3. PROBLEM MODELING

The main focus of this research is to propose an evident model to predict binding affinity for spinocerebellar ataxia in human which occurs due to repeat mutations. The research problem of predicting binding affinity is formulated as regression task and suitable model is proposed to build using traditional machine learning and deep learning architectures. This chapter portrays the approach of problem modelling that facilitates the objectives.

The research work is carried out in two stages with traditional machine learning and deep learning approaches to build predictive models. In the first stage, the traditional machine learning approach is utilized to build the prediction models and the idea here is to identify and extricate the distinct features from the simulated complexes. Features like energy calculations, sequence descriptors, physio-chemical properties are derived from the docked complexes and mutation induced docked complexes. Three independent datasets with these features are equipped and the models are developed using traditional machine learning techniques to predict the binding affinity for spinocerebellar ataxia.

In the second stage, the contemporary deep learning approach is adopted for building the predictive models wherein the deep learning architectures self extracts the hidden features by high level representation learning. The idea here is to learn the significance of the features and signalling between them that leads in accurate prediction of binding affinity. The datasets used in traditional machine learning approach are used again in deep learning approach to facilitate representation learning from the user defined features.

3.1 OVERALL FRAMEWORK OF BINDING AFFINITY PREDICTION MODEL

The overall framework of binding affinity prediction model is divided into four phases (i) corpus creation (ii) feature extraction and dataset development (iii) building predictive models and (iv) performance evaluation of the models. Binding affinity prediction models are constructed with three corpuses namely protein-ligand corpus, protein-mutated-ligand corpus and protein-protein corpus. To prepare the corpuses, the protein structures are gathered from PDB and ligands from various literatures. Mutational information for repeat mutation are collected from HGMD database and induced to the protein structures. These protein structures and ligands are docked using three different docking approaches to produce docked complexes. From these complexes, features such as energy calculations, scoring functions, sequence descriptors and physio-chemical properties are identified and extracted. Three respective datasets are developed and named as PLD dataset, PMLD dataset and PPD dataset. These datasets are trained by traditional machine learning and contemporary deep learning techniques. Binding affinity predictive models are built by implementing regression algorithms such as linear regression, support vector regression, random forest, artificial neural network and deep learning algorithms such as sequential deep neural network, functional deep neural network, customized layers with DNN. These machine learning and deep learning approaches have been chosen to improve the prediction rate of the models. The models are evaluated with various metrics such as explained variance score, R2 score, mean squared error, root mean squared error, mean absolute error and median absolute error and the prediction rate is determined. The proposed framework of binding affinity prediction model is depicted in Fig. 3.1. Corpus development is explained in section 3.2, feature extraction and preparation of datasets is described in section 3.3. Various evaluation metrics is elucidated in section 3.4.

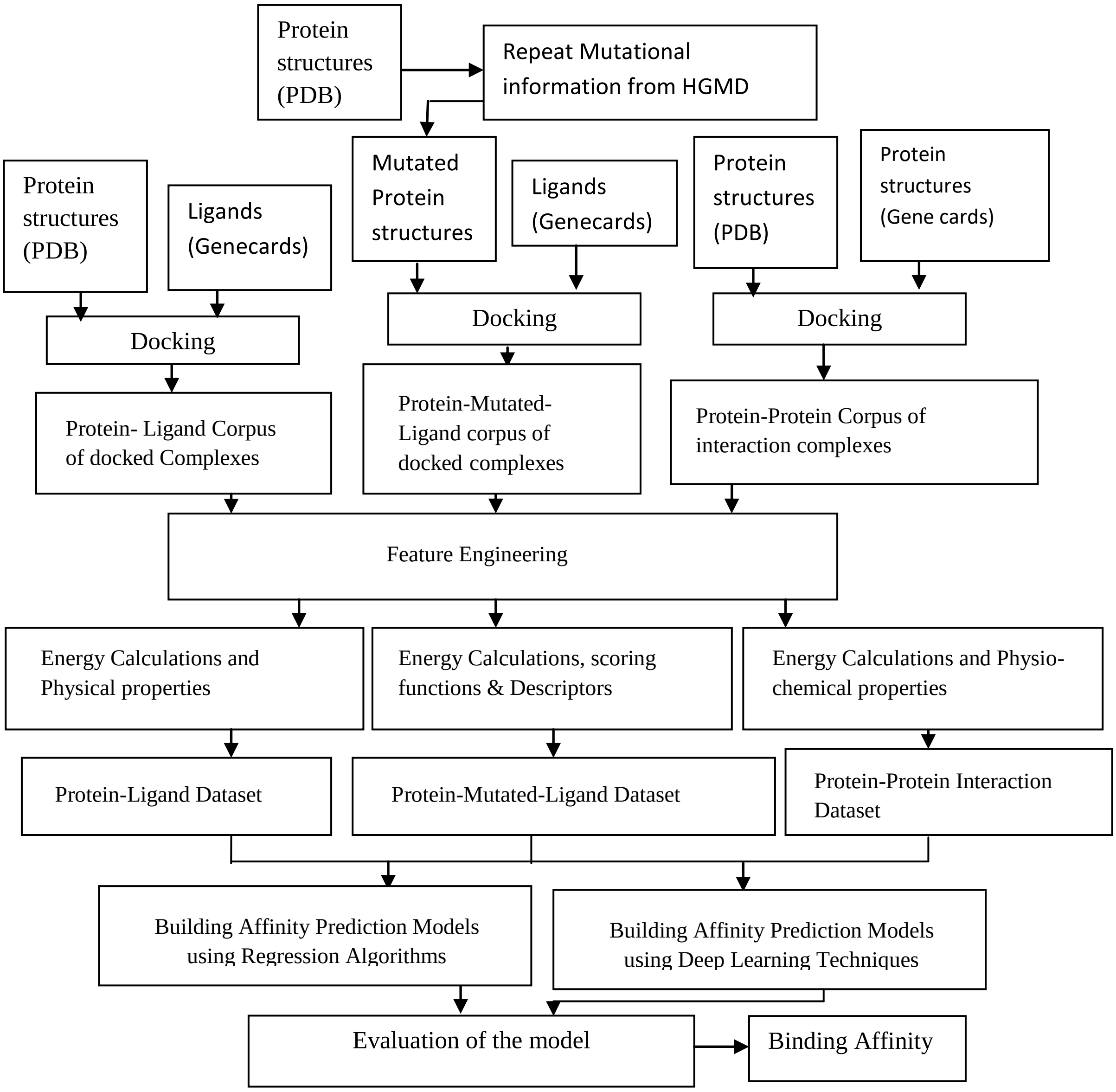

Protein structures (PDB)
Repeat Mutational information from HGMD
Protein structures (PDB)
Ligands (Genecards)
Mutated Protein structures
Ligands (Genecards)
Protein structures (PDB)
Protein structures (Gene cards)
Docking
Docking
Docking
Protein- Ligand Corpus of docked Complexes
Protein-Mutated-Ligand corpus of docked complexes
Protein-Protein Corpus of interaction complexes
Feature Engineering
Energy Calculations and Physical properties
Energy Calculations, scoring functions & Descriptors
Energy Calculations and Physio-chemical properties
Protein-Ligand Dataset
Protein-Mutated-Ligand Dataset
Protein-Protein Interaction Dataset
Building Affinity Prediction Models using Regression Algorithms
Building Affinity Prediction Models using Deep Learning Techniques
Evaluation of the model
Binding Affinity

Fig. 3.1 Proposed Framework of Binding Affinity Prediction Model

3.2 DEVELOPMENT OF CORPUSES

Data collection is a procedure of gathering information from all the appropriate sources to discover resolution to the research problem, testing hypothesis and assessing the outcome. The main focus of data collection is to capture the superior data to bring the desired outcome for the problems. Accurate prediction of binding affinity is a challenging task as the structure changes due to mutation and also binding affinity changes for each and every structure. The protein structures for six types of SCA that commonly occur due to repeat mutation presented in Section 1.2 and the structures of ligands shown in Fig. 1.10 are considered for docking. The information of repeat mutation given in Table 1.4 is used for inducing mutation.

