



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Drug Fusion Recommendation Analysis Using CGEFA & GCN

Dr D J Samatha Naidu, K.Venkata Ramya, P.Samitha Reddy

Department of MCA, Annamacharya PG College of Computer Studies, Rajampet, Andhra Pradesh, India

Department of MCA, Annamacharya PG College of Computer Studies, Rajampet, Andhra Pradesh, India

Department of MCA, Annamacharya PG College of Computer Studies, Rajampet, Andhra Pradesh, India

ABSTRACT: PREDICTING drug-target binding affinity (DTA prediction) is crucial in new drug development as well as drug repurposing. The gold standard to determine the binding affinity is by experimental assays but this is prohibitively expensive as a rapid screening tool as there are over 100 million drug-like compounds and over 5000 potential protein targets. Therefore, it is necessary to have alternative computational methods using simulation or machine learning to predict the binding affinity of novel drug-target pairs. Machine learning methods are particularly attractive because they offer cheap and fast alternatives with reasonable performance thanks to the large DTA databases that we can leverage on.

With the advance of machine learning, many computational prediction methods [5]–[8] have been proposed to tackling DTA. In recent existing works, the protein is typically represented as a string of amino acids denoted by letters [7]–[9]. The drawback of using protein sequence is that it can not represent the 3D structure of the protein which is crucial information for determining the binding affinity between protein and drug in practice. However, obtaining the high-resolution 3D structure is a challenging task. A more practical solution is using the 2D pairwise distance or contact maps to represent tertiary protein structure. These maps can now be determined with reasonable accuracy from deep learning powered algorithms.

In proposed work, Designed a novel deep learning method, called GEFA (Graph Early Fusion for binding Affinity prediction) for target-drug affinity prediction, a crucial task for rapid virtual drug screening and drug repurposing. To improve the power of protein representation, we use self-supervised to take advantage of a large amount of unlabeled target sequences. To address the latent representation change due to conformation change during the binding process, the early fusion between drug and target is proposed. Unlike the late fusion approach extracting representation separately, the early fusion approach integrates drug representation info into protein representation learning phase. The self-attention value of the target sequence is used as edge weight connecting drug node and residue node in the target protein graph. Self-attention allows the model more interpretable as it shows which residues contribute to the binding process and the level of contribution of each residue. The quantitative experiments show that the early fusion approach has advantages over the late fusion approach. Using the embedding feature as target node feature has advantages over using one-hot encoding. Residual block design allows stacking multiple GCN layers for better learning representation capability.

KEYWORDS: Drug-target binding affinity, Graph neural network, Early fusion, Representation change

I. INTRODUCTION

Computational with the use of image analysis techniques and machine learning representations of data at several degrees of complexity made possible by deep learning. These methods have considerably improved the state of the art in many other domains, including object identification, visual object recognition, voice recognition, connecting DNA to drug discovery, and object identification. By employing a reverse propagation strategy to suggest changes to a tool's internal parameters that are used to define the model in each layer from the representation in the preceding layer, it may uncover sophisticated architectures in enormous data sets. Recurring nets have shed more insight into certain data categories including speech and text whereas deep convolutional networks have improved at analyzing pictures, video, voice, and audio. selecting relevant search results, identifying objects in pictures, text-to-speech conversion, matching news stories, posts, or items with users' interests, and more. Such applications employ more and more of the deep training approach. The potential of conventional Data that was natural in its raw form could only be analyzed using limited machine learning techniques. Over the years, developing a feature extractor that transformed the raw data (such as the pixel Building pattern recognition or artificial intelligence system requires understanding the values of a picture

system. This needed thorough design and in-depth subject-matter expertise. The knowledge Often referred to as the classifier, the subsystem may identify or categorize patterns in the data input into an appropriate internal representation or feature vector.

II. RELATED WORK

Drug Re-purposing as an Alternative Medication for Novel Disease Drug re-purposing [18] is the process of identifying well-established medications for the novel target disease. The advantages of this drug re-purposing over developing a completely novel drug are lower risk and fast-track development [19]. The process of drug re-purposing consists of three key steps: identifying the candidate molecules given the target disease, drug effect assessment in the preclinical trial, and effectiveness assessment in clinical trial [20]. The first step, hypothesis generation, is critical as it decides the success of the whole process. Advanced computational approaches are used for hypothesis generation. Computational approaches in drug re-purposing can be categorized into six groups [20]: genetic association [21], [22], pathway pathing [23]–[25], retrospective clinical analysis [26]–[28], novel data sources, signature matching [29]–[31], molecular docking [32]–[34].

III. LITERATURE SURVEY

[1] M. Thafar, A. B. Raies, S. Albaradei, M. Essack, and V. B. Bajic, “Comparison study of computational prediction tools for drug–target binding affinities,” *Frontiers in Chemistry*, vol. 7, 2019.

