2. REVIEW OF LITERATURE

In recent years, the neuro disorders started to increase rapidly. The researchers significantly identified the proteins that cause the disorder due to mutations and strived hard to found the drugs for neuro disorders that prolong the symptoms. It is revealed that repeat mutations play a major role in neuro disorder called spinocerebellar ataxia and the several researches have been carried out in binding affinity prediction that aids drug design. Some of the important related research works in binding affinity prediction have been described.

This chapter presents the literature review of binding affinity prediction methods used in both general approaches and traditional methods. Section 2.1 begins with the traditional approaches that are used in binding affinity prediction and section 2.2 describes the computational methods for predicting binding affinity using machine learning and deep learning.

2.1 TRADITIONAL APPROACHES

In traditional approaches the features were taken from readily available dataset called PDBBind, PDBculled, PDBBind-CN etc. In these approaches, the prediction of binding affinity was calculated based on the features like scoring functions, molecule descriptors etc. Some of the research works using traditional approaches are reviewed below:

Hongjian Li et al., in [57] proposed a method for prediction of binding affinity in docked complex. They collected data from PDBBind. To predict affinity, features like RF score, physiochemical properties and gauss have been extracted. RF-Score features are elemental occurrence counts of a set of protein-ligand atom pairs in a complex. To calculate these features, atom types were selected so as to generate features that are as dense as possible, while considering all the heavy atoms commonly observed in PDB complexes (C, N, O, F, P, S, Cl, Br, I). As the number of protein-ligand contacts is constant for a particular complex, the more atom types were considered, the sparser the resulting features will be. A minimal set of atom types was selected by considering atomic number only. A smaller set of interaction features has the additional advantage of leading to computationally faster scoring functions. Random forest predicted binding of protein-ligand complexes with the prediction rate of 0.80 and proved that it was superior to other models in prediction of binding affinity.

Xueling Li et al., in [58] proposed a method for automatic protein-protein affinity binding based on svr-ensemble. Two-layer Support Vector Regression (TLSVR) model was employed to implicitly capture binding contributions and the input features for TLSVR in the first layer were scores of 2209 interacting atom pairs within each distance bin. The dataset for this predicting model contained 1056 heterogeneous protein complexes that were obtained from PDBbind-CN, 2010 version. Those complexes include single residue mutation or multiple residue mutations. The dataset was then filtered with sequence similarity <50% by PDBculled with complex entity criteria and other default parameters. Thus 180 protein-protein interaction complexes were formed as a dataset. The base SVRs were combined by the second to infer the final affinities. Protein-protein binding affinity was predicted by using two-layer SVR (TLSVR). Each input vector at first layer of TLSVR is 2209-dimensional. Each real value of a vector represents a score of an atom pair in interface within each of 71 bins, i.e. 1.8 Å: 0.2 Å: 16 Å of a protein complex. Here the contact atom pairs with distance below 1.8 Å of atom clashes were disregarded. The 71 individual SVR models were included at first layer. The predicted values from the individual SVR modes of the first layer were input into the second layer SVR (the combiner). The output of the combiner was the final predicted affinity. Parameters were default in individual SVR models. All the computational experiments were carried out with LIBSVM. TLSVR method obtained better result of 0.90 as correlation coefficient.

Yu Su et al., in [59] planned a way for qualitative prediction of protein-protein affinity by volume correction. The dataset used here was X-ray structures of macromolecule-protein complexes from PDB. To evaluate the prediction ability for protein–protein complexes, six test sets were examined. Sets 1–5 were used as test and the set 6 was the union of sets 1-5 with a whole of eighty six protein–protein complexes and traditional approach of Potential Mean Force (PMF) were applied. Some approaches to calculate PMF were based on the radial distribution function (RDF) in the statistical mechanics of simple liquids. In those approaches, the frequency was normalized in the manner of dividing occurrence numbers in a sphere volume without any correction. Therefore, when normalizing the occurrence frequency of atom pairs, the whole sphere volume was not a good indicator of the actual occupied volume. The authors normalized the occurrence numbers with the numbers in a whole sphere volume ($4\pi r^2 dr$) and also analyzed in detail, the distribution tendency of the occurrence numbers of residue pairs in protein systems with increasing distance and compared it with the occurrence numbers in a whole sphere. This abnormality was due to the occupied volume of atoms in protein complexes deviating significantly from r^2 proportionality. The results obtained were more than 0.73 for all the five test sets.