In this research work, the homosapiens protein structures are considered. The types of spinocerebellar ataxia are SCA1, SCA2, SCA3, SCA6, SCA8, SCA10 and the respective 3d protein structures are ataxin-1, ataxin-2, ataxin-3, voltage-dependent P/Q-type calcium channel subunit alpha-1A, ataxin-8, ataxin-10. These six types of protein structures contain many protein structures that are related to these protein structures as listed in Table 1.3. The total of 17 protein structures related to six types of SCA, 609 interacting protein structures and 18 ligands are considered here for developing corpuses. Three corpuses are developed and the respective datasets are constructed for building the predictive models.

Commonly available ligands in various literatures are taken into account for molecular docking. Also the ligands which were used in docking studies of the species mouse affected through various types of spinocerebellar ataxia are also included since the DNA pattern of human is similar as that of mouse. For instance, the ligand amantadine was previously docked with mice protein structures of type SCA1 [76] but in this work the ligand is docked with other 5 types of SCA along with SCA1 for human structures. Similarly the ligands used previously for neurodegenerative disorders and experimented with animal models are considered for this research [77].

Polyglutamine repeats are expanded repeats of CAG nucleotides which encodes the amino acid glutamine. Each structure is mutated with the mutational information, collected from HGMD and the mutated protein structures are prepared. Mutated protein structures are docked with ligands to capture the changes that occur in the protein structure. Binding affinity from protein-protein interaction aids in interpreting unknown biological function since both proteins are macromolecules, some functions of proteins are unknown. Interacting proteins for protein-protein interaction are collected from gene cards. Rigid docking is performed for protein-protein interaction and flexible docking is performed for protein-ligand docking.

Protein-Ligand (PL) corpus

The optimal binding between a small compound and protein is required to find appropriate binding affinity. Hence it is proposed to develop a corpus of docked complexes by docking seventeen protein structures and eighteen ligands. Each protein is docked with all the eighteen ligands to produce collection of docked complexes. Docking is executed through autodock which is based on genetic algorithm (GA) [78].

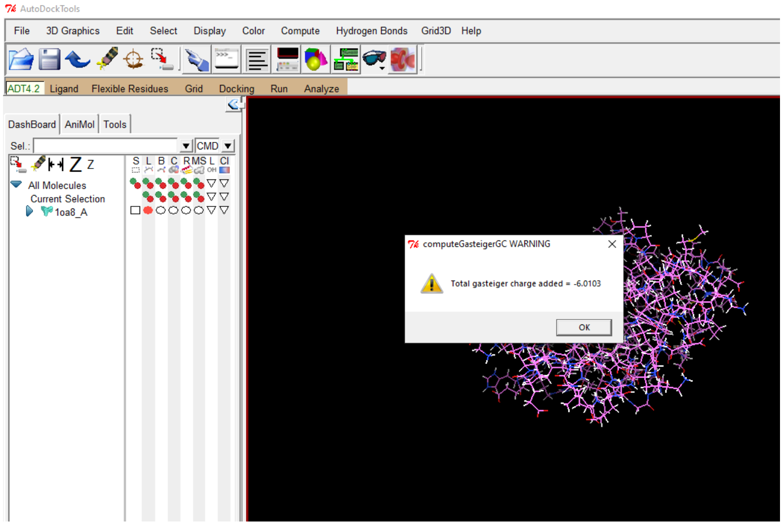
In autodock, the PDB file of protein structures are given as input and it is prepared by adding kollman and computing gasteiger charges. Protein structures are covered by grid to generate Grid Parameter File (GPF) and Docking Parameter File (DPF). The GPF specifies the 3D search space by setting the number of points in each dimension, the center of the grids and the spacing between points. It also specifies the types of probe atoms to use, the filename of the receptor and the names of each output gridmaps. The DPF docking parameter file gives information about, map files to use, the ligand molecule to move, center of ligand and number of torsions. It also provides the start of the ligand, the flexible residues in the side chain of the receptor which is to be modelled, docking algorithm and the number of runs. Ligand is prepared by finding the root to dock with the protein. Genetic algorithm is used to generate Grid Log File (GLG) and Dock Log File (DLG). GLG is output from the file autogrid whereas DLG is the output from autodock. This produces various conformation values and the lowest conformation value is chosen as it has the strong binding values.

For example, consider the protein structure 1oa8 and the ligand amantadine. The macromolecule 1oa8 has four chains A, B, C and D. Since all the chains have same residues, A chain is considered for docking. The protein structure is fed as PDB file to autodock Protein and ligand is prepared by adding charges, creating grid and dock files. The docking involves the following five steps.

Step 1: Adding and Computing Charges - Initially the PDB file corresponding to the protein structure 1oa8 shown in Fig. 3.2 is given as input to autodock. The hydrogen bonds are added and the charges like kolman, gasteiger are computed for 1oa8 protein structure. This file is saved as PDBQ and shown in Fig. 3.3.

```
ATOM      1  N   GLY A 562     -14.689 -38.398  -7.187  1.00 29.90           N
ATOM      2  CA  GLY A 562     -15.570 -37.442  -6.452  1.00 30.20           C
ATOM      3  C   GLY A 562     -14.739 -36.463  -5.628  1.00 30.06           C
ATOM      4  O   GLY A 562     -13.535 -36.344  -5.865  1.00 29.96           O
ATOM      5  N   SER A 563     -15.374 -35.783  -4.669  1.00 28.52           N
ATOM      6  CA  SER A 563     -14.755 -34.666  -3.940  1.00 29.39           C
ATOM      7  C   SER A 563     -14.715 -34.931  -2.448  1.00 26.88           C
ATOM      8  O   SER A 563     -15.760 -35.026  -1.837  1.00 26.07           O
ATOM      9  CB  SER A 563     -15.519 -33.360  -4.199  1.00 29.85           C
ATOM     10  OG  SER A 563     -14.954 -32.693  -5.326  1.00 35.62           O
ATOM     11  N   PRO A 564     -13.522 -35.024  -1.841  1.00 24.79           N
ATOM     12  CA  PRO A 564     -13.505 -35.385  -0.422  1.00 24.26           C
ATOM     13  C   PRO A 564     -13.980 -34.240   0.489  1.00 22.98           C
```