The drug development is generally arduous, costly, and success rates are low. Thus, the identification of drug–target interactions (DTIs) has become a crucial step in early stages of drug discovery. Consequently, developing computational approaches capable of identifying potential DTIs with minimum error rate are increasingly being pursued. These computational approaches aim to narrow down the search space for novel DTIs and shed light on drug functioning context. Most methods developed to date use binary classification to predict if the interaction between a drug and its target exists or not. However, it is more informative but also more challenging to predict the strength of the binding between a drug and its target. If that strength is not sufficiently strong, such DTI may not be useful. Therefore, the methods developed to predict drug–target binding affinities (DTBA) are of great value. In this study, we provide a comprehensive overview of the existing methods that predict DTBA. We focus on the methods developed using artificial intelligence (AI), machine learning (ML), and deep learning (DL) approaches, as well as related benchmark datasets and databases. Furthermore, guidance and recommendations are provided that cover the gaps and directions of the upcoming work in this research area. To the best of our knowledge, this is the first comprehensive comparison analysis of tools focused on DTBA with reference to AI/ML/DL.

[2] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, “Drug–target interaction prediction: databases, web servers and computational models,” *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 696–712, Jul. 2016.

Identification of drug–target interactions is an important process in drug discovery. Although high-throughput screening and other biological assays are becoming available, experimental methods for drug–target interaction identification remain to be extremely costly, time-consuming and challenging even nowadays. Therefore, various computational models have been developed to predict potential drug–target associations on a large scale. In this review, databases and web servers involved in drug–target identification and drug discovery are summarized. In addition, we mainly introduced some state-of-the-art computational models for drug–target interactions prediction, including network-based method, machine learning-based method and so on. Specially, for the machine learning-based method, much attention was paid to supervised and semi-supervised models, which have essential difference in the adoption of negative samples. Although significant improvements for drug–target interaction prediction have been obtained by many effective computational models, both network-based and machine learning-based methods have their disadvantages, respectively. Furthermore, we discuss the future directions of the network-based drug discovery and network approach for personalized drug discovery based on personalized medicine, genome sequencing, tumor clone-based network and cancer hallmark-based network. Finally, we discussed the new evaluation validation framework and the formulation of drug–target interactions prediction problem by more realistic regression formulation based on quantitative bioactivity data.

[3] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al., “PubChem 2019 update: improved access to chemical data,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

PubChem (<https://pubchem.ncbi.nlm.nih.gov>) is a key chemical information resource for the biomedical research community. Substantial improvements were made in the past few years. New data content was added, including spectral information, scientific articles mentioning chemicals, and information for food and agricultural chemicals. PubChem released new web interfaces, such as PubChem Target View page, Sources page, Bioactivity dyad pages and Patent View page. PubChem also

released a major update to PubChem Widgets and introduced a new programmatic access interface, called PUG-View. This paper describes these new developments in PubChem

[4] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1045–D1053, 2016.

BindingDB, www.bindingdb.org, is a publicly accessible database of experimental protein-small molecule interaction data. Its collection of over a million data entries derives primarily from scientific articles and, increasingly, US patents. BindingDB provides many ways to browse and search for data of interest, including an advanced search tool, which can cross searches of multiple query types, including text, chemical structure, protein sequence and numerical affinities. The PDB and PubMed provide links to data in BindingDB, and vice versa; and BindingDB provides links to pathway information, the ZINC catalog of available compounds, and other resources. The BindingDB website offers specialized tools that take advantage of its large data collection, including ones to generate hypotheses for the protein targets bound by a bioactive compound, and for the compounds bound by a new protein of known sequence; and virtual compound screening by maximal chemical similarity, binary kernel discrimination, and support vector machine methods. Specialized data sets are also available, such as binding data for hundreds of congeneric series of ligands, drawn from BindingDB and organized for use in validating drug design methods. BindingDB offers several forms of programmatic access, and comes with extensive background material and documentation. Here, we provide the first update of BindingDB since 2007, focusing on new and unique features and highlighting directions of importance to the field as a whole.

[5] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, and T. Aittokallio, "Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors," *PLOS Computational Biology*, vol. 13, no. 8, e1005678, 2017.

