information about target properties are crucial considerations when deciding on a strategy.

Other medication-related concerns have also been identified that have a negative impact on the identification and prediction of drug interactions. The effects of a single medication might be extremely variable. There are numerous consequences, both positive and negative, that are difficult to trace down and quantify. It is possible that a drug will have different effects on different people, even if their genes are essentially the same. Making matters worse, the pathways in the body are extremely complicated and difficult to comprehend. That is why identifying statistically meaningful relationships among them is so difficult [18].

In this paper, we develop an Improved Frequent Subsequence Mining(IFSM) based transformer binding phase (TBP) for the pre-processing and extraction of features for DTI. At the initial phase, we use IFSM to extract the meaningful frequent subsequence from the input datasets that forms an intuitive pattern expression. The study aims at extraction of the instances by initially transforming the datasets to the sub-structures using Improved Frequent Subsequence Mining (IFSM). The TBP enables the extraction of semantic relations between the sub-structures extracted from previous IFSM phase, where it is of an unlabelled biomedical data.

RELATED WORKS

However, while there have been numerous evaluations of DTI prediction, none of this research has focused specifically on machine learning techniques. The model in [18] gives an overview of the techniques based on similarity approach for the prediction of DTI in general and in particular, the prediction of DTI is conducted using proper feature extraction procedure. Specifically, the review in [19] focuses on approaches for predicting DTIs that take into account both the structure of the drug and the sequence of the target protein.

Mousavian et al. [20] investigated it from the perspectives of supervised/semi-supervised learning. On other hand, Chen et al. [21] studied several databases, computational models and web servers have been investigated. This study covers the computational approaches of machine learning and network-based computation.

Ezzat et al. [22] investigated prediction of chemogenomic DTI using databases and methods, and they provided an empirical review of their findings. In their research, they employ five forms of chemogenomic methodology: matrix factorization, neighbourhood, network diffusion, bipartite local, and feature extraction models, among others.

Chen et al. [23] investigated the application of chemogenomic approaches to DTI prediction as part of their evaluation of machine learning algorithms and datasets. As a result, chemogenomics techniques is split into two groups based on how negative samples are handled: supervised learning approaches based on similarity and features, as well as semi-supervised learning approaches and unsupervised learning approaches.

Sachdev et al. [24] investigated the efficacy of feature-based chemogenomic strategies for DTI prediction (excluding similarity-based approaches). There are three types of strategies included in this survey: (1) SVM methods, (2) ensemble-based methods (such as decision trees or random forests), and (3) alternative techniques. Sercinoglu et al. [25] conducted a thorough investigation into drug repurposing databases.

Once the feature space has been generated, a variety of algorithms that are utilised to carry out the DTI prediction task once the feature space has been constructed. Because membrane proteins do not have three-dimensional (3D) structures, it is hard to extract the critical features that would ordinarily result in higher prediction performances.

PROPOSED METHODOLOGY

Almost all DTI feature extraction is given in Figure 1.

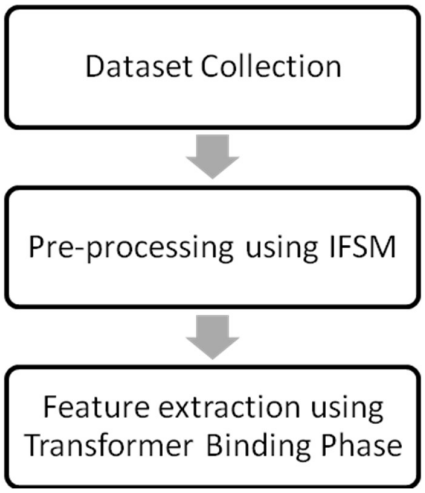

Figure 1: Proposed Method

In this model, all drug-target pairings will be represented by feature vectors of a specific length that are of binary labels and it splits the vector pairs into two different groups based on whether or not they interact positively or negatively. This is the most frequently encountered representation. Assuming the target and drug feature vectors are defined as

$F=\{f:=d\oplus t||d=[d_1,d_2,\cdots,d_n]\ \&t=[t_1,t_2,\cdots,t_m]\}$,

where

d – target feature vector with a length and

t - drug feature vectors with a length m, respectively.