From these general approaches, it is observed that the prediction of binding affinity was time consuming and it gives low accuracy in terms of prediction rate. Predicting binding affinity was carried out through scoring functions in general approaches. At the same time

scoring functions were not able to forecast binding affinity or binding free energy for a couple of reasons.
The reasons mainly determined enthalpic terms, and disregard entropy, especially of the protein. Entropy is needed, certainly, to calculate binding free energy.
Scoring functions merely is familiar with the bound state of the protein-ligand system, not the unbound states of the protein and the ligand. Binding free energy can only be estimated using knowledge of the bound state and the unbound states of the binding partners. The issues in traditional approaches can be solved along with scoring functions, molecule descriptors to give better results. The machine learning and deep learning algorithms are used to train these features and better prediction rate is achieved. The computational methods based on machine learning and deep learning are reviewed below.

2.2 COMPUTATIONAL METHODS

In these computational methods various machine learning and deep learning techniques are considered for prediction of binding affinity. The features considered here are scoring functions, molecule descriptors and energy calculations that helps in accurate prediction. Some of the research works in computational methods are briefed below:

Jacob D. Durrant and J. St. Andrew McCammon in [60] projected a neural network score to characterize the binding affinity of protein-ligand. Dataset was prepared using the protein-ligand complexes obtained from PDB. The X-ray crystal and NMR structures were used to prepare a dataset. The structures from PDB contain K_d values from MOAD and PDBbind databases. These multiple K_d values were averaged to give one value for protein-ligand complex. Along with this value, features like energy calculations and scoring functions were used. Neural Network (NN) score along with traditional scoring functions successfully characterized the binding affinity of protein-ligand complexes. NN score were not only able to distinguish well docked complexes but also the true ligands docked with decoy compounds.

Volkan Uslan et al., in [61] anticipated the way for significant prediction of HLA-B*2705 compound. The authors projected the prediction of domain-peptide binding affinity models based on support vector regression. The models were applied to yeast bmh 14-3-3 and syh GYF pattern recognition domains. The features from the amino acid datasets CISAPS that has the physio-chemical and bio chemical properties were considered. Support vector regression was able to predict domain-peptide recognitions better than the partial least square algorithm.

Tammy Man-Kuang Cheng et al., in [62] planned a scheme of protein-protein interaction. The accurate scoring of rigid-body docking orientations represents one of the major difficulties in protein–protein docking prediction. They explored a technique called pyDock for rigid docking. It was based on Coulombic electrostatics with distance dependent dielectric constant, and implicit desolvation energy with atomic solvation parameters previously adjusted for rigid-body protein–protein docking. The method was able to detect a near-native solution from 12,000 docking poses and place it within the 100 lowest-energy docking solutions in 56% of the cases, in a completely unrestricted manner and without any other additional information.

Solène Grosdidier and Juan Fernández-Recio in [63] proposed a method for identifying protein hot spots. The structural prediction of protein-protein binding mode, and the identification of the relevant residues for the interaction was considered. Unfortunately, large-scale experimental measurement of residue contribution to the binding energy, based on alanine-scanning experiments, was costly and thus data was fairly limited. Recent computational approaches for hot-spot prediction have been reported. They had applied computational docking approach called Normalized Interface Propensity (NIP) values derived from rigid-body docking with electrostatics and desolvation scoring for the prediction of interaction hot-spots. This parameter achieved upto 80% positive predictive value other than existing methods. The NIP values derived from rigid-body docking can reliably identify a number of hot-spot residues whose contribution to the interaction arises from electrostatics and desolvation effects. This method can propose residues to guide experiments in complexes of biological or therapeutic interest, even in cases with no available 3D structure of the complex.