Due to relatively high costs and labor required for experimental profiling of the full target space of chemical compounds, various machine learning models have been proposed as cost-effective means to advance this process in terms of predicting the most potent compound-target interactions for subsequent verification. However, most of the model predictions lack direct experimental validation in the laboratory, making their practical benefits for drug discovery or repurposing applications largely unknown. Here, we therefore introduce and carefully test a systematic computational-experimental framework for the prediction and pre-clinical verification of drug-target interactions using a well-established kernel-based regression algorithm as the prediction model. To evaluate its performance, we first predicted unmeasured binding affinities in a large-scale kinase inhibitor profiling study, and then experimentally tested 100 compound-kinase pairs. The relatively high correlation of 0.77 ($p < 0.0001$) between the predicted and measured bioactivities supports the potential of the model for filling the experimental gaps in existing compound-target interaction maps. Further, we subjected the model to a more challenging task of predicting target interactions for such a new candidate drug compound that lacks prior binding profile information. As a specific case study, we used tivozanib, an investigational VEGF receptor inhibitor with currently unknown off-target profile. Among 7 kinases with high predicted affinity, we experimentally validated 4 new off-targets of tivozanib, namely the Src-family kinases FRK and FYN A, the non-receptor tyrosine kinase ABL1, and the serine/threonine kinase SLK. Our subsequent experimental validation protocol effectively avoids any possible information leakage between the training and validation data, and therefore enables rigorous model validation for practical applications. These results demonstrate that the kernel-based modeling approach offers practical benefits for probing novel insights into the mode of action of investigational compounds, and for the identification of new target selectivities for drug repurposing applications

IV. PROPOSED ALGORITHM

Gradient boosting

Gradient boosting is a versatile machine learning technique employed in regression and classification tasks, among others. It constructs a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When using decision trees as the weak learner, the resulting algorithm is referred to as gradient-boosted trees, often surpassing the performance of random forests. The construction of a gradient-boosted trees model occurs in a stage-wise manner, similar to other boosting methods, but it stands out by allowing the optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a straightforward yet highly effective classification algorithm that operates based on a similarity measure. It is non-parametric and employs lazy learning, meaning it does not "learn" until presented with a test example. Whenever there is a new data point to classify, KNN identifies its K-nearest neighbors from the training data and determines its classification based on their majority vote or weighted vote.

Logistic regression

Logistic regression analysis explores the relationship between a categorical dependent variable and a set of independent variables. The term "logistic regression" is applied when the dependent variable has only two values, such as 0 and 1, or Yes and No. On the other hand, "multinomial logistic regression" is used when the dependent variable has three or more unique values, like Married, Single, Divorced, or Widowed. While the nature of data for the dependent variable differs from that of multiple regressions, the practical application of the procedure remains similar.

Logistic regression serves as a competitor to discriminant analysis in analyzing categorical-response variables. Many statisticians favor logistic regression due to its versatility and suitability for modeling various situations compared to discriminant analysis. This preference arises from logistic regression's ability to not assume that the independent variables follow a normal distribution, unlike discriminant analysis.

NAIVE BAYES

The naive Bayes approach is a supervised learning method founded on a simple assumption: it presumes that the presence or absence of one feature of a class is independent of the presence or absence of any other feature. Despite its simplicity, it demonstrates robustness and efficiency comparable to other supervised learning techniques. One explanation often highlighted in the literature is based on representation bias.

The naive Bayes classifier operates as a linear classifier, akin to linear discriminant analysis, logistic regression, or linear support vector machines (SVMs). However, the distinction lies in the method used to estimate the classifier's parameters, known as the learning bias. Although the naive Bayes classifier finds extensive use in the research community due to its ease of programming, parameter estimation simplicity, rapid learning even with large datasets, and reasonably good accuracy compared to other methods, it remains less popular among practitioners seeking practical results. Researchers appreciate its simplicity and efficacy. However, practitioners often struggle with its interpretability and deployment, as they may not grasp its relevance or utility.

Random forests

Random forests, also known as random decision forests, represent an ensemble learning technique used for classification, regression, and other tasks. They function by constructing numerous decision trees during training. For classification tasks, the output of the random forest is determined by the class selected by the majority of trees. Conversely, for regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests aim to mitigate the issue of decision trees overfitting to their training set.

In general, random forests tend to outperform individual decision trees, although they may have lower accuracy compared to gradient boosted trees. Nonetheless, the performance of random forests can be influenced by the characteristics of the data.

The concept of random decision forests was first introduced in 1995 by Tin Kam Ho, who utilized the random subspace method. This method, as formulated by Ho, serves as an implementation of the "stochastic discrimination" approach to classification initially proposed by Eugene Kleinberg.

Support Vector Machine (SVM)

Support Vector Machine (SVM) represents a discriminant machine learning technique commonly used in classification tasks. It aims to find a discriminant function, based on an independently and identically distributed training dataset, that accurately predicts labels for newly acquired instances. Unlike generative machine learning approaches, which necessitate computations of conditional probability distributions, a discriminant classification function assigns a data point x to one of the classes involved in the classification task. Compared to generative approaches, discriminant methods may be less powerful, particularly in outlier detection scenarios. However, they require fewer computational resources and less training data, especially in multidimensional feature spaces and when only posterior probabilities are necessary. Geometrically, learning a classifier equates to identifying the equation for a multidimensional surface that optimally separates the different classes in the feature space.

