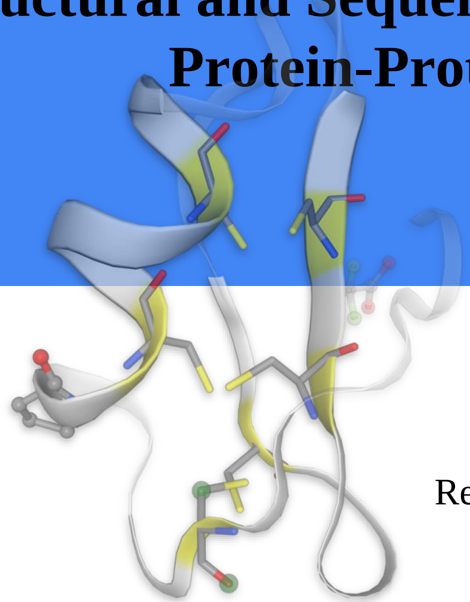
Predictive Analysis of protein structures: Parsing PDB data for Structural and Sequential Insight and Active Site Prediction in Protein-Protein Interaction at residue level



Submitted by

Godhuli Das
Examination Roll: M6TCT24023
Registration No:160099 of 2021 – 2024
Class Roll: 02110504028

Under the guidance of

Prof. (Dr.) Subhadip Basu

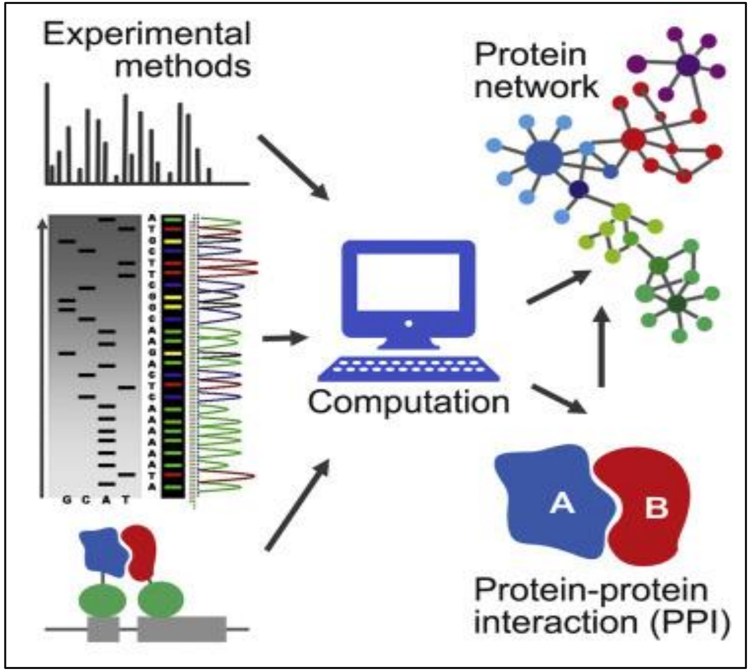
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, JADAVPUR UNIVERSITY

Outline

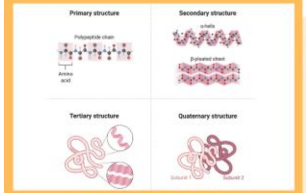
1. Introduction
2. Literature Review
3. Research Objectives
4. Proposed Methodology
5. Detailed Discussion
6. Experiments and Evaluation
7. Conclusions and Future Work
8. References

Introduction

- Computational methods and deep learning technologies have shown great promise in predicting active sites in protein complexes. Recent studies have demonstrated the efficacy of these approaches, but challenges remain, including the need for improved interpretability, scalability, and generalization across diverse protein structures.
- To address these challenges, future research may focus on advancing model architectures, integrating data from multiple sources, and developing new algorithmic innovations.
- The integration of computational methodologies and deep learning technologies offers promising avenues for accelerating drug discovery and protein engineering endeavours.
- Finally, the use of trusted PDB data from RCSB PDB and the incorporation of both structural and sequential information can improve the accuracy of active site prediction structures.



Literature Review



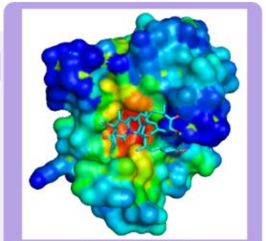
Protein Strcuture studies

1970s

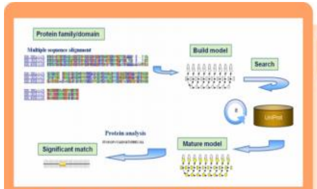
- Early beginnings:
- Establishment of PDB
- Initial protein structure studies

1980s-1990s

- Structural bioinformatics:
- Algorithm development
- PPI and binding site analysis