B. Improved Frequent Sub-sequence Mining(IFSM)

The main aim of IFSM is to find the subsequence, which often occur typically in multiple sequences of input. This is considered to be a challenging problem that enables finding the total distance sub-sequences of a single sequence of input *T* and this is considered exponential with length for the input *T*. However, it poses problems like expensive frequent mining of sub-sequences and these sub-sequences may find limited applicability on various applications. To reduce such constraints, in this paper, we use IFSM model that focus specifically on constraints associated with the sub-sequences of DTIs. This controls the sub-sequences for mining and hence it develops IFSM to attain improved efficacy.

The constraints associated with subsequence derives the subsequences from input sequence and this is considered for IFSM. The main aim of the study is to provide an option to express the subsequence constraints. Consider a subsequence predicate*P* where the $Sup_P(S, D)$ is the P-support for a sequence $S \in \Sigma$ in database *D* is considered as the multiset over entire sequence of *D*. The frequency P for the sequence *S* in the database *D* is expressed as below:

$$f_P(S, D) = |Sup_P(S, D)|.$$

Where the entire sequencesT in the database D for $S \subseteq T$ and P (S,T) holds.

C. Transformer Binding Phase

As with proteins, the TBP begins by searching for sequence-specific motifs, separating the results into grids. Creating a compound token is accomplished by sending the fingerprints of the compounds through a series of layers that are interconnected to form a network. Once this compound token has been associated with the corresponding protein grid encoding, the process continues. A feature that are concatenated represents sequence and chemical with the size of

$$H \times (1+\lceil l_p/S_g \rceil),$$

where

S_g – Gridsize

l_p– Proteinlength

H - Hidden dimensionsize.

Positional encoding are added with the grid feature in order to help the transformers to interpret the positioning information. The compound-grid attributes that are inputs into the transformer blocks. During the continuous transformer blocks, the information is sent through to describe the whole interaction between the compound token and the target protein, as well as the interactions and selectivity to ligands, and this information is passed through to describe the total interaction.

$$(c_g, w_g, p_g) = \sigma(f_{BR}(TB_{BR}(h_g))),$$

where

h_g –grids of protein in compoundgrids feature,

$f_{BR}(\cdot)$- Dense Layers for the extraction of BR features, and

c_g – centerfor each protein grid,
σ(·) - Sigmoid Function
w_g –widthfor each protein grid,
p_g –confidence scorefor each protein grid.

DTIs extraction of features hence can be expressed as:

$$P_{DTI}=\sigma(f_{DTI}(TB_{DTI}(TB_{BR}(h_c))))$$

h_c- compound token in compound–grids feature, and
$f_{DTI}(\cdot)$ - dense layers for DTI prediction.

C. Finding Binding Region.

In order to locate the binding sites, it is necessary to first identify the assays in themolecular parts list. Findings of assays (protein or cell-based) is a technique is useful in identifying and classifying the assays in the molecular part list.The setof assays are divided into grids based on its bioactivity values that consists of a confidence score predicted over each assay grid. These confidence score involves the target type and the confidence rate that tells that the mapped target is accurate.

The prediction model helps in detection of assays by employing a similar design to that of the original model. The prediction model helps in predicting the confidence ratings for each individual. Another feature of the programme is a prediction score that ranges from 0 to 9 as in Table 1.

Table 1: Confidence score and comment

Score	Comment
0	Default Value – Unknown Target
1	Target assigned is non-molecular
2	Target assigned is subcellular fraction
3	Target assigned is non-protein molecular target
4	Target assigned is multiple homologous protein target
5	Target assigned is multiple direct protein target
6	Target assigned is homologous protein subunits complex
7	Target assigned is direct protein complex subunits
8	Target assigned is homologous single protein target
9	Target assigned is direct single protein

These are then reconstructed to their original values and saved back into the database for future use.

$C_g=s_g+S_g \cdot c_g$,

where

c_g - BR locationprediction, and
s_g - gridstarting index.
C_g –Feature Extracted BR location.

The widths of BRs (W_{ig}) is hence represented as:

$W_{ig}=r_i e w_g$,

where

w_g - BR widthprediction, and
r_i - predefined BR width,
e - Euler number.
W_{ig} - predicted BRwidth.