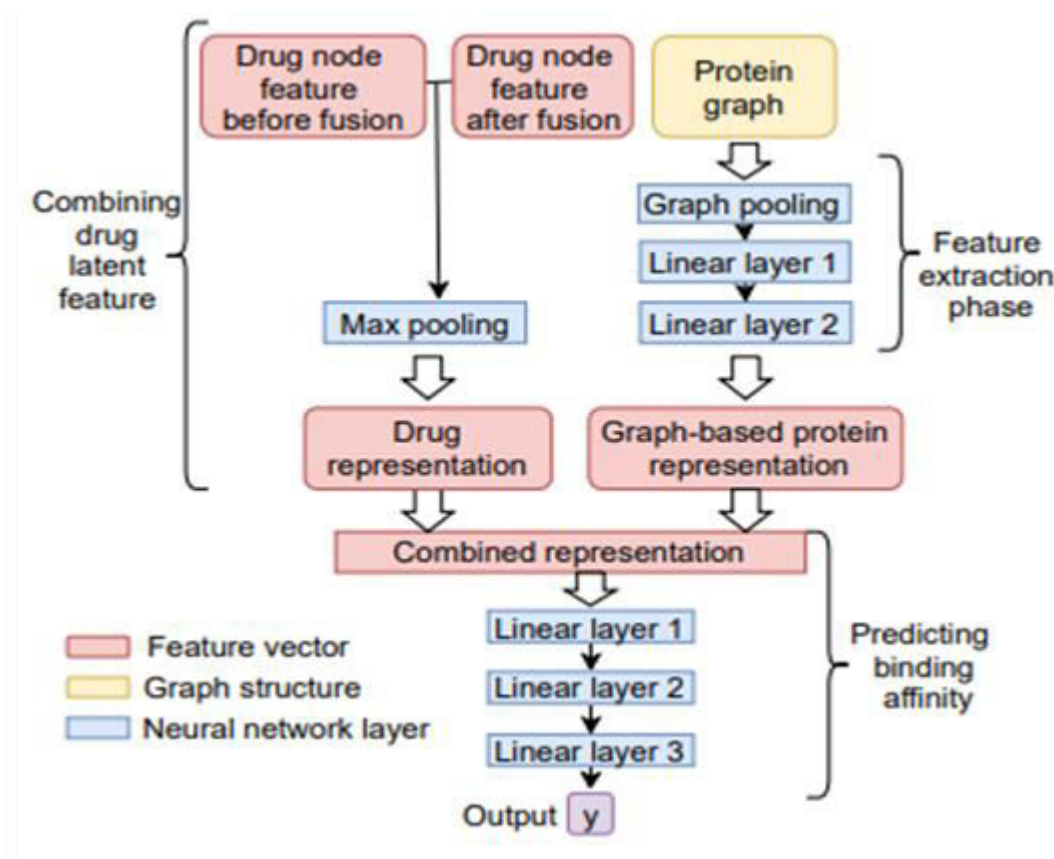
SVM is a discriminant technique that solves convex optimization problems analytically, consistently yielding the same optimal hyperplane parameters. In contrast, genetic algorithms (GAs) and perceptrons, both widely used for

classification in machine learning, may produce solutions highly dependent on initialization and termination criteria. With a specific kernel transforming data from the input space to the feature space, SVM training returns uniquely defined model parameters for a given training set, whereas perceptron and GA classifier models vary with each training iteration.