Structural BioInformatics



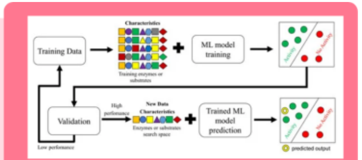
Sequential Analysis of protein chains

Early 2000s

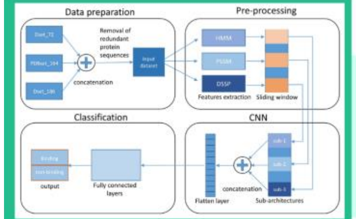
- Sequential analysis:
- PCA and HMMs for protein sequences
- Functional property prediction

Mid-2000s

- Active site prediction:
- Machine learning methods
- Enzyme active site identification



Machine learning methods in Active site prediction

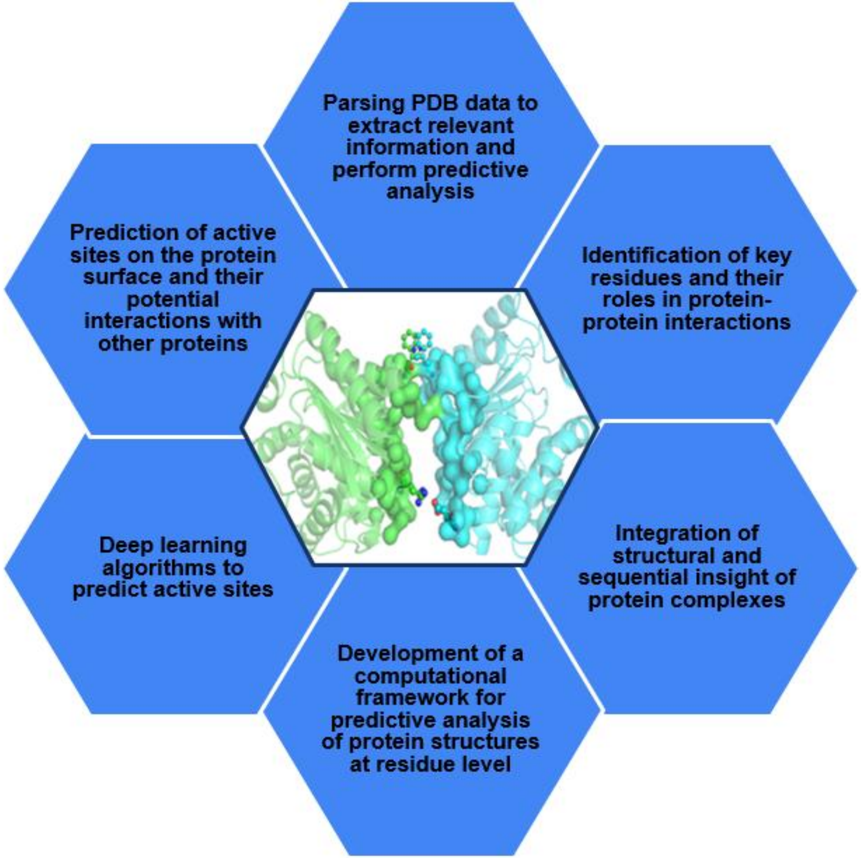


Deep Learning Approach in PPI and Active site prediction

Recent years

- Residue-level analysis:
- Molecular dynamics simulations
- Detailed PPI and binding site prediction

Research Objectives



Parsing PDB data to extract relevant information and perform predictive analysis

Prediction of active sites on the protein surface and their potential interactions with other proteins