Fig. 3.2 PDB File of 1oa8



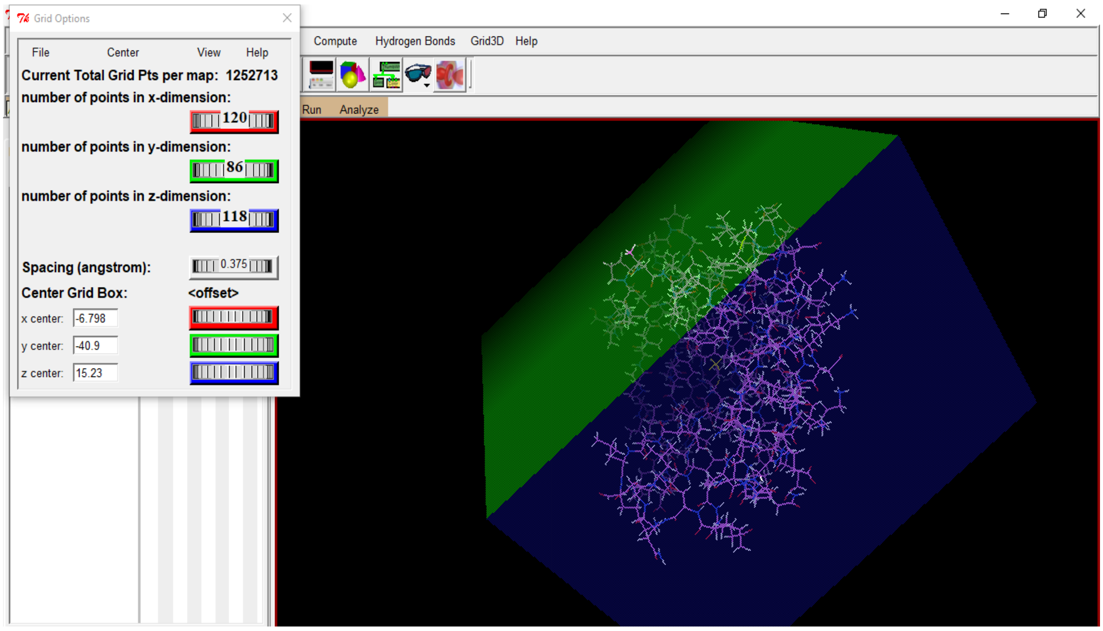
computeGasteigerGC WARNING

Total gasteiger charge added = -6.0103

OK

Fig. 3.3 Computing Charges for PDB File

Step 2: Preparation of GPF – Protein PDBQT file corresponding to 1oa8 is prepared by opening PDBQ file where the charges of protein are not preserved and GPF is prepared. GPF of 1oa8 is created by making grid boxes around the protein where the protein should be fully covered and saved as output.gpf. The making of GPF is shown in Fig.3.4.



Current Total Grid Pts per map: 1252713

number of points in x-dimension: 120

number of points in y-dimension: 86

number of points in z-dimension: 118

Spacing (angstrom): 0.375

Center Grid Box: <offset>

x center: -6.798

y center: -40.9

z center: 15.23

Fig. 3.4 Grid Parameter File for 1oa8

Step 3: Preparation of DPF - Once protein structures are prepared for docking, the ligand amantadine is prepared to dock with protein. Initially the ligands are in .mol format where the .mol format is changed to pdb format. Ligand amantadine is loaded into autodock where its root is detected and saved as PDBQT format. DPF is created by opening the PDBQT files of both protein and ligand. Search algorithm GA is used in docking parameter where the GA runs is given as 50.

Step 4: GLG & DLG Preparation – GPF of 1oa8, amantadine and DPF of 1oa8, amantadine are edited and saved to launch the GLG. GLG and DLG are created by running autogrid and autodock respectively. Docking is performed with GLG and DLG files which generates a list of conformation values. Binding energy conformation is based on the lowest energy. Among the clusters of binding energy, the lowest energy value is taken which implies strong binding. The cluster of conformation is shown in Fig. 3.5 and the sample output of binding energy with other energy values is shown in Fig. 3.6.

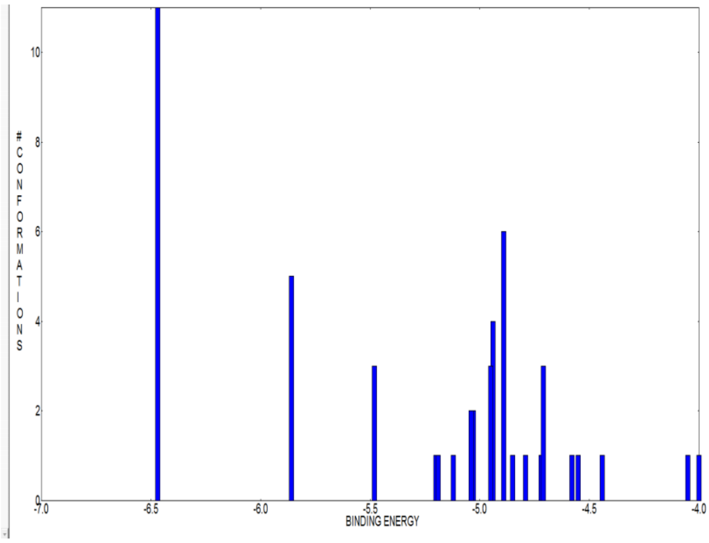
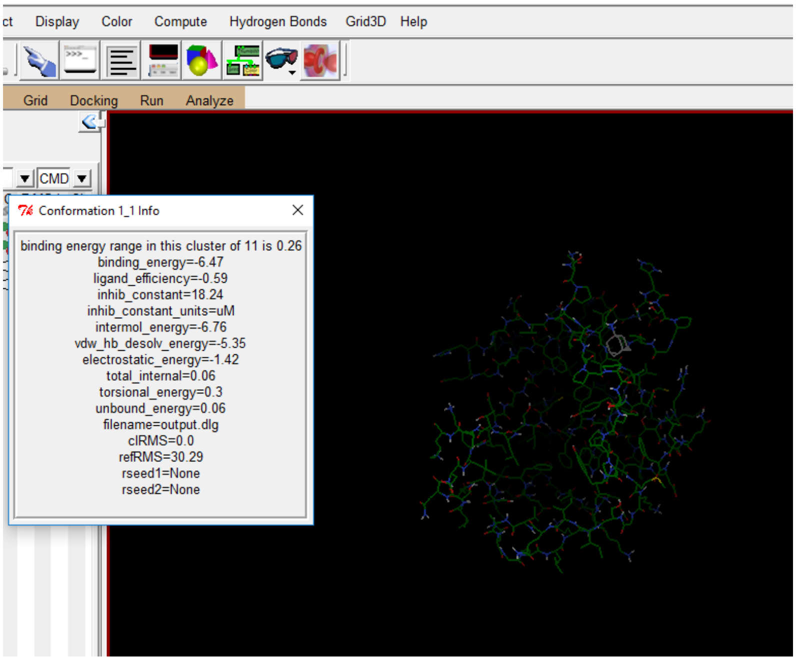


Fig. 3.5 Cluster of Confirmation



```
binding energy range in this cluster of 11 is 0.26
binding_energy=-6.47
ligand_efficiency=-0.59
inhib_constant=18.24
inhib_constant_units=uM
intermol_energy=-6.76
vdw_hb_desolv_energy=-5.35
electrostatic_energy=-1.42
total_internal=0.06
torsional_energy=0.3
unbound_energy=0.06
filename=output.dlg
clRMS=0.0
refRMS=30.29
rseed1=None
rseed2=None
```

Fig. 3.6 Sample Output of Binding Energy

Step 5: Binding Site Identification – The docked complex and the binding site of ligand with protein structure 1oa8 is viewed through pymol. The ligand amantadine is docked to glutamine amino acid in the position 644, 648 and 652. The docked complex and the binding site of ligand amantadine with protein 1oa8 is shown in Fig. 3.7

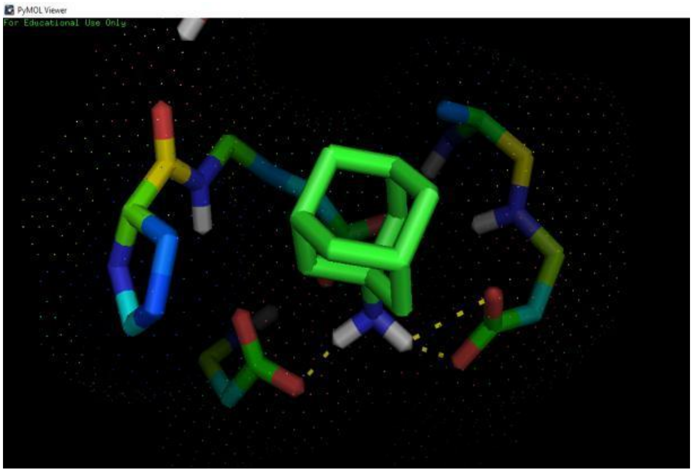


Fig. 3.7 Docked Complex of 1oa8 with Amantadine

In this manner, the protein-ligand docking is performed for seventeen protein structures and eighteen ligands. This docking process developed the collection of 307 docked complexes and the corpus is named as PL corpus.