2.3 PROPOSED MODULES

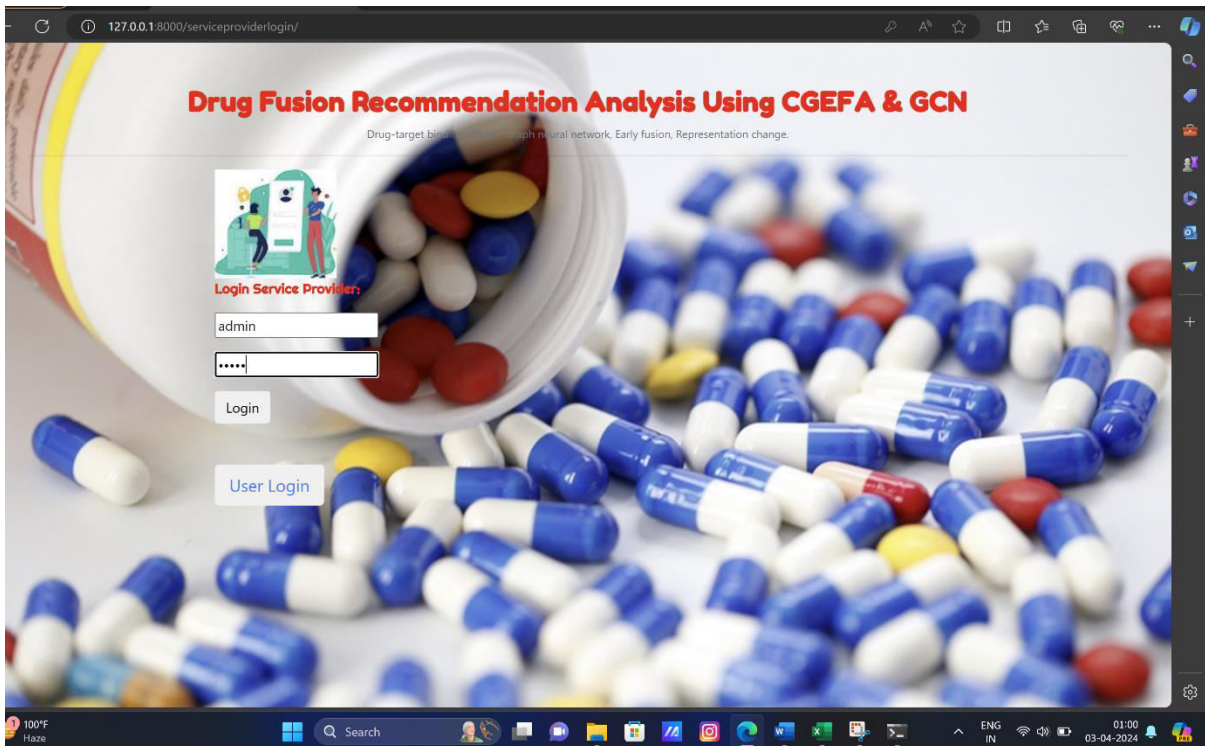
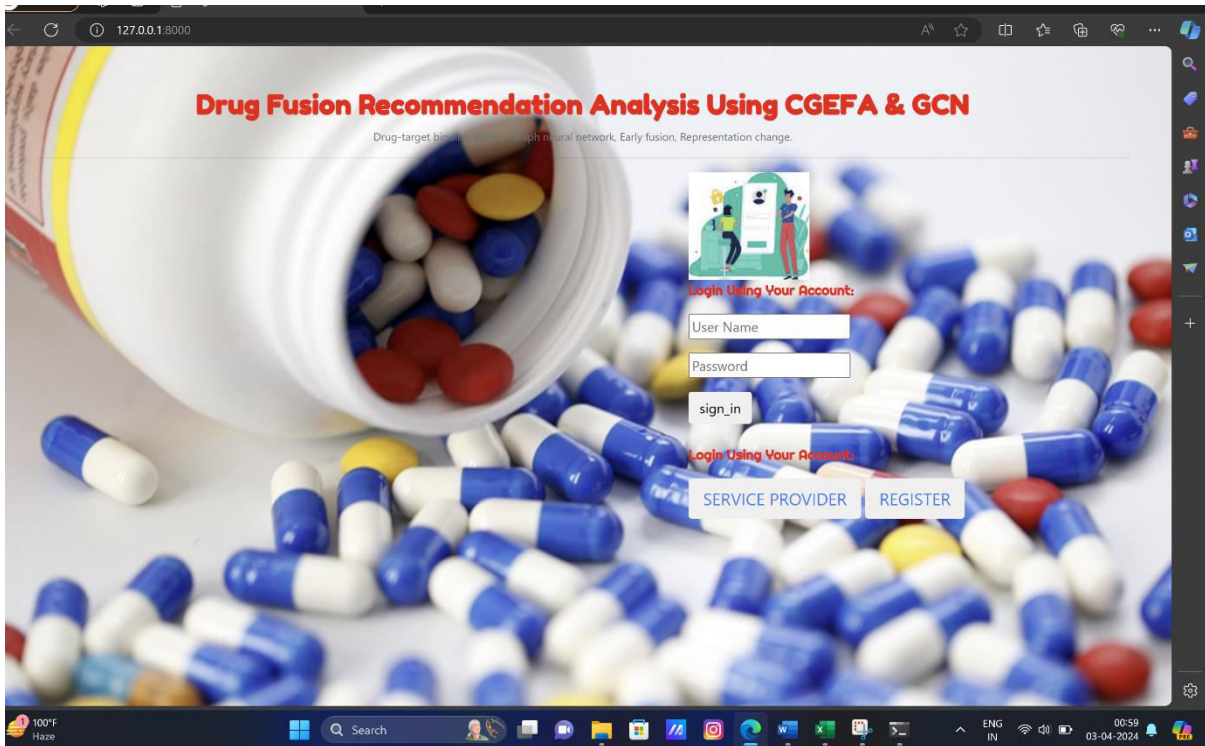
2.3.1 Service Provider