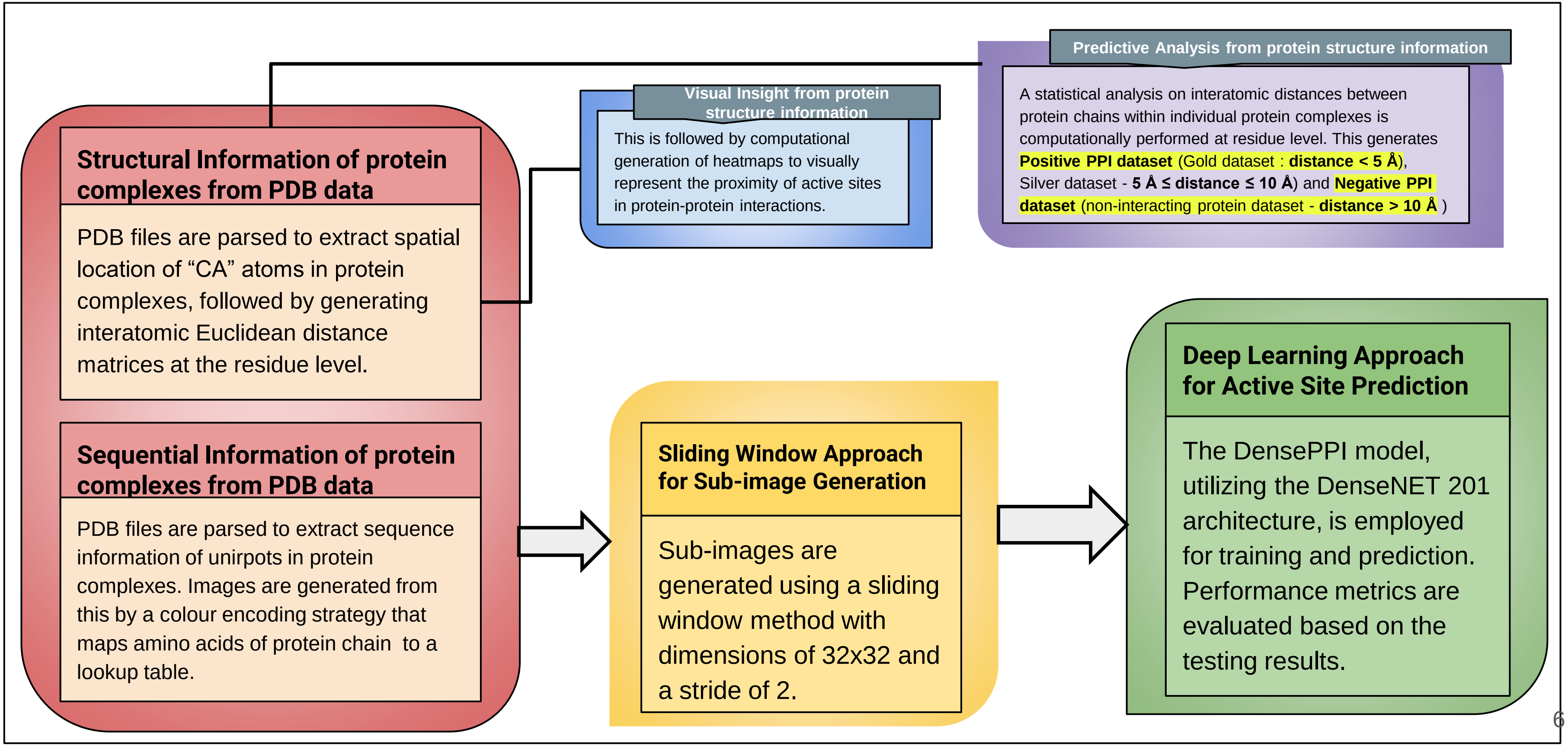
Identification of key residues and their roles in protein-protein interactions

Deep learning algorithms to predict active sites

Integration of structural and sequential insight of protein complexes

Development of a computational framework for predictive analysis of protein structures at residue level

Proposed Methodology



Structural Information of protein complexes from PDB data

PDB files are parsed to extract spatial location of “CA” atoms in protein complexes, followed by generating interatomic Euclidean distance matrices at the residue level.

Visual Insight from protein structure information

This is followed by computational generation of heatmaps to visually represent the proximity of active sites in protein-protein interactions.

Predictive Analysis from protein structure information