Protein-Mutated-Ligand (PML) corpus

In the previous case, the changes in protein structure due to mutation are not analyzed for finding the binding affinity. But binding affinity value changes due to variation in the protein structure. Analyzing the changes facilitates in capturing the sequence descriptor changes, binding site changes etc., to determine the accurate binding affinity value. So the change in the protein structure that occurs due to repeat mutation is taken into account in this case to revise the corpus of docked complexes. The same seventeen protein structures and eighteen ligands are considered here also for docking. Docking is executed through autodock and the same procedure is followed. The steps followed to create docked complex for the sample protein structure 1oa8 and ligand amantadine are given below.

Step 1: Inducing Mutation - The seventeen protein structures are mutated with repeat mutation according to the information from the HGMD database given in Table 1.4. The general count of repeat mutation corresponding to protein sequence of 1oa8 associated with ataxin-1 protein is 40-100. Here the protein sequence of 1oa8 shown in Fig. 3.8a is induced with 40 repeats of glutamine amino acid. The mutated protein sequence is depicted in Fig.

3.8b. The structure of protein also changes due to repeat mutation, the normal protein structure overlapped with mutated protein structure is illustrated in Fig. 3.9 to project the structural differences in 1oa8.

```
GSPAAAPPTLPPYFMKGSIIQLANGELKKVEDLKTED
FIQSAEISNDLKIDSSTVERIEDSHSPGVAVIQFAVG
EHRAQVSVEVLVEYPFFVFGQGWSSCCPERTSQLFDL
PCSKLSVGDVCISLTLKNLKNG
```

```
GSPAAAPPTLPPYFMKGSIIQQQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQQQQQQQQQLANGELKKVEDLKTEDFIQS
AEISNDLKIDSSTVERIEDSHSPGVAVIQFA
VGEHRAQVSVEVLVEYPFFVFGQGWSSCCPE
```

Fig. 3.8a Sequence 1oa8: Before Mutation **Fig. 3.8b Sequence 1oa8: After Mutation**

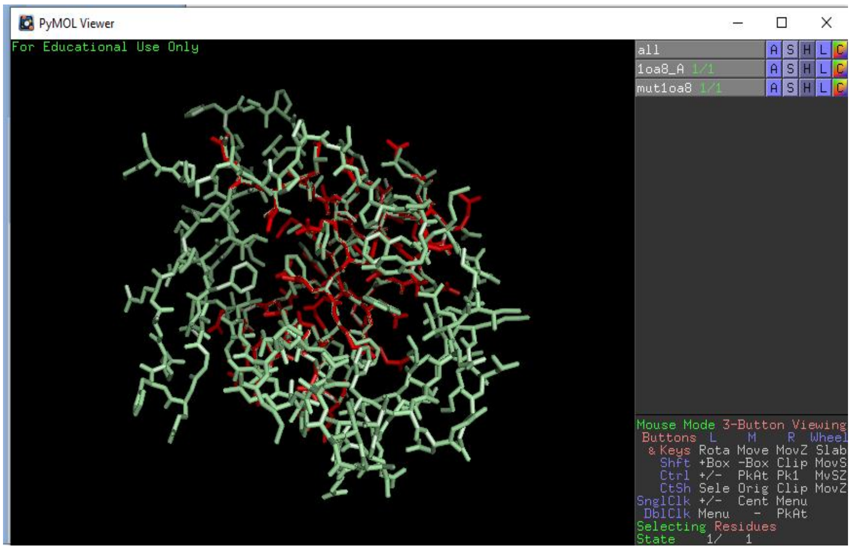
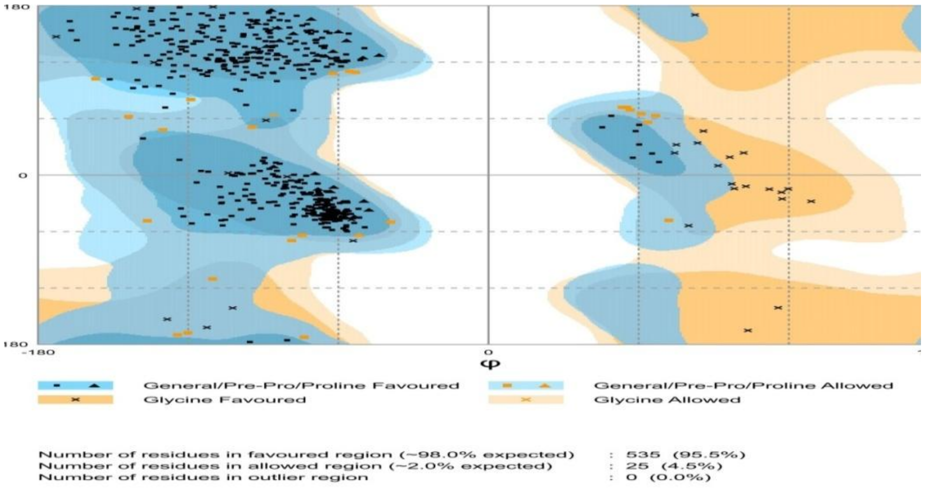


Fig. 3.9 Structures Overlapped with Normal 1oa8 and Mutated 1oa8

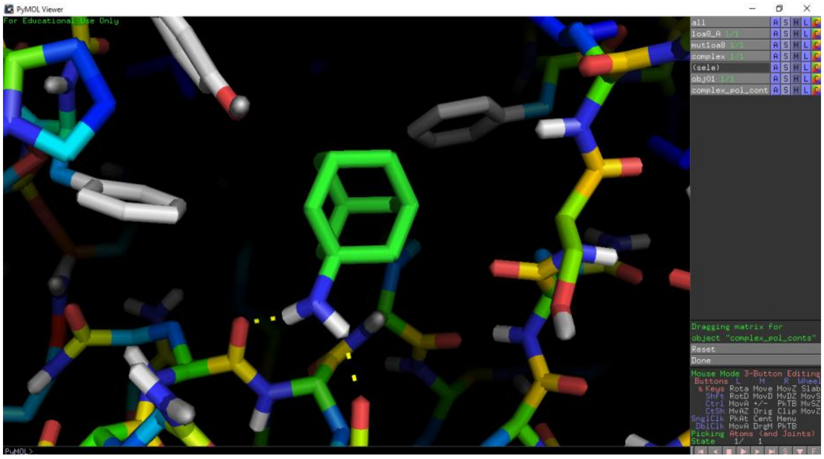
Step 2: Validation - Each mutated protein structures are checked with the ramachandran plot for validation as some protein structures lose their validity due to mutation. The plot illustrates favoured and allowed regions. The residues in favoured region and allowed region are considered to be valid. The amino acids in unfavored region are made to fit in the allowed region and the validity is checked again. The amino acid glycine can fall in the unfavored region. The ramachandran plot of mutated 1oa8 is shown in Fig 3.10. This plot shows 95.5% residues are in favoured region and 4.5% residues are in allowed region. There is no residue in outlier region and the protein is valid for docking.



```
Number of residues in favoured region (~98.0% expected)   :  535  (95.5%)
Number of residues in allowed region (~2.0% expected)     :  25  (4.5%)
Number of residues in outlier region                      :  0  (0.0%)
```

Fig. 3.10 Ramachandran plot of Mutated 1oa8

Step 3: Docking - After the validation check for each mutated protein structure, it is docked with the ligands using autodock. The procedure of docking adopted in previous corpus is used here also to produce the collection of mutated docked complexes. The mutated protein structure docked with ligand amantadine is shown in Fig. 3.11 and it shows that the ligand amantadine is docked with two amino acids in the mutated protein structure 1oa8. The amino acids binded with ligand are Glycine (G) and Leucine (L). The ligand binding with the normal protein structure and mutated protein structure differs and this difference is reflected in physio-chemical properties which are captured during feature extraction process.



Dragging matrix for object "complex_pol_conts"

Fig. 3.11 Docked Complex of Mutated Protein Structure 1oa8 with Ligand Amantadine

By this way, the seventeen mutated protein structures are docked with eighteen ligands and the pool of 307 mutated docked complexes is developed. This corpus is named as PML corpus.