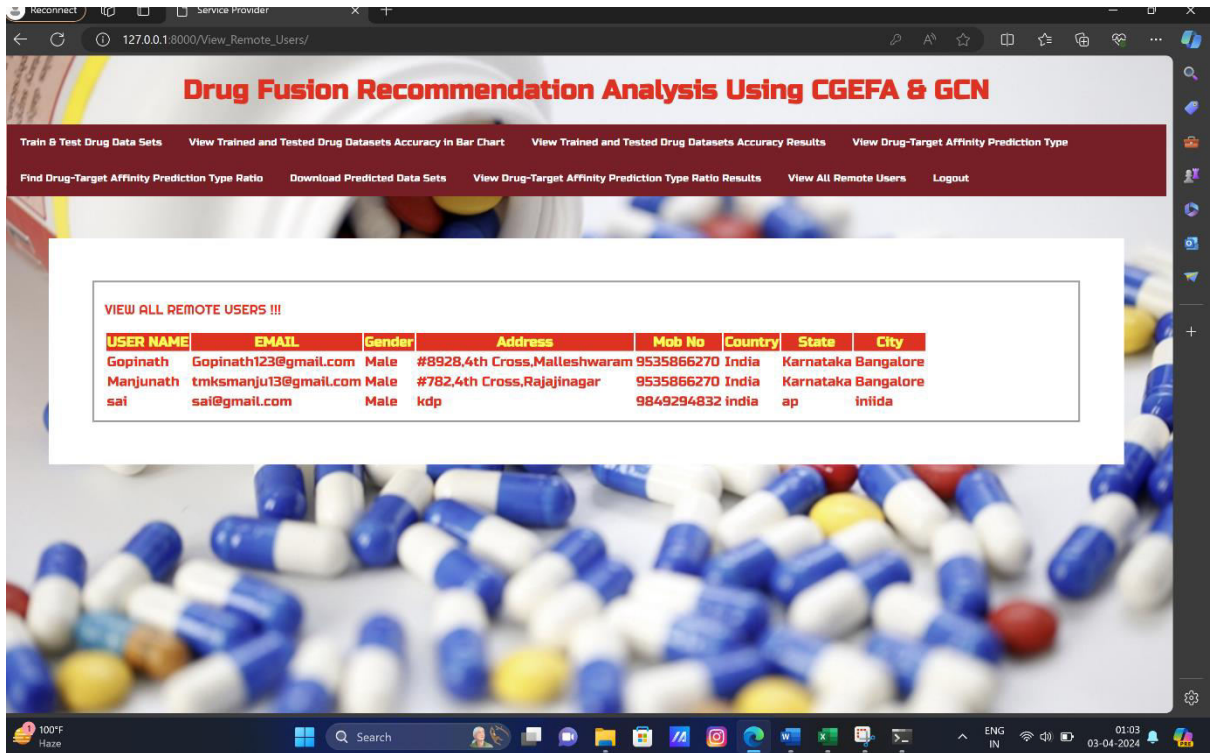
SYSTEM ARCHITECTURE



V. SIMULATION RESULTS

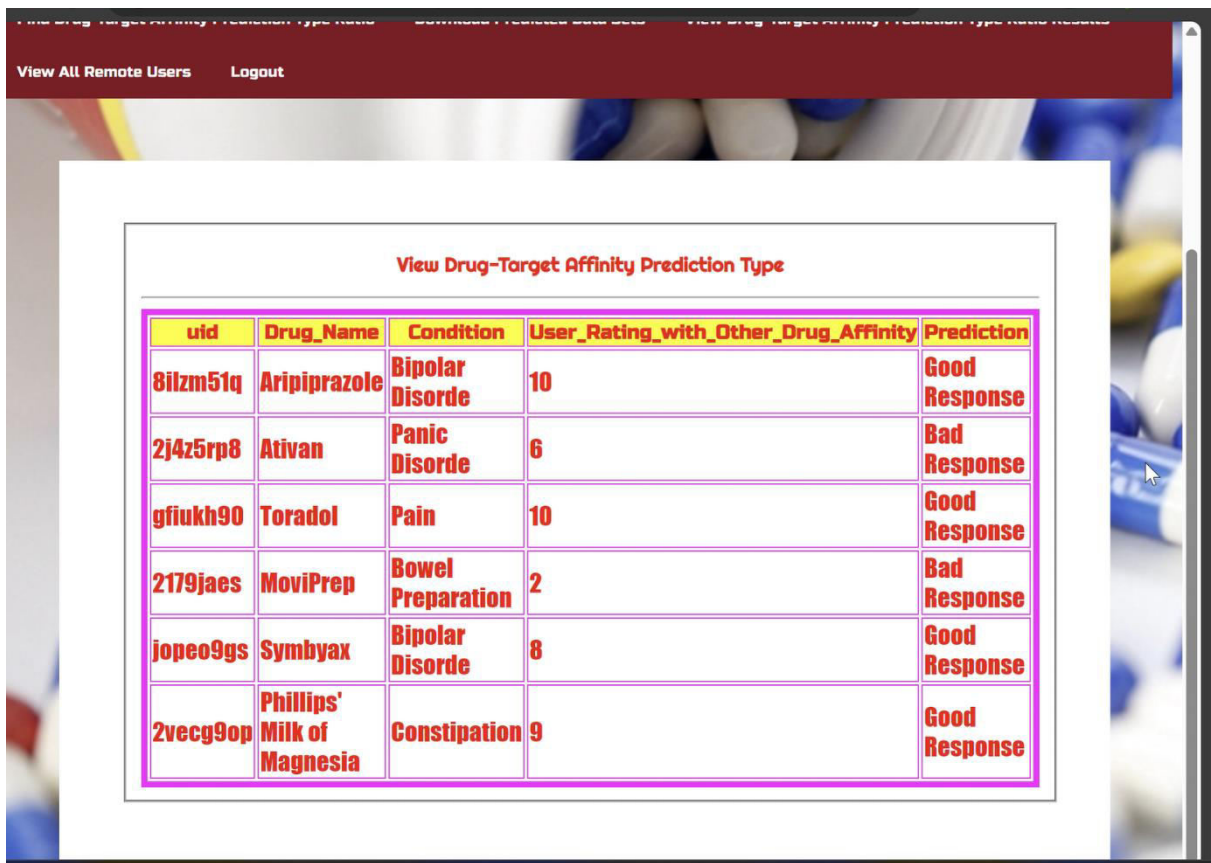
We report our late fusion approach, GLFA, and early fusion approach, GEFA, with previous works in Davis benchmark on four settings in Table 1 and in PDBBind dataset (Table 2). Our proposed method GEFA consistently outperforms previous works in four settings of Davis dataset and generalrefined setting of PDBBind dataset. Our proposed methods achieve state-of-the-art performance across all four settings. Between two late fusion based methods DGraphDTA [15] and GLFA, our proposed GLFA method also outperforms DGraphDTA. This follows our expectations as the embedding feature contains richer information than one-hot encoding and PSSM. This also demonstrates the