A statistical analysis on interatomic distances between protein chains within individual protein complexes is computationally performed at residue level. This generates **Positive PPI dataset** (Gold dataset : **distance < 5 Å**), Silver dataset - **5 Å ≤ distance ≤ 10 Å**) and **Negative PPI dataset** (non-interacting protein dataset - **distance > 10 Å**)

Sequential Information of protein complexes from PDB data

PDB files are parsed to extract sequence information of unirpots in protein complexes. Images are generated from this by a colour encoding strategy that maps amino acids of protein chain to a lookup table.

Sliding Window Approach for Sub-image Generation

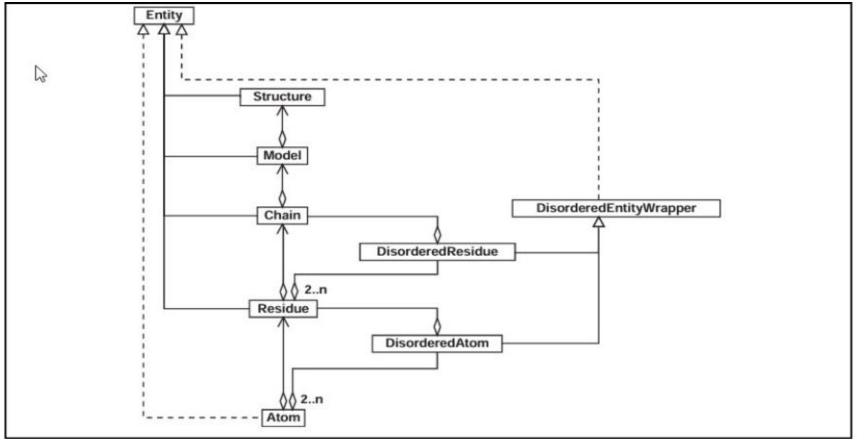
Sub-images are generated using a sliding window method with dimensions of 32x32 and a stride of 2.