Protein-Protein (PP) corpus

Protein-protein interaction is essential because the hidden biological process of proteins can be known when two macromolecules interact. The corpus is developed using seventeen protein structures of SCA and 609 interacting protein structures. The proteins for interaction are gathered from genecards. Protein-protein interaction is performed through haddock software and prodigy in haddock predicts the binding energy of protein-protein complexes [79]. Binding site of protein structures are identified using CNN before the interaction. Protein-protein interaction is performed with the known binding sites as both the proteins are macromolecules. Since both the proteins are macromolecules, rigid docking is performed in haddock.

In haddock, the PDB file of protein structures with chain is given as input. In this stage, the interacting proteins are treated as rigid bodies, meaning that all geometrical parameters such as bonds lengths, bond angles, and dihedral angles are frozen. The proteins are separated in space and rotated randomly about their centres of mass. This is followed by a rigid body energy minimization step, where the protein structures are allowed to rotate and translate to optimize the interaction. The second stage of the docking protocol is flexibility to the interacting partners through a three-step molecular dynamics-based refinement in order to optimize the protein structures. The flexibility in torsion angle space defines that bond lengths and angles are still frozen. The interacting partners are first kept rigid and only their orientations are optimized. Flexibility is then introduced in the interface, which is automatically defined based on intermolecular contacts within a 5Å cut-off. Rigid body minimization produces different binding poses. Residues belonging to this interface region are then allowed to move their side-chains in a second refinement step.

Finally, both backbone and side-chains of the flexible interface are made flexible. The final stage of the docking protocol immerses the complex in a solvent shell to improve the energies of the interaction. The final models are automatically clustered based on a specific similarity measure where the cluster with lower binding energy is chosen. The steps carried out in protein-protein interaction are given below.

Step 1: Identification of Interacting proteins – The interacting proteins are identified from genecards with respective to the seventeen protein structures. The sample interaction profile for protein ataxin-1 is shown in Fig. 3.12. The interaction profile of proteins for six types of SCA is listed in Table 3.1.

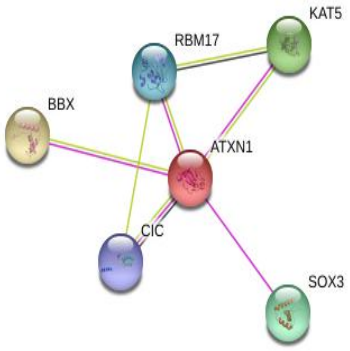


Fig. 3.12 Sample Interaction Pathway for Atxn1

Table 3.1 Interaction Profiles

Protein	Interacted Proteins
Ataxin-1	4j2l,4j2j,2m41,2gzk,1j46,1yqb,2jy6,2knz,4xos,4kdi,2pjh,1s3s,5ftn,5ftj,5c19,3cfo,1u8f,2xxn,2f1x,2f1z,1nbf,5fwi,2kbr,5jtv,4pyz,3u3o,4wpi,4y0c,3bzh,1y61
Ataxin-2	31py,2cqb,2r99,4pjo,3jcr,5mf9,1d3b,1n54,1h2t,1h6k,1h2v,1n52,3p8b,2ckk,3fe2,4pxa,4lk2,2i4i,4kbg,4kbf,3kx2,1n52,1cbj,2k8g,5ifn,1cvj,4f02,2xa6,5elt,5vl3,2bl5,3qhe
Ataxin-3	5ijo,4zol,4tv9,5fnv,5iy4,3vht,4kdi,2pjh,1s3s,5ftn,5ftj,5c19,3cfo,4v3l,3u3o,4ksl,1gjz,5gjq,3b08,3low,2w9n,5b83,2znv,3zn2,2qho,5gjq,5hpl,5koy,4k2x,4uq5,3o65,4xkh,2kl2,2mkg,4wth,4kbq,3q4a,2c2l,2oxq,1p1a,1oel,1ify,2f4m,2qsf,1dvo0,1iyf,5c1z,4inf,2jm0,4p50,2n7k,2brf,3zvn,3zvl,4rck,1jey,1jeq,1jjr,1e17,2k86,
CACNA1A	4l9m,2vrw,4l9u,5cm8,1xd2,5kbt,1nvv,2yuu,4dex,3dvk,ebxl,3bxk,3dvj,2ws7,3w14,3w11,1g7a,4oga,1jk8,2kqp,1toc,4y19,4qsz,2w44,1b9y,1m56,5kd0,4q5q,1aqg,3mpx,1xd4,4f7z,3c5h,3h5h,2ee5,3ah8,2bcj,3pvu,2rmk,5hzh,1x86,5c2k,3cx8,3ab3,1zca,3uzs
Ataxin-10	2bcj,3uzs,1xhm,3ny8,3a8y,1xqs,1yuw,4wv7,4po2,3lof,1hx1,3c7n,1ckr,4kbq,2p32

Step 2: Binding Site/Hot-spot Identification - Binding site is crucial for rigid docking and hotspot identification is indispensable for interaction of proteins. The binding site of a molecule that bound with another protein is identified to analyze resultant chemical reaction and the corresponding biological process. The residues in the active site form temporary bond with enzyme and residues catalyse a reaction of that enzyme. To recognize the active site, 3D protein structure is conceded into convolutional neural network. The protein structure is applied with some learnable weights, bias and non-linear function. The binding score of protein structure is derived through four layers namely convolutional layers, pooling layers,

activation layers and other layers. The protein structure is passed to convolutional layers with max pooling and dropout after every two convolutional layers, followed by one regular fully connected layer. Exponential linear unit activation function is used at the last layer.

The output of the network is applied with sigmoid function and the threshold values are ranging from 0 to 1. The process of CNN for active site identification is shown in Fig. 3.13. The threshold value 0.5 is considered for binding site identification and the amino acid with the threshold value 0.5 is taken as binding site. The binding site prediction is evaluated with two criteria such as distance to the center of the binding site and discretized volumetric overlap. The process is freely available in the website as www.playmolecule.org [80]. The threshold values obtained for sample protein structures are revealed in Table 3.2. The binding site identification of protein structure 1j46 obtained using convolutional neural network is shown in Fig. 3.14.