The focal loss can be used to dynamically modify object classification weights and prevent class imbalances by adjusting the weights of individual objects. The following are the weights of the focus loss used to address the detection of the BR class imbalance:

$$FL(p_t) = (p_t - 1)^{\gamma} \log(p_t)$$

where,

$$p_t = \begin{cases} p & if\ y = 1 \\ (1-p) & if\ y = 0 \end{cases}$$

where loss weights are regulated dynamically to reduce the rate of class imbalance. When calculating the BR centre and width loss, the mean absolute error is employed. To determine the total loss associated with the prediction of BR, the following formula is used:

$L(c,w,p)=\lambda_{reg}(|c_t-c_p|_1+|w_t-w_p|_1)+\lambda_c FL(p_t)$,

where

$|\ |_1$ - L1 loss,
λ_{reg} -regression loss weights
λ_c - focal lossweights

EXPERIMENTAL RESULTS

A. Specification of Dataset

The information in the ChEMBL database was gathered from published literature and entered into the database by the database administrators. The EMBL-European Bioinformatics Institute published these records in 2002, and they are available online (EMBL-EBI). Since its inception in 1998, this database has contained more than 1.9 million chemical compounds. ChEMBL has information on more than 10,000 drugs as well as more than 12,000 targets.

B. Results

Metrics can be used to compare and contrast different techniques. They assist in the comparison of numerous ways in order to determine which is the most appropriate for implementation.

Accuracy: An accuracy is assessed by the percentage of interactions that it correctly predicts, and its precision is measured by the percentage of interactions that it correctly predicts. Precision is measured by the percentage of interactions that it correctly predicts. Calculate the precision using the following formula:

$$Acc = (TN+TP)/(TN+TP+FN+FP)$$

Recall: The term recall refers to the process of effectively identifying a successful interaction. It can be deduced in the following way:

$$Recall = (TP)/(TP+FN)$$

Precision: Precision is a term that is used to describe an additional usual evaluation metric.

$$Precision = (TP)/(TP+FP)$$

Time: When analysing and comparing alternative techniques, the training and prediction times for various classifiers can be used as a metric to evaluate and compare them.
where

TP - True Positive,
FP - False Positive,
FN - False Negative, and
TN -True Negative.

Instead of FP, the non-interactive drug target pair, TP is projected to interact with the interacting drug target pair, whereas FP does not interact with the drug target pair (of non-interacting one) (TP). The non-interactive pairs of drug target that are predicted for not interacting and hence it is denoted as the TN and FN, while the interacting pairs of drug target and that gets predicted for interaction, where it is denoted by FN and TN.