Deep Learning Approach for Active Site Prediction

The DensePPI model, utilizing the DenseNET 201 architecture, is employed for training and prediction. Performance metrics are evaluated based on the testing results.

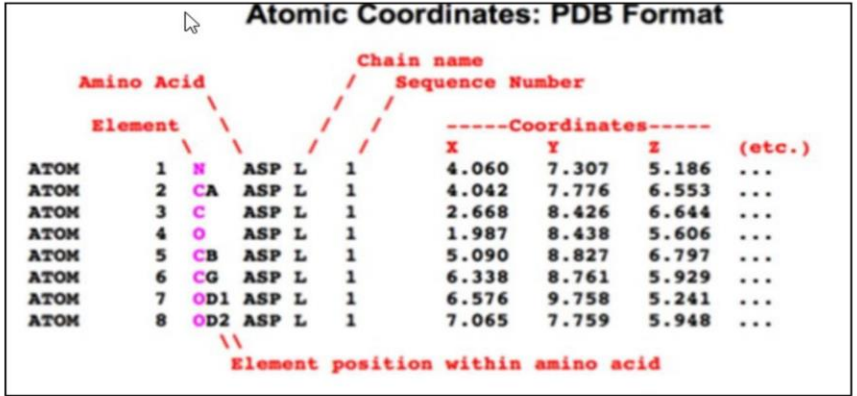
PDB Parsing for Protein Structure Information Generation by Biopython library

The PDB parsing process for generating protein structure information is efficiently handled by the Biopython library. Through Biopython, a robust suite of tools for protein data handling and analysis is provided.



Detailed protein structure information is extracted from PDB files using the `PDBParser` class, which delves into the SMCRA (Structure, Model, Chain, Residue, Atom) data structure.

Atomic coordinates, residue names, and chain identifiers are accessed, enabling the creation of distance matrices between atoms or residues for structural analysis (demonstrated in upcoming slide).



Atomic Coordinates: PDB Format

```
                              Chain name
         Amino Acid          /  Sequence Number
                   \        /  /
      Element       \      /  /        -----Coordinates-----
             \       \    /  /          X        Y        Z       (etc.)
ATOM       1  N    ASP L   1          4.060    7.307    5.186    ...
ATOM       2  CA   ASP L   1          4.042    7.776    6.553    ...
ATOM       3  C    ASP L   1          2.668    8.426    6.644    ...
ATOM       4  O    ASP L   1          1.987    8.438    5.606    ...
ATOM       5  CB   ASP L   1          5.090    8.827    6.797    ...
ATOM       6  CG   ASP L   1          6.338    8.761    5.929    ...
ATOM       7  OD1  ASP L   1          6.576    9.758    5.241    ...
ATOM       8  OD2  ASP L   1          7.065    7.759    5.948    ...
                 \\
                  Element position within amino acid
```

PDB files are downloaded from the RCSB PDB repository, streamlining protein structure research workflows.

Pairwise Interatomic Distance Calculation of protein complexes and Matrix Generation

With the extracted structural information from PDB data parsed, pairwise distances between alpha carbon atoms (`CA`) of protein chains are calculated.

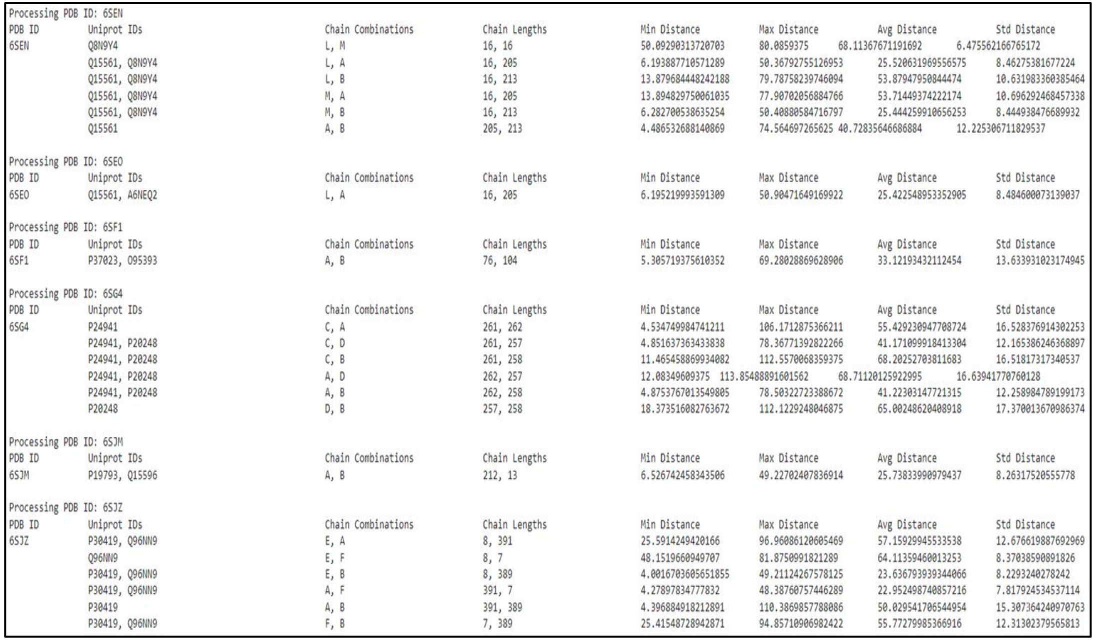
The specific function iterates over all possible combinations of protein chain pairs within each structure. For each pair, the **Euclidean distance between corresponding alpha carbon coordinates** is computed to construct a distance matrix. This matrix quantitatively represents the spatial separation between residues in the protein chains.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (coord1_i - coord2_i)^2}$$