advantage of using the residual block. DGraphDTA [15], GLFA, and GEFA outperform GINConvNet in all four settings. GINConvNet and GCNConvNet [8] only use sequence and CNN to learn the target representation. On the other hand, DgraphDTA [15], GLFA, and GEFA use the graph built from the protein contact map and learn the target representation using GCN. This demonstrates the advantage of using the graph representation of the contact map.





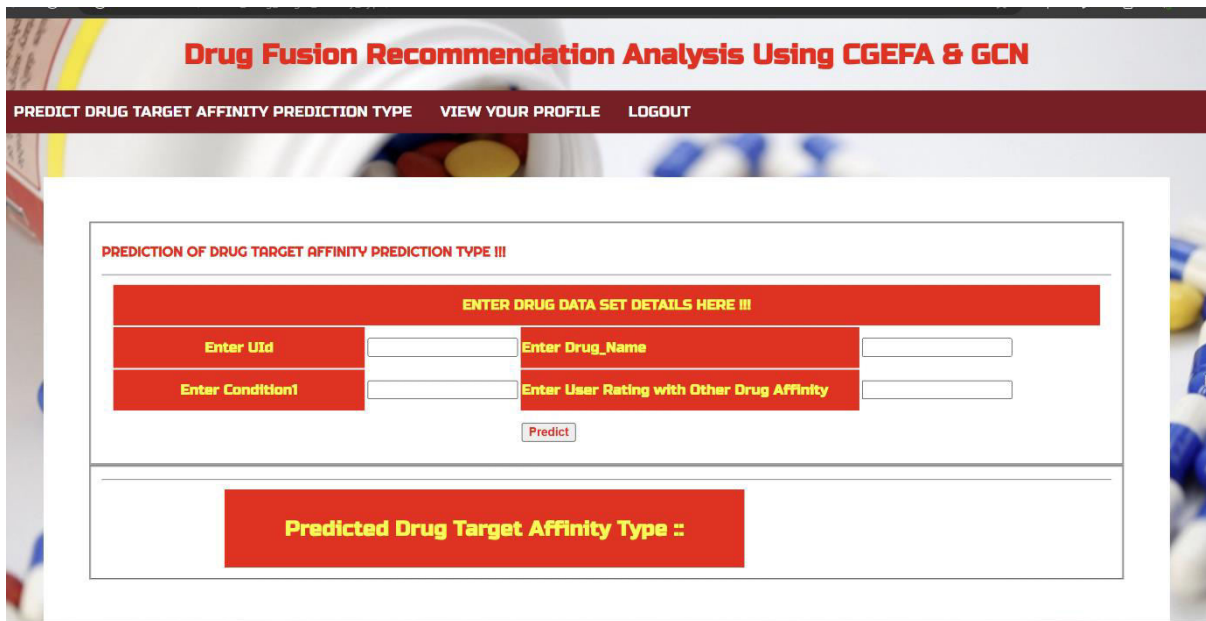
VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Gopinath	Gopinath123@gmail.com	Male	#8928,4th Cross,Malleshwaram	9535866270	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	Male	#782,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
sai	sai@gmail.com	Male	kdp	9849294832	india	ap	intida