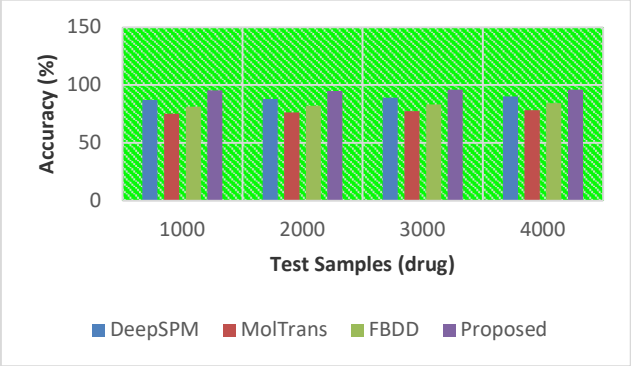

Figure 2 Accuracy

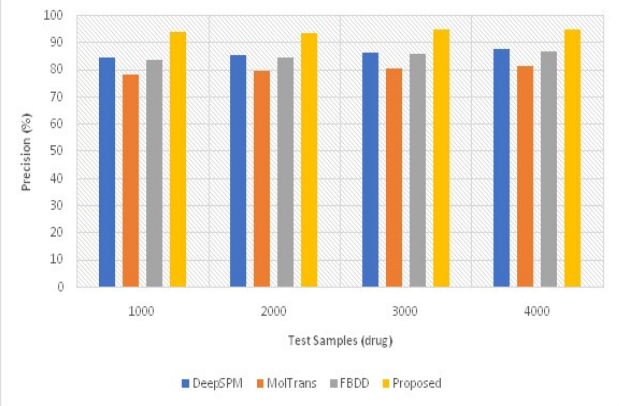

Figure 3 Precision

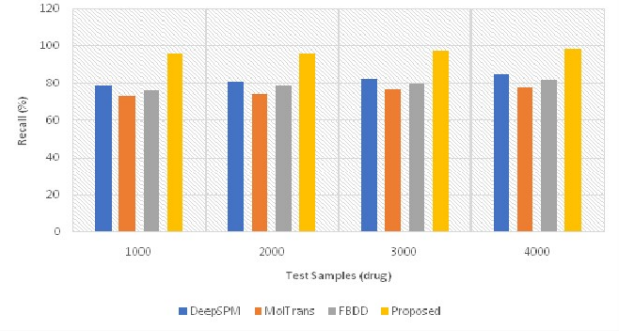

Figure 4Recall

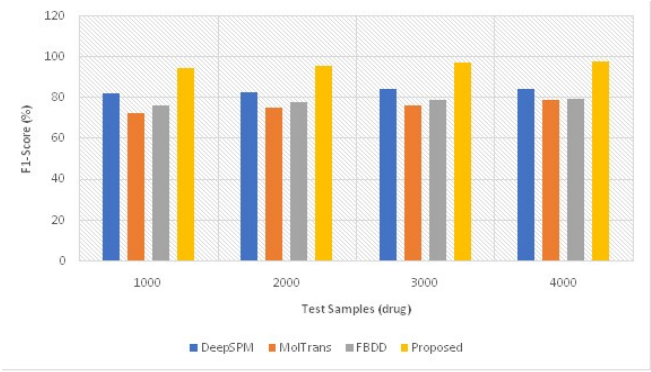

Figure 5 F1 Measure

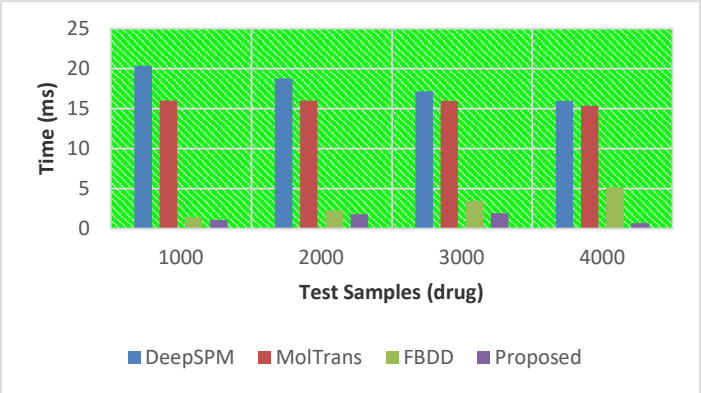

Figure 6 Computational Time

A second test of dependability was performed in this experiment, in which the proposed method was applied to the prediction of drug-target interactions on four different datasets using the five-fold cross-validation method. By using this strategy, each of the model is trained on a different dataset. Finally, the feature vectors from the testing datasets were fed into the trained prediction model in order to create a prediction score (Figure 2 – Figure 6) that measures the likelihood that a specific drug-target pair will interact with each other.

Prediction of DTIs can be accomplished through the use of the proposed feature extraction method, allowing for further examination of the method prediction outputs. In this experiment, a five-fold cross validation procedure was used to compare the accuracy of the prediction with existing state-of-the-art methodologies. This is demonstrated by a comparison of the feature extraction results obtained from different approaches using the same feature descriptor. The prediction performance of the proposed feature extraction method is significantly better than that of the existing method.

Each substructure of a molecule is represented by a single bit in the binary vectors used to represent it. Drug molecules that include specific substructures can be identified by employing substructure fingerprints, which encode structural information about a given medicinal component into a series of binary bits and can be used to identify the presence of a specific substructure.

CONCLUSION

In this study, a TBP based on the IFSM was created for the preprocessing and extraction of DTI aspects from clinical data. We use the IFSM to extract the relevant frequent subsequences that may be combined to generate a comprehensible pattern expression from the input data set. It is necessary to restructure the datasets into sub-structures using the IFSM before extracting any instances from them in this investigation. TBP can be used to extract semantic associations between sub-structures created from the earlier IFSM phase from unlabelled biological data using a variety of techniques. The model is assessed using state-of-the-art DTI feature extraction algorithms in order to determine its capacity to create contextual structural binding with high accuracy and consistency. It is necessary to put the model accuracy, precision, recall, and f-measure through their paces.

References

[1] Hu, F., Hu, Y., Zhang, J., Wang, D., & Yin, P. (2020, December). Structure Enhanced Protein-Drug Interaction Prediction using Transformer and Graph Embedding. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1010-1014). IEEE.

[2] Kim, Y., & Shin, B. (2021, October). An Interpretable Framework for Drug-Target Interaction with Gated Cross Attention. In *Machine Learning for Healthcare Conference* (pp. 337-353). PMLR.