Quantitative Insights from PDB data parsed

Once the distances between protein chains are computed, various statistical measures are derived to characterize their spatial relationships. This process includes:

- Computing **minimum, maximum, average distances, and the standard deviation of distances**.
- Providing **quantitative insights** into the proximity and distribution of protein chains.
- Facilitating the identification of key structural features and interaction patterns.
- These steps offer a comprehensive understanding of the spatial relationships within protein structures.

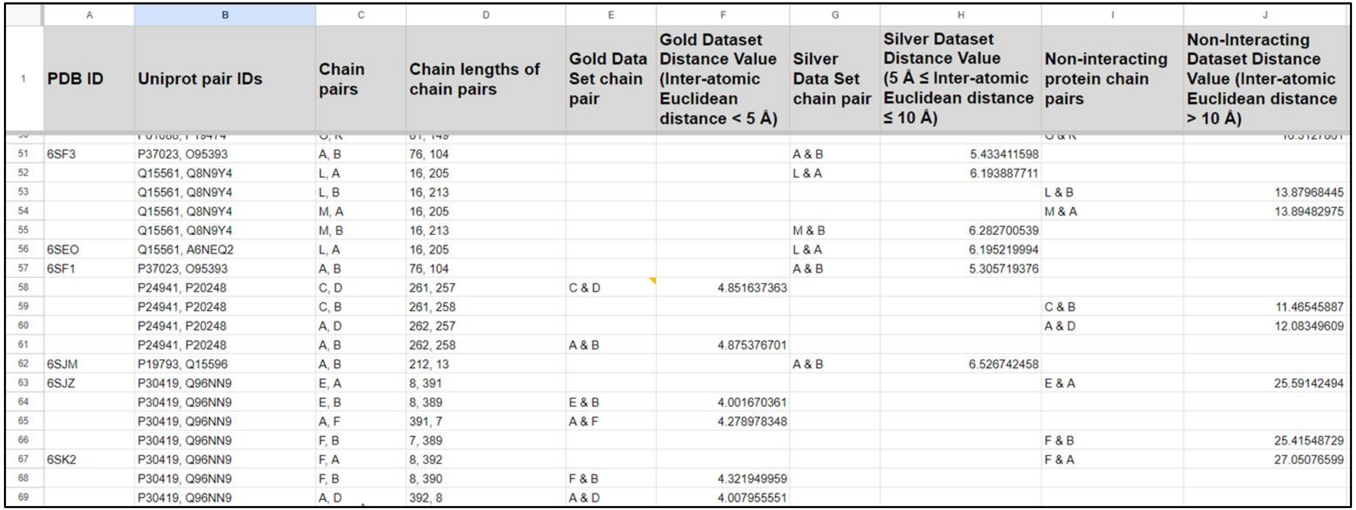


```
Processing PDB ID: 6SEN
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SEN        Q8N9Y4                            L, M                  16, 16               50.09290313720703    80.0859375      68.11367671191692       6.475562166765172
            Q15561, Q8N9Y4                    L, A                  16, 205              6.193887710571289    50.36792755126953   25.520631969556575  8.46275381677224
            Q15561, Q8N9Y4                    L, B                  16, 213              13.879684448242188   79.78758239746094   53.87947950844474   10.631983360385464
            Q15561, Q8N9Y4                    M, A                  16, 205              13.894829750061035   77.90702056884766   53.71449374222174   10.696292468457338
            Q15561, Q8N9Y4                    M, B                  16, 213              6.282700538635254    50.40880584716797   25.444259910656253  8.444938476689932
            Q15561                            A, B                  205, 213             4.486532688140869    74.564697265625 40.72835646686884       12.225306711829537

Processing PDB ID: 6SEO
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SEO        Q15561, A6NEQ2                    L, A                  16, 205              6.195219993591309    50.90471649169922   25.422548953352905  8.484600073139037

Processing PDB ID: 6SF1
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SF1        P37023, 095393                    A, B                  76, 104              5.305719375610352    69.28028869628906   33.12193432112454   13.633931023174945

Processing PDB ID: 6SG4
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SG4        P24941                            C, A                  261, 262             4.534749984741211    106.1712875366211   55.429230947708724  16.528376914302253
            P24941, P20248                    C, D                  261, 257             4.851637363433838    78.36771392822266   41.171099918413304  12.165386246368897
            P24941, P20248                    C, B                  261, 258             11.465458869934082   112.5570068359375   68.20252703811683   16.51817317340537
            P24941, P20248                    A, D                  262, 257             12.08349609375  113.85488891601562      68.71120125922995       16.63941770760128
            P24941, P20248                    A, B                  262, 258             4.8753767013549805   78.50322723388672   41.22303147721315   12.258984789199173
            P20248                            D, B                  257, 258             18.373516082763672   112.1229248046875   65.00248620408918   17.370013670986374

Processing PDB ID: 6SJM
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SJM        P19793, Q15596                    A, B                  212, 13              6.526742458343506    49.22702407836914   25.73833990979437   8.26317520555778

Processing PDB ID: 6SJZ
PDB ID      Uniprot IDs                       Chain Combinations    Chain Lengths        Min Distance         Max Distance        Avg Distance        Std Distance
6SJZ        P30419, Q96NN9                    E, A                  8, 391               25.5914249420166     96.96086120605469   57.15929945533538   12.676619887692969
            Q96NN9                            E, F                  8, 7                 48.1519660949707     81.8750991821289    64.11359460013253   8.37038590891826
            P30419, Q96NN9                    E, B                  8, 389               4.0016703605651855   49.21124267578125   23.636793939344066  8.2293240278242
            P30419, Q96NN9                    A, F                  391, 7               4.27897834777832     48.38760757446289   22.952498740857216  7.817924534537114
            P30419                            A, B                  391, 389             4.396884918212891    110.3869857788086   50.029541706544954  15.307364240970763
            P30419, Q96NN9                    F, B                  7, 389               25.41548728942871    94.85710906982422   55.77279985366916   12.31302379565813
```