View Drug-Target Affinity Prediction Type

uid	Drug_Name	Condition	User_Rating_with_Other_Drug_Affinity	Prediction
8ilzm51q	Aripiprazole	Bipolar Disorder	10	Good Response
2j4z5rp8	Ativan	Panic Disorder	6	Bad Response
gfiukh90	Toradol	Pain	10	Good Response
2179jaes	MoviPrep	Bowel Preparation	2	Bad Response
jopeo9gs	Symbyax	Bipolar Disorder	8	Good Response
2vecg9op	Phillips' Milk of Magnesia	Constipation	9	Good Response



```

C:\Windows\System32\cmd.e  X  +  v
      0      0.50      0.00      0.00      3624
      1      0.60      1.00      0.75      5534

accuracy
macro avg      0.55      0.50      0.38      9158
weighted avg   0.56      0.60      0.46      9158

CONFUSION MATRIX
[[ 1 3623]
 [ 1 5533]]
Logistic Regression
ACCURACY
60.42804105699935
CLASSIFICATION REPORT
      precision      recall      f1-score      support
      0      0.50      0.00      0.00      3624
      1      0.60      1.00      0.75      5534

accuracy
macro avg      0.55      0.50      0.38      9158
weighted avg   0.56      0.60      0.46      9158

CONFUSION MATRIX
[[ 1 3623]
 [ 1 5533]]
Good Response
    
```

VI. CONCLUSION AND FUTURE WORK

We have proposed a novel deep learning method, called GEFA (Graph Early Fusion for binding Affinity prediction) for target-drug affinity prediction, a crucial task for rapid virtual drug screening and drug repurposing. To improve the power of protein representation, we use self-supervised to take advantage of a large amount of unlabeled target sequences. To address the latent representation change due to conformation change during the binding process, the early fusion between drug and target is proposed. Unlike the late fusion approach extracting representation separately, the early fusion approach integrates drug representation info into protein representation learning phase. The self-attention value of the target sequence is used as edge weight connecting drug node and residue node in the target protein graph. Self-attention allows the model more interpretable as it shows which residues contribute to the binding process and the level of contribution of each residue. The quantitative experiments show that the early fusion approach has advantages over the late fusion approach. Using the embedding feature as target node feature has advantages over using one-hot encoding. Residual block design allows stacking multiple GCN layers for better learning representation capability. If we can learn the edge change, we can express the conformation change caused by the drug-target binding.

In addition, in case the target protein has multiple binding pocket at different regions, the drug molecule may only bind at one pocket at one time. However, in our model, the drug node links to all possible binding sites indicated by self-attention mask. The binding process modeling will be more accurate if we can combine drug info into the finding drug-residues edges process. Our framework can be applied for RNA with binding sites capable of binding drug-like molecules. However, in case of drugs binding to the secondary structure of RNA, the binding mechanism can be different and the target graph may require modifications to represent the secondary structure interaction.

REFERENCES

- 1) M. Thafar, A. B. Raies, S. Albaradei, M. Essack, and V. B. Bajic, "Comparison study of computational prediction tools for drug-target binding affinities," *Frontiers in Chemistry*, vol. 7, 2019.
- 2) X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug-target interaction prediction: databases, web servers and computational models," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 696–712, Jul. 2016.
- 3) S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al., "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1102–D1109, 2019.
- 4) M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1045–D1053, 2016.
- 5) A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, and T. Aittokallio, "Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors," *PLOS Computational Biology*, vol. 13, no. 8, e1005678, 2017.
- 6) A. Cichonska, T. Pahikkala, S. Szedmak, H. Julkunen, A. Airola, M. Heinonen, T. Aittokallio, and J. Rousu, "Learning with multiple pairwise kernels for drug bioactivity prediction," *Bioinformatics*, vol. 34, no. 13, pp. i509–i518, 2018.
- 7) H. Oztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.
- 8) T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: prediction of drug-target binding affinity using graph convolutional networks," *Bioinformatics*, Oct. 2020.
- 9) H. Oztürk, E. Ozkirimli, and A. Özgür, "WideDTA: pre-diction of drug-target binding affinity," *arXiv preprint arXiv:1902.04166*, 2019.
- 10) Z.-C. Li, M.-H. Huang, W.-Q. Zhong, Z.-Q. Liu, Y. Xie, Z. Dai, and X.-Y. Zou, "Identification of drug-target interaction from interactome network with 'guilt-by-association' principle and topology features," *Bioinformatics*, vol. 32, no. 7, pp. 1057–1064, 2016.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details