Pedro J. Ballester, John B. O. Mitchell in [64] proposed a technique for predicting binding affinity of protein-ligand complexes using computational approach. The scoring functions that attempt such computational prediction are essential for analysing the outputs of molecular docking, which in turn is an important technique for drug discovery, chemical biology and structural biology. Each scoring function assumes a predetermined theory-inspired functional form for the relationship between the variables that characterize the complex, which also include parameters fitted to experimental or simulation data and its predicted binding affinity. The inherent problem of this rigid approach was that it leads to poor predictivity for those complexes that do not conform to the modelling assumptions. Moreover, resampling strategies, such as cross-validation or bootstrapping, are still not systematically used to guard against the overfitting of calibration data in parameter estimation for scoring functions. They had proposed a novel scoring function called RF-score. Dataset which was used by them was PDBbind benchmark. Intermolecular interaction features were extracted and RF score was used. RF score was superior to other scoring functions. RF score with random forest achieves the correlation of 0.8. Other scoring functions such as drug score, x-score, HM-score etc., were considered and compared with RF score. Random forest was able to achieve the highest score among all the scores.

Thomas Unterthiner et al., in [65] proposed a model for drug target prediction. In this work they used chEMBL database. In chEMBL database, 13 M compound descriptors, 1.3 M compounds, and 5 k drug targets, compared to the Kaggle dataset with 11 k descriptors, 164 k compounds, and 15 drug targets. Performance of deep learning was compared with seven target prediction methods, including two commercial predictors, three predictors deployed by pharma, and machine learning methods were used to scale this dataset. Deep learning outperformed all the methods with respect to the area under ROC curve and was significantly better than all commercial products. Deep learning surpassed the threshold to make virtual compound screening possible and has the potential to become a standard tool in industrial drug design.

Rhys Heffernan et al., in [66] scrutinized how deep neural network architectures can be used to predict the secondary structure of a protein from genetic sequence. Authors used deep neural network in three iterations and achieved 82% accuracy in predicting the secondary structure. In this work, they had also predicted local backbone angles and solvent accessible surface area of protein. Iterative features had been used for solvent accessible surface area and backbone angles and dihedrals based on Cα atoms. First iteration was used for only seven representative physical chemical properties of amino acid residues and position specific scoring matrix (PSSM) from PSIBLAST. It was employed to predict SS, angles, and ASA, separately. In the second iteration, PSSM/PP plus predicted SS, angles, and ASA were employed from the first iteration. Additional iterations can be followed by using SS, angles, and ASA from the previous iteration in addition to PSSM and PP. Each iteration had three separate predictors and each predictor utilizes one stacked auto-encoder deep neural network.

Babak Alipanahi et al., in [67] authors presented the prediction of sequence specificities of DNA and RNA binding proteins by deep learning. DeepBind approach was used and built standalone software tool that was fully automatic which handles millions of sequences per experiment. DeepBind can be used to predict deleterious SNVs in promoters, by training a deep neural network to discriminate between high-frequency derived alleles (neutral or negative) and simulated variants (putatively deleterious, or positive) from the

CADD framework44. The scores of ~600 DeepBind transcription factor models for the wild type and mutant sequences were used as inputs (~1,200 inputs; Supplementary Fig. 9). The rationale is that a true transcription factor binding site is likely to be located with other transcription factor binding sites, and so these additional scores collectively provide context. When evaluated using held-out test data, the neural network, called DeepFind, achieved an AUC of 0.71, which increased to 0.73 when we included as inputs the distance to the closest transcription start site and a transversion/transition flag. In this approach a set of sequences was used and for each sequence, an experimentally determined binding score was considered. Sequences had varying lengths of 14–101 and binding scores can be real-valued measurements or binary class labels. Deep learning outperforms other state-of-art methods, even when training on in vitro data and testing on in vivo data.