Predictive Analysis from PDB data parsed

The dataset summary is being prepared that includes euclidean distances between chain pairs based on the overall minimum distance thereby analysing a prediction on the interaction classifications as : **Positive PPI dataset** (*Gold dataset : distance < 5 Å*), *Silver dataset : 5 Å ≤ distance ≤ 10 Å*) and **Negative PPI dataset** (*non-interacting protein dataset : distance > 10 Å*)

This computational process, supported by functions to retrieve UniProt IDs and chain information from mmCIF files, offers valuable insights into protein structural characteristics.

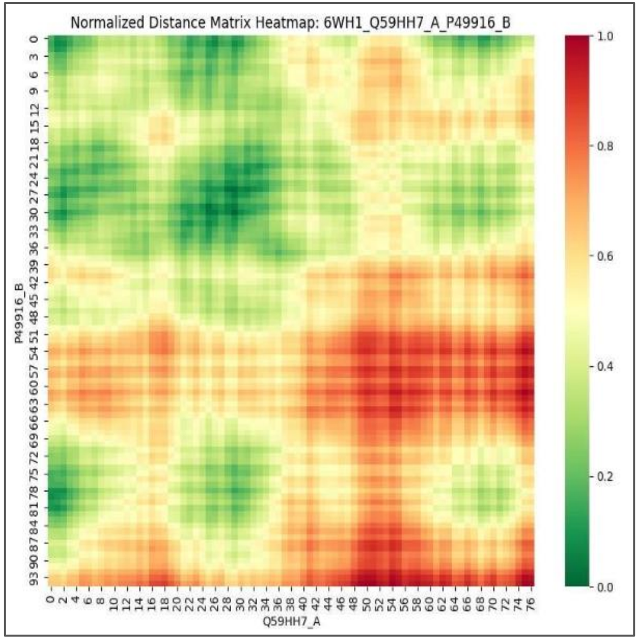


	A	B	C	D	E	F	G	H	I	J
1	PDB ID	Uniprot pair IDs	Chain pairs	Chain lengths of chain pairs	Gold Data Set chain pair	Gold Dataset Distance Value (Inter-atomic Euclidean distance < 5 Å)	Silver Data Set chain pair	Silver Dataset Distance Value (5 Å ≤ Inter-atomic Euclidean distance ≤ 10 Å)	Non-interacting protein chain pairs	Non-Interacting Dataset Distance Value (Inter-atomic Euclidean distance > 10 Å)
51	6SF3	P37023, O95393	A, B	76, 104			A & B	5.433411598		
52		Q15561, Q8N9Y4	L, A	16, 205			L & A	6.193887711		
53		Q15561, Q8N9Y4	L, B	16, 213					L & B	13.87968445
54		Q15561, Q8N9Y4	M, A	16, 205					M & A	13.89482975
55		Q15561, Q8N9Y4	M, B	16, 213			M & B	6.282700539		
56	6SEO	Q15561, A6NEQ2	L, A	16, 205			L & A	6.195219994		
57	6SF1	P37023, O95393	A, B	76, 104			A & B	5.305719376		
58		P24941, P20248	C, D	261, 257	C & D	4.851637363				
59		P24941, P20248	C, B	261, 258					C & B	11.46545887
60		P24941, P20248	A, D	262, 257					A & D	12.08349609
61		P24941, P20248	A, B	262, 258	A & B	4.875376701				
62	6SJM	P19793, Q15596	A, B	212, 13			A & B	6.526742458		
63	6SJZ	P30419, Q96NN9	E, A	8, 391					E & A	25.59142494
64		P30419, Q96NN9	E, B	8, 389	E & B	4.001670361				
65		P30419, Q96NN9	A, F	391, 7	A & F	4.278978348				
66		P30419, Q96NN9	F, B	7, 389					F & B	25.41548729
67	6SK2	P30419, Q96NN9	F, A	8, 392					F & A	27.05076599
68		P30419, Q96NN9	F, B	8, 390	F & B	4.321949959				
69		P30419, Q96NN9	A, D	392, 8	A & D	4.007955551				

Visual Insight from Protein Structure Data by Heatmaps

For each pair of protein chains within a given structure, a distance matrix heatmap is generated using seaborn and matplotlib libraries.

The heatmap visualizes the distances between alpha carbon atoms, with a color gradient representing varying degrees of proximity. Specifically, the **heatmap employs a color spectrum ranging from green (indicating shorter distance PPIs) to red (indicating longer distances PPIs)**, allowing for intuitive interpretation of spatial relationships within protein structures

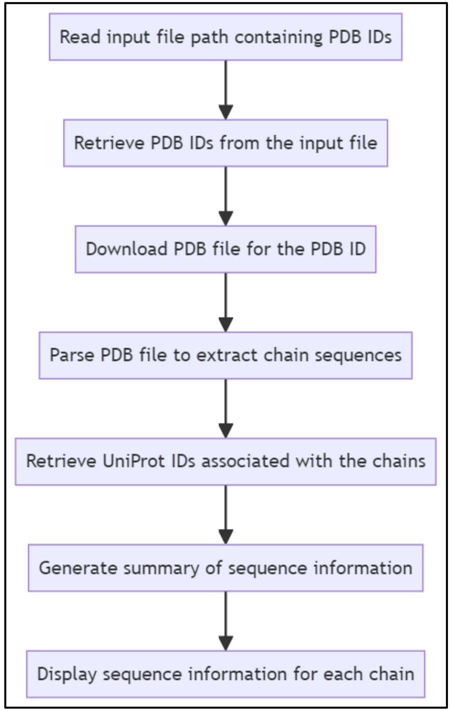


PDB Parsing for Protein Sequence Information Generation

- Python script leverages Biopython package for the retrieval and analysis of protein sequences from the Protein Data Bank (PDB).

- Aids in understanding primary protein structures, including amino acid sequences and associated UniProt IDs.

- This protein sequence data facilitates downstream analyses (e.g., sequence alignment, functional annotation) for the current or other research in structural biology and bioinformatics.