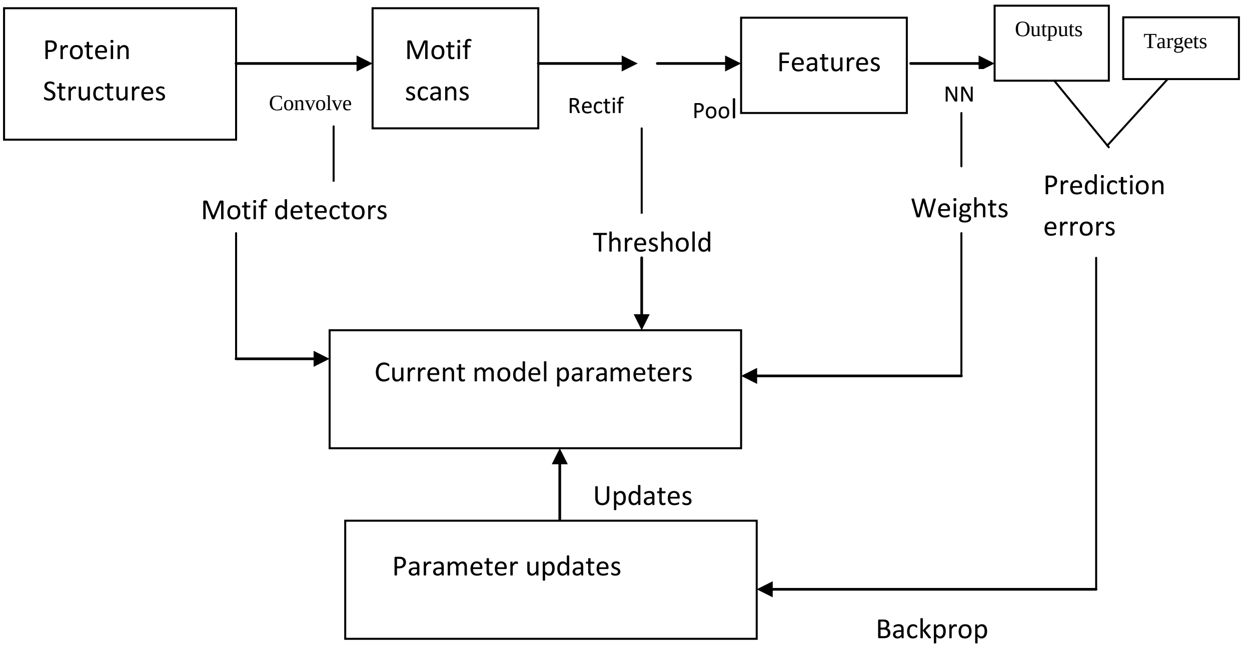


Fig. 3.13 Process of CNN for Active Site Identification

Table 3.2 Thresholds of Sample Protein Structures

Protein	**<.5**	**>.5**
1oa8	0.4	0.9
1j46	-	1.0

1yzb	0.5	0.9
4v3l	-	0.8

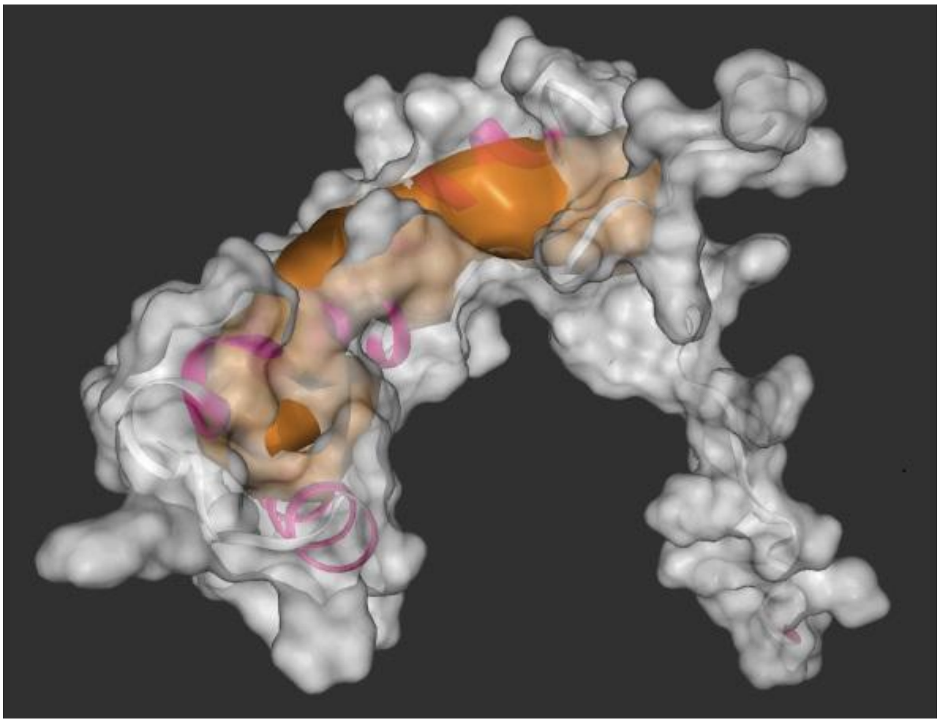


Fig. 3.14 Active Site of 1j46

Step 3: Protein-protein Interaction - Binding site identified through convolutional neural network facilitates in interaction with protein, ligand, ion etc. In this work, protein-protein interaction is performed using haddock software to construct the interacted complexes. Each protein structure among seventeen protein structures is interacted with each protein of 609 interacting proteins. Totally 626 protein structures are interacted and 313 interacted complexes are produced.

The complexes obtained from haddock are clustered and the cluster with the minimum energy and minimum score is chosen for further feature extraction process. The interaction of protein 1oa8 with 2jy6 is shown in Fig. 3.15 and the cluster of 1oa8 is shown in Fig. 3.16.

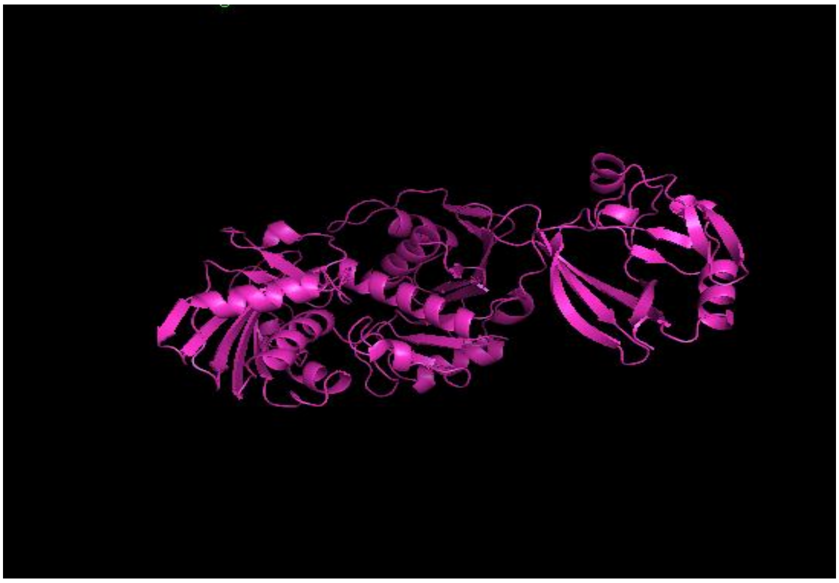


Fig. 3.15 Protein-protein Interaction of 2jy6 with 1oa8

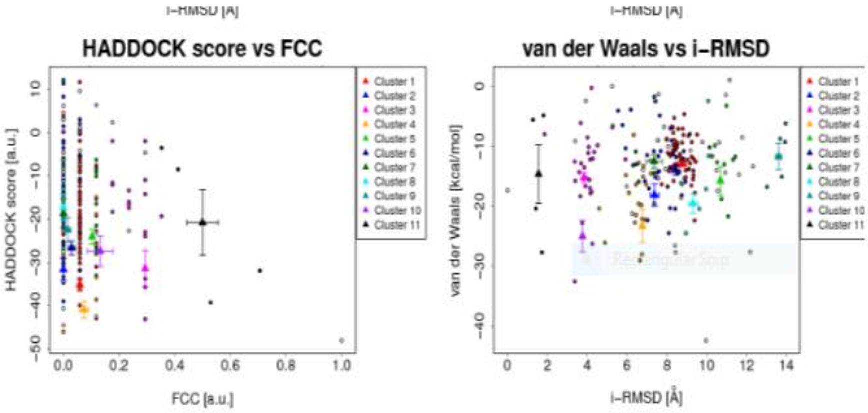
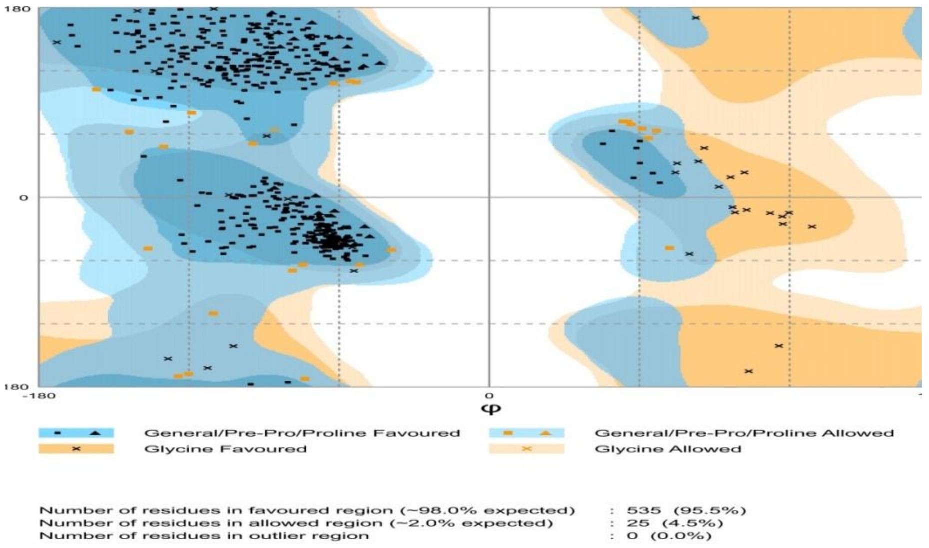