Youjun Xu et al., in [68] proposed prediction models named Drug Induced Liver Injury (DILI) and a data set from the U.S. Food and Drug Administration (FDA's) National Center for Toxicological Research (DL-NCTR) were developed in deep learning. Models were trained on 475 drugs and predicted with an accuracy of 86.9%, sensitivity of 82.5%, specificity of 92.9%, and area under the curve of 0.955. Authors used NCTR, Liew, Xu and combined datasets for both DILI positives and negatives. The performance of DL-NCTR model with 190 drugs performed well with accuracy of 80.5% and with sensitivity 70.3%, specificity 88.2%. DL-view model produced accuracy of 70% and 70% in sensitivity and 70% in specificity. DL- combined model performs best with accuracy of 88.9% and 89.9%, 87% of sensitivity and specificity accordingly. Undirected Graph Recursive Neural Networks (UGRNN) encoding approach with large datasets was developed for the prediction of DILI drugs and small compounds. The DL-combined model performs best with highest accuracy.

Bharath Ramsundar et al., in [69] proposed the use of massively multitask networks for drug discovery. Dataset was created by gathering large amount of data from public sources, more than 200 biological targets. Models with random forest, linear regression, single task model were trained on 259 datasets gathered from public sources. These datasets were divided into four groups PubChem Bio Assay (PCBA), Maximum Unbiased Validation (MUV), Directory of Useful Decoys, Enhanced (DUD-E), and Tox21. Three datasets with random forest achieved the highest enrichment scores with 40. PCBA group contained 128 experiments in the PubChem, BioAssay database whereas MUV group contained 17 challenging datasets. The DUD-E group contained 102 datasets and Tox21 datasets were used. Models were built using multitasking network like logistic regression, random forest, single task neural network, max, pyramidal, one-hidden layer multitask neural network and pyramidal multitask neural network.

Haiping Zhang et al., in [70] proposed deep learning based drug designing for novel corona virus. In this work, they considered 2019-nCov_3C-like protease as a potential target and built a structural model after systematically analyzing its sequence features. The authors built a pipeline with a deep learning based method developed in our group by representing molecules as vectors to identify potential drugs (peptides or small ligands) against the protein target of the 2019-nCoV virus. This method was extremely fast in virtual drug screening and it takes less than a day to finish the virtual screening over millions of protein–ligand or protein-peptide predictions, whereas traditional docking methods take several weeks with the help of a supercomputer. The dataset was prepared using virus RNA sequences from Global Initiative on Sharing All Influenza Data (GISAID) database. The amino acid sequence was translated from the RNA sequence by Translate web tool. Multiple sequence alignment is performed by using Clustal Omega program. They used Dense Fully Convolutional Neural Network (DFCNN) deep learning model to reverse search drug targets. Since the method was shown to have relatively higher accuracy and efficiency, it was suitable for applying to such an emerging disease outbreak. The DFCNN was a densely fully connected neural network, and the densely network allows deep layer without the gradient vanishing problem. The deeper layers make it to learn more abstract features from the data and concluded that this model works faster.

Indra kundu et al., in [71] presented machine learning methods towards prediction of binding affinity using fundamental molecular properties. The prerequisites of this prediction are sufficient and unbiased features of training data and a prediction model which can fit the data well. In this study, they have applied Random forest and Gaussian process regression algorithms from the Weka package on protein–ligand binding affinity, which encompasses protein and ligand binding information from PdbBind database. The models were trained on the basis of selective fundamental information of both proteins and ligand, which can be effortlessly fetched from online databases or can be calculated with the availability of structure. The assessment of the models was made on the basis of correlation coefficient (R^2) and root mean square error (RMSE). The Random forest model produced R^2 and RMSE of 0.76 and 1.31 respectively. The authors concluded that the features used for prediction outperformed the existing ones.