Read input file path containing PDB IDs

Retrieve PDB IDs from the input file

Download PDB file for the PDB ID

Parse PDB file to extract chain sequences

Retrieve UniProt IDs associated with the chains

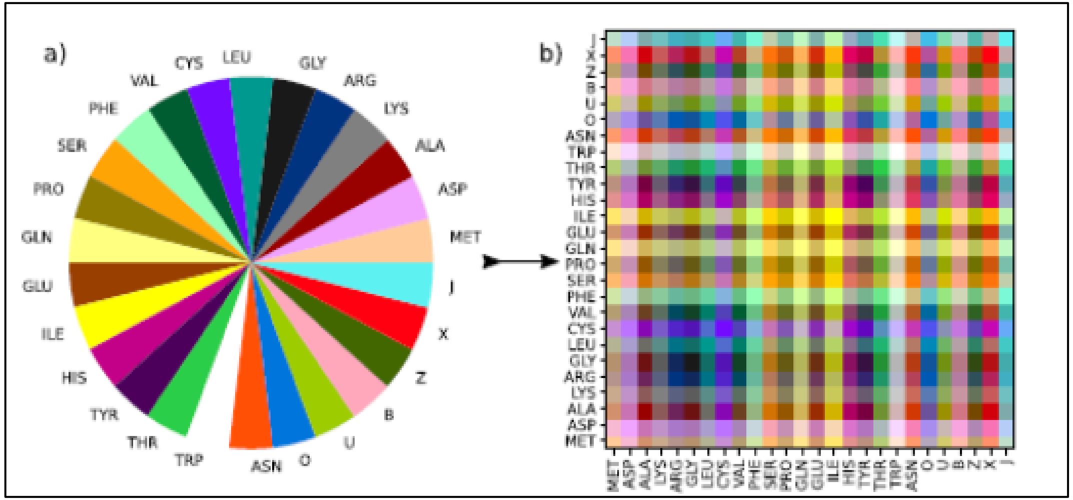
Generate summary of sequence information

Display sequence information for each chain

Image Generation from Protein Sequence Information

Bigram Interaction Mapping:

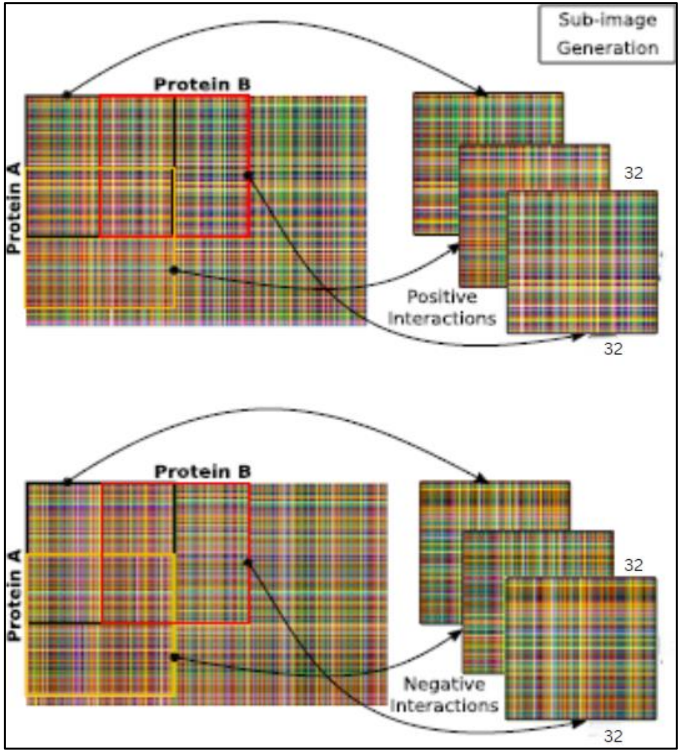
- A color encoding strategy maps bigram interactions of amino acids to a lookup table, visualizing their interactions.
- This method assigns each amino acid one of 26 distinct RGB colors, ensuring equal importance.
- The resulting 26 × 26 color matrix (CMAP) represents residue interactions.



Sliding Window Approach for Sub-image Generation

- **Sliding Window Approach**: A sliding window approach with fixed dimensions (32x32) and stride (2) is used to create sub-images from the original PPI images.

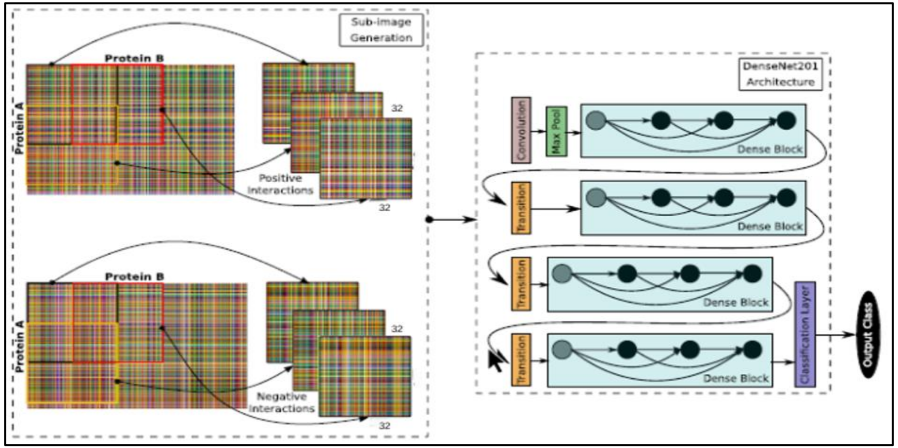
- **Sub-Image