[3] Zollanvari, A., Kunanbayev, K., Bitaghsir, S. A., & Bagheri, M. (2020). Transformer fault prognosis using deep recurrent neural network over vibration signals. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-11.

[4] de Souza, J. G., Fernandes, M. A., & de Melo Barbosa, R. (2022). A Novel Deep Neural Network Technique for Drug–Target Interaction. *Pharmaceutics*, *14*(3), 625.

[5] Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., ... & Zheng, M. (2020). TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, *36*(16), 4406-4414.

[6] Chen, C., Zhang, Y., Chen, Z., Yang, H., & Gu, Z. (2021). Cellular transformers for targeted therapy. *Advanced drug delivery reviews*, *179*, 114032.

[7] Kalakoti, Y., Yadav, S., & Sundar, D. (2022). TransDTI: Transformer-Based Language Models for Estimating DTIs and Building a Drug Recommendation Workflow. *ACS Omega*.

[8] Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H., & Yang, Y. (2022). AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Briefings in Bioinformatics*.

[9] Yang, L., Yang, G., Bing, Z., Tian, Y., Niu, Y., Huang, L., & Yang, L. (2021). Transformer-Based Generative Model Accelerating the Development of Novel BRAF Inhibitors. *ACS omega*, *6*(49), 33864-33873.

[10] Du, B. X., Qin, Y., Jiang, Y. F., Xu, Y., Yiu, S. M., Yu, H., & Shi, J. Y. (2022). Compound–protein interaction prediction by deep learning: databases, descriptors and models. *Drug Discovery Today*.

[11] Chen, C., Qiu, Z., Yang, Z., Yu, B., & Cui, X. (2021, December). Jointly Learning to Align and Aggregate with Cross Attention Pooling for Peptide-MHC Class I Binding Prediction.

In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 18-23). IEEE.

[12] Rani, P., Dutta, K., & Kumar, V. (2022). Artificial intelligence techniques for prediction of drug synergy in malignant diseases: Past, present, and future. *Computers in Biology and Medicine*, 105334.

[13] Hu, F., Wang, L., Hu, Y., Wang, D., Wang, W., Jiang, J., ... & Yin, P. (2021). A novel framework integrating AI model and enzymological experiments promotes identification of SARS-CoV-2 3CL protease inhibitors and activity-based probe. *Briefings in bioinformatics*, *22*(6), bbab301.

[14] Liu, Q., & Xie, L. (2021). TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS computational biology*, *17*(2), e1008653.

[15] Wang, X., Zhang, Z., Zhang, C., Meng, X., Shi, X., & Qu, P. (2022). TransPhos: A Deep-Learning Model for General Phosphorylation Site Prediction Based on Transformer-Encoder Architecture. *International Journal of Molecular Sciences*, *23*(8), 4263.

[16] Abdel-Basset, M., Hawash, H., Elhoseny, M., Chakrabortty, R. K., & Ryan, M. (2020). DeepH-DTA: deep learning for predicting drug-target interactions: a case study of COVID-19 drug repurposing. *Ieee Access*, *8*, 170433-170451.

[17] Lee, C. Y., & Chen, Y. P. P. (2021). New Insights Into Drug Repurposing for COVID-19 Using Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

[18] Ding, H., Takigawa, I., Mamitsuka, H., & Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in bioinformatics*, *15*(5), 734-747.

[19] Yamanishi, Y. (2013). Chemogenomic approaches to infer drug–target interaction networks. *Data Mining for Systems Biology*, 97-113.

[20] Mousavian, Z., & Masoudi-Nejad, A. (2014). Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert opinion on drug metabolism & toxicology*, *10*(9), 1273-1287.

[21] Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*, *17*(4), 696-712.

[22] Ezzat, A., Wu, M., Li, X. L., & Kwoh, C. K. (2019). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, *20*(4), 1337-1357.

[23] Chen, R., Liu, X., Jin, S., Lin, J., & Liu, J. (2018). Machine learning for drug-target interaction prediction. *Molecules*, *23*(9), 2208.

[24] Sachdev, K., & Gupta, M. K. (2019). A comprehensive review of feature based methods for drug target interaction prediction. *Journal of biomedical informatics*, *93*, 103159.

[25] Serçinoğlu, O., & Sarica, P. O. (2019). In Silico Drug Design.