Derek Jones et al., in [72] presented fusion models that combine features and inference from complementary representations to improve binding affinity prediction. It was the first comprehensive study that uses a common series of evaluations to directly compare the performance of three-dimensional (3D)-convolutional neural networks (3D-CNNs), spatial graph neural networks (SG-CNNs), and their fusion. The authors used temporal and structure-based splits to assess performance on novel protein targets. To test the practical applicability of those models, the authors examined their performance in cases that assume that the crystal structure was not available. In those cases, binding free energies were predicted using docking pose coordinates as the inputs to each model. Comparison was done with those deep learning approaches based on docking scores and molecular mechanic/generalized born surface area (MM/GBSA) calculations. The results showed that the fusion models made more accurate predictions than their constituent neural network models as well as docking scoring and MM/GBSA rescoring, with the benefit of greater computational efficiency than the MM/GBSA method.

Nguyen et al., (2019) in [73] presented prediction of drug-target based binding affinity prediction models using graph neural networks. Multiheaded input CNNs have been used in these regression problems in which the drug (small molecule) and protein were input separately, usually as SMILES strings and character based amino acid sequences respectively, passed through convolutional blocks, merged, then passed through dense layers. Here, the authors, replace SMILES strings with various graph convolutional layers. In this method, molecules were represented as mathematical graphs. The node feature vector were constituted of five types of atom features such as atom symbol, atom degree – number of bonded neighbors plus number of hydrogen, total number of hydrogen, implicit value of atom, and aromatic or not. Those atom properties constitute a multi-dimensional binary feature vector. An edge was set to a pair of atoms if there exists a bond between them. As a result, an indirect, binary graph with attributed nodes was built for each input SMILES string. GraphDTA can not only predict the affinity of drugs-targets better than non-deep learning models, but also outperform competing deep learning methods. GraphDTA performed consistently well across two separate benchmark databases in all the evaluation measures. The result suggested that representing molecules as graphs can improve the performance considerably. Also, it was confirmed that deep learning models were appropriate for drug-target binding affinity problems.

Majumdar et al., in [74] presented Deep learning-based potential ligand prediction framework for COVID-19 with drug–target interaction model. The authors implemented an architecture using 1D convolutional networks to predict drug–target interaction (DTI) values. The network was trained on the KIBA (Kinase Inhibitor Bioactivity) dataset. With this network, the authors predicted the KIBA scores (which gives a measure of binding affinity) of a list of ligands against the S-glycoprotein of 2019-nCoV. Based on those KIBA scores, 33 ligands were proposed that had a high binding affinity with the S-glycoprotein of 2019-nCoV and thus can be used for the formation of drugs. This research is of utmost importance where the proposed new compound, if validated by biochemists as an effective solution, can help mankind survive this tough time. In this work, the authors have trained a machine learning model for the prediction of KIBA scores for a pair of protein–ligand. Using that model, the top 33 ligands were identified that can be used to find a potential cure for SARS-CoV-2.

Shim J et al., in [75] proposed that a similarity based model that applies two dimensional [2D] convolutional neural network [CNN] to the outer products between column vectors of two similarity matrices for the drugs and targets to predict DT binding affinities. This was the first application of 2D CNN in similarity DT binding affinity prediction. The validation results on multiple public datas showed that the proposed model was an effective approach for DT binding affinity prediction and can be quite helpful in drug development process. Continouus value provide more information about the actual strength of DT binding. Experimental results showed that SimCNN-DTA outperformed other existing methods such as KronRLS and DeepDTA in prediction performance on the Davis and KIBA datasets. The case study found that drug candidates targeting EGFR showed that SimCNN-DTA included all existing EGFR drugs as 100 top ranked candidates among 1018 candidates. SimCNN-DTA can be futher improved by adjusting the architecture of CNN according to the data structure. The SimCNN-DTA was an effective approach for DTA prediction and can be quite helpful in drug development process.The summary of literature review is given in Table 2.1.