Fig. 3.16 Clusters of 1j46

Step 4: Validation - The complexes obtained through protein-protein interactions are validated by ramachandran plot. This plot is used to check the quality of the protein and it is analyzed by examining the amino acids in allowed region and disallowed region. The ramachandran plot for the complex shown in Fig. 3.15 is depicted in Fig. 3.17. This validation with ramachandran plot proves that the complex is valid and all the complexes are validated in this manner.



```
Number of residues in favoured region (~98.0% expected)   : 535 (95.5%)
Number of residues in allowed region (~2.0% expected)     : 25 (4.5%)
Number of residues in outlier region                      : 0 (0.0%)
```

Fig. 3.17 Ramachandran plot of 2jy6 with 1oa8

In this manner, a set of 313 validated interacted complexes is developed and referred as PP corpus.

3.3 DESIGN OF FEATURES AND DATASETS

Data preparation is significant because the raw data is converted to a final dataset that can be trained for developing the models which gives high prediction rate. The main idea of feature engineering is to extricate the defined features from docked complexes to build predictive models. Feature extraction is one of the crucial steps as it influences the development of predictive models in machine learning task. Three datasets have been developed to assist traditional machine learning and contemporary deep learning approaches and to provide suitable solution for the objectives considered.

Protein-Ligand Dataset (PLD)

The discriminative features like energy calculations and physical properties of protein and ligand are identified from 307 docked complexes of PL corpus. The features are extracted using autodock, autodock vina and pymol. Energy calculations such as vanderwaals energy, desolvation energy, torsional energy, electrostatic energy, inhibition constant, ligand efficiency etc., are extracted using autodock. The physical properties like molecular weight of both protein, ligand, complex; surface area of solvent for protein, ligand and complex are extricated using pymol and vina.

Energy calculations are important in predicting binding affinity as binding energy facilitates in calculating binding affinity. Binding energy is determined by other energies like vanderwaals, salvation etc., where the vanderwaals energy plays role in attraction and repulsion of atoms, molecules and surfaces. Vanderwaals is the weakest attraction, along with this bond covalent and ionic bonds helps in proper function of protein. Solvation energy is important because the complex is fed to a solvent where the chemical reaction occurs between the bonds and the energy is measured. Torsional energy is measured as the dihedral angles in the protein structure where the energy changes due to rotation and the best value of energy is calculated. Electrostatic energy is the energy that occurs due to reaction of charged atoms and it is measured to calculate the changes in the energy. Inhibition constant measures the potency of inhibitor.

Binding affinity is the measure of the strength between the structure and ligand. Binding energy is the energy of complex that indicates the bond of ligand with the protein. Physical properties are measured for ligand, complex and protein structure where the changes occur in molecular weight, atom count, charge and surface area of all the solvents are changed due to docking where they are collected from pymol. Thus a total of 27 features are defined and extracted to form feature vectors of the dataset. The feature values are normalized using min-max normalization. Binding affinity values are derived from autodock and augmented with feature vectors to facilitate supervised learning of regression. This dataset with 307 instances is significant in building accurate binding affinity predictive model, since the binding affinity from docked complexes assist in drug potency and it is named as PLD.

The detailed description of feature extraction and dataset creation will be explained in chapter 4, section 4.2.

Protein-Mutated-Ligand Dataset (PMLD)

In previous case, the protein structures are not mutated where the changes in the structure and sequences cannot be monitored. It is essential to monitor the changes in the sequences, due to repeat mutation and the respective contributors are analyzed and captured. This can be achieved using scoring functions and sequence descriptors in binding affinity prediction. The significant features such as energy calculations, sequence descriptors and scoring functions are recognized from 307 mutated docked complexes of PML corpus. Energy profiles like binding energy, inhibition constant, intermolecular energy, desolvation energy, electrostatic energy, total internal energy and torsional energy are defined. Sequence descriptors consist of amino acid composition, autocorrelation, Composition-Transition-Distribution (CTD), Quasi-sequence-order descriptors, Pseudo amino acid composition and profile-based descriptors. The scoring functions include cyscore and rfscore where cyscore posses hydrophobic free energy, cyscore, van der waals interaction energy, hydrogen-bond interaction and ligand’s conformational entropy. Rfscore consists of thirty six values and each feature will denote the number of occurrences of a particular protein-ligand atom type pair interacting within a certain distance range. Autodock vina scores have $\Delta G_{gauss,}$ $\Delta G_{repulsion,}$ $\Delta G_{hydrophobic}$ and ΔG_{Hbond}. The features are extricated using autodock, autodock vina, R script.

Energy calculations are significant to predict binding affinity and the energy calculations defined in the previous case are considered here also. Sequence descriptors are measured where the changes occur in sequence of amino acid, structure, protein folding and binding gets changed due to mutation. Cyscore is extracted for interaction energy profiles like hydrogen-bond, vanderwaals and cyscore. Scores from autodock vina are squeezed for free energy profiles of repulsion, hydrogen bond and hydrophobic. Rf score contains 36 features where the commonly occurred atoms in both the ligand and structure are computed. Cyscore, rfscore and autodock vina scores are squeezed through unix where the sequence descriptors are extracted using R script. Thus a total of 509 features are defined and extracted to form feature vectors of the dataset. The features are normalized using min-max normalization. Binding affinity values are derived from autodock and amplified with feature vectors to facilitate regression task. This dataset is named as PMLD. The number of features for PMLD dataset is high when compared with PLD and PPD dataset, as the sequences of the protein structures are mutated and the protein structure changes due to mutation. The changes are monitored along with the sequence descriptors and scoring functions.The detailed description of feature extraction and dataset creation will be explained in chapter 5, section 5.2.

Protein-Protein Dataset (PPD)

Protein-protein interaction is essential as it aids in knowing hidden functions between the macromolecules. The hidden functions NIS, interfacial contacts with binding affinity assist in drug development for disorders. The discriminative features such as energy calculations, interfacial contacts and physiochemical properties are identified and defined. These features are extracted from 313 interacted complexes of PP corpus. Energy features include haddock score, cluster size, Root Mean Squared Deviation (RMSD), vanderwaals energy, desolvation energy, electrostatic energy, Z-score, Buried surface area, violation energy. Energy values of desolvation, vanderwaals, and electrostatic energy are significant to get the binding affinity score. The interfacial contacts comprises number of interface pairwise contacts and NIS properties. Physio-chemical properties like amino acid composition, molecular weight, theoretical pl, negatively charged residues, positively charged residues, carbon, hydrogen, nitrogen, oxygen, sulfur, instability index, aliphatic index, aromaticity, Grand Average of Hydropathy (GRAVY) are considered here. Energy values are taken from haddock and physiochemical properties are derived using R script.

Haddock score is the unit of Energy terms is given as kcal/mol and 1.0, 0.2 and 1.0 are a weighted sum of intermolecular energies (vdw, elec, desolvation). Vanderwaals intermolecular energy (EVDW), electrostatic intermolecular energy (EELEC), EDesolvation desolvation energy. The haddock score with the minimum energy value is considered as a feature value. The size of the cluster is the number of amino acid compositions in the most populated cluster. The cluster which occurs first with the minimum energy is chosen for finding the cluster size. The root mean squared deviation is used to validate the docking with respect to biological configuration. RMSD is the measure of the average distance between the atoms. Desolvation energy is the static van der Waals energy, were the lose of the interaction between substance or organic compound and solvent upon binding describes the energy. For example electro-statically bound particles, dissociate by releasing water in an aqueous solution. Vanderwaals energy is the attraction of intermolecular forces between molecules. Hydrogen bonding, dipole interactions are the examples of vanderwaals energy.

Electrostatic energy is the long term interaction between charged atoms. The example of electrostatic energy is, to hold balloon against ceiling. The z-score represents the standard deviations the haddock score of a given cluster, is separated from the mean of all clusters. Buried surface area predicts different measures of flexibility. Violation energy is calculated based on dihedral angle, distance, RDC, etc. Interfacial contacts calculate number of interface residue pair wise contacts, for each complex. NIS properties such as percentage of polar, apolar charged residue are used here. Physical and chemical properties are extracted to identify the changes in the structure, owing to interaction. Physical properties like molecular weight, number of aminoacids, theoritcal pl etc., are taken for consideration. Chemical properties such as negatively charged residues, positively charged residues, carbon, hydrogen, nitrogen, oxygen, sulfur, instability index, aliphatic index, aromaticity and GRAVY. Physio-chemical properties along with energy calculations facilitate in predicting binding affinity. Thus a total of 56 features are defined and extracted to form feature vectors of the dataset. Binding affinity values are derived from haddock and augmented with feature vectors to facilitate regression task. The features are normalized using min-max normalization. The dataset with 313 instances of 56 dimensions is developed and named as PPD.

The detailed description of feature extraction and dataset creation dataset will be explained in chapter 6, section 6.2. The profile of datasets is given in Table 3.3.

Table 3.3 Profile of Datasets

Datasets	Approach	Count of Protein Structures and Ligand	Total number of Features	Total number of Instances
PLD	Protein-Ligand Docking	17 structrues and 18 Ligands	27	307
PMLD	Protein Mutated-Ligand	17 structures and 18 Ligands	509	307
PPD	Protein-protein Dataset	626 protein structures	56	313

3.4 TRAINING AND TESTING

The datasets mentioned in the above section are used to train the predictive models to predict the binding affinity of spinocerebellar ataxia. The dataset is split into training and testing as 90% and 10% respectively. The algorithm learns the data through annotations and predicts the output if unknown data is given. In this research work the problem is considered as regression task, the output variable is solitary and the input variables are supplementary. The dataset contains independent variables (X) and dependant variable (Y). The features from three datasets are trained and validated using regression algorithms where the training parameters are modified while training and validated with evaluation metrics. The supervised regression algorithms like support vector regression, random forest, artificial neural network and linear regression are engaged to build the predictive models employing PLD, PMLD and PPD datasets. In deep learning approach the same three datasets are used to build the predictive models by training sequential DNN, functional DNN and DNN with customized layers. The hyper parameters such as learning rate, epochs, dropout, optimizers etc., are used in DNN models to fine tune the predictive models.

The common technique used to evaluate the prediction rate of a regression algorithm is k-fold cross validation. In k-fold cross validation the dataset is split into training and testing. Initially the datasets is split into 10 fold cross validation where the training data is 90% and testing data is 10%. In this work, the k value is fixed as 10 and the entire dataset is divided into 10 folds where 9 folds are used for training set and 1 fold for testing. 10 fold cross validation is used to evaluate the performance of the models. The performance metrics is averaged across all 10 folds. As every data goes into testing, best parameter combination is found that reduces the error rate. The test set is used to evaluate the performance of the models with various metrics. Evaluation metrics used in this work are explained variance score, mean squared error, R2 score, mean absolute error, median absolute error, root mean squared error, correlation coefficient and p value. These eight evaluation metrics are chosen as these are the common evaluation metrics for evaluating regression models. Each metric is explained below.

Explained variance score

Explained variance score is the measure of the difference between observed values and the average of predicted values. It is calculated using the equation given in 3.1.

$$\text{Explained\_variance}\ (\mathrm{y}, \mathrm{y1}) \ = \ 1 - \frac{\mathrm{var}\{\mathrm{y}-\mathrm{y1}\}}{\mathrm{var}\{\mathrm{y}\}} \qquad \textbf{(3.1)}$$

where y is the true value, y1 is the predicted value. The higher the explained variance score, the higher is the prediction rate.

Mean squared error

Mean squared error is the average of the square of the errors and it is calculated using the equation given in 3.2. The lower the error rate, the closer to the best fit of the model.

$$MSE(y, y1) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}\ -1} (\mathrm{y_i} - \mathrm{y1_i})^2 \qquad \textbf{(3.2)}$$

R2 score

R2 score is the proportion of the variance in the dependent variable that is predictable from the independent variable where the score is calculated using the equation given in 3.3. The higher the R2 score, the model will be better.

R2 = 1 – (First Sum of Errors / Second Sum of Errors) **(3.3)**

Root mean squared error (RMSE)

Root mean squared error is the standard deviation of the predicted errors. Error can be calculated using the equation given in 3.4.

$$RMSE = \sqrt{}(\mathrm{f} - \mathrm{o})^2 \qquad \textbf{(3.4)}$$

where f = forecasts (expected values or unknown results), o = observed values (known results). Lower values of RMSE indicate best fit for the model.

Mean absolute error (MAE)

Mean absolute error is the average of all absolute errors and it is calculated using the equation given in 3.5. Mean absolute error should be lower for the best model.

$$MAE = 1/n \sum_{i=1}^{n} |\mathrm{x_i} - \mathrm{x}| \qquad \textbf{(3.5)}$$

where n = the number of errors, $|x_i – x|$ = the absolute errors.

Median absolute error

Median absolute error is a robust measure of the data spread out and it is calculated using the equation given in 3.6.

$$\text{MAD} = \text{median}(|Yi - \text{median}(Yi|) \qquad \textbf{(3.6)}$$

The lower the error rate, the better is the model. In general the error rate should be minimal and the scores should be high, to achieve the highest prediction rate.

Correlation coefficient

Correlation coefficient is used in statistics to measure the relationship between two variables. Pearson's correlation is commonly used in linear regression. The value of correlation coefficient lies between -1 to 1. The value 1 indicates the strong relationship and the value -1 indicates strong negative relationship. 0 indicates there is no relationship between two variables. The correlation coefficient is calculated using the equation given in 3.7.

$$r = \frac{\mathrm{n}(\sum \mathrm{xy}) - (\sum \mathrm{x})(\sum \mathrm{y})}{\sqrt{[\mathrm{n}\sum \mathrm{x}^2 - (\sum \mathrm{x}^2)][\mathrm{n}\sum \mathrm{y}^2 - (\sum \mathrm{y}^2)]}} \qquad \textbf{(3.7)}$$

where n is the number of instances, x is the independent variable and y is the dependent variable.

P-value

P-value is used in hypothesis testing to support or reject null hypothesis. It is the evidence against null hypothesis. The alpha value is set as 0.05. If the p value is smaller than 0.05 then the null hypothesis is rejected, then there is significance between the values. If the value is greater than 0.05 then hypothesis is weak, then the null is not rejected and it shows that there is no significance between the values. If the value lies below 0.01 or equal to 0.01 then there is highly significance between the values.

4.3 SUMMARY

The main component of research is problem modelling and it has been explained in detail in this chapter with various tasks such as corpus development, features and dataset creation, training and testing. The corpus development process for three corpuses and the composition of respective datasets has been presented. The training and testing methods adopted in this research has been elucidated. The performance metrics used for evaluating the predictive models are also described in this chapter with the methods. Various predictive models built with PLD dataset using regression algorithms such as linear regression, support vector regression, random forest, artificial neural network will be presented in chapter 4. The predictive models built with PMLD dataset using traditional regression algorithms like linear regression, support vector regression, random forest, artificial neural network will be discussed in chapter 5. The predictive models built with PPD dataset using supervised regression algorithms such as linear regression, support vector regression, random forest, artificial neural network will be elucidated in chapter 6. The implementations of deep neural network architectures for building predictive models are explained in chapter 7.