Table 2.1 Summary of Literature Review

Authors	Objective	Dataset	Algorithm	Prediction Rate
Tammy Man-Kuang Cheng, Tom L. Blundell, Juan Fernandez-Recio (2007)	Protein-protein interaction using rigid body docking	ICM and ICM-DISCO benchmark datasets	PyDock	It predicted the 56% cases from 80 unbound docking poses
Solène Grosdidier and Juan Fernández-Recio (2008)	Identification of protein hotspots using computational approach	Structures from PDB	Normalised Interface Propensity	0.80
Yu Su, Ao Zhou,	Prediction of	X-ray structures	Traditional	It obtained

Xuefeng Xia, Wen Li, Zhirong Sun (2009)	protein-protein binding affinity	from PDB	approach of PMF	more than 0.73 for all the six test sets
Jacob D. Durrant and J. Andrew McCammon (2010)	Characterizing the protein-ligand binding affinity	Protein-Ligand complexes from PDB	Neural Network	Neural network combined with traditional functions successfully characterized the binding affinity of protein-ligand
Pedro J. Ballester, John B. O. Mitchell (2010)	Predicting binding affinity of protein-ligand complexes using scoring functions	PDBbind database	Random Forest	It produced the correlation coefficient of 0.953
Li X., Zhu M., Li X., Wang HQ., Wang S (2012)	Prediction of protein-protein binding affinity using SVR-ensemble	Complexes from PDBbind-CN	Two layered SVR ensemble	SVR ensemble produced the correlation coefficient as 0.9
Thomas Unterthiner, Andreas Mayr, Gunter KlambauerJesse, Marvin Steijaert, Jorg K. Wegner, Hugo Ceulemans, Sepp Hochreiter, (2014)	Drug-target prediction by deep neural networks	chEMBL database	Deep neural network	0.83
Li H., Leung KS., Wong MH., Ballester P.J (2015)	Prediction of binding affinity from docked complexes	PDBbind	Random Forest	0.80
Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, YuedongYang & Yaoqi Zhou (2015)	Secondary structure prediction from genetic sequences	TR4590 dataset	Deep neural network	82% of accuracy in secondary structure prediction
Babak Alipanahi,	Predicting the	Protein binding	Deep neural	DeepBind

Andrew Delong, Matthew T Weirauch & Brendan J Frey (2015)	sequence specificities of RNA and DNA- binding proteins	microarrays (PBMs), RNAcompete assays, chromatin immunoprecipitation (ChIP)-seq	network	models trained in vitro data worked well for in vivo data
Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai (2015)	Prediction of drugs for liver injury	Four public datasets of DILI-positive and DILI-negative properties of drugs	Deep neural network	DILI prediction model predicted the drugs with 80% of accuracy
Bharath Ramsundar, Steven Keames, Patrick Riley, Dale Webster, David Konerding, Vijay Pande (2015)	Multitask networks for drug discovery	259 publicly available datasets	Deep neural network	Multitask networks performed well than single task models
Huseyin Seker, Volkan Uslan (2016)	Binding affinity prediction of domain-peptide recognition	Amino acid datasets	Support Vector regression	SVR predicts better than the partial least square
Zhang, H., Saravanan, K.M., Yang, Y. (2020)	Binding affinity prediction for drug designing	Virus RNA dataset from GISAID	DFCNN	DFCNN predicts the binding affinity in a precise manner
Indra kundu, goutam paul and raja banerjee (2018)	Binding affinity prediction of protein-ligand complexes	PDBBind	Random forest and Gaussian regression	Random forest gives best score of 0.76 as R^2
Derek Jones, Hyojin Kim, Xiaohua Zhang, Adam Zemla, Garrett Stevenson, W. F. Drew Bennett, Daniel Kirshner, Sergio E. Wong, Felice C. Lightstone, and Jonathan E. Allen (2021)	Binding affinity prediction	PDBBind	3D CNN and SG-CNN	Fusion model gave the best prediction rate

Nguyen et al., (2019)	Drug-target based binding affinity prediction	SMILES	Graph neural networks	Graph DTA model predicts the model accurately
Majumdar, S., Nandi, S.K., Ghosal, et al., (2021)	Deep learning based potential drug prediction	KIBA dataset	1D CNN	33 ligands are found for curing covid
Shim, J., Hong, ZY., Sohn, I. et al (2021)	Prediction of Drug-target binding affinity using similarity-based CNN	DAVIS and KIBA datasets	2D CNN	SIMCNN outperforms all the other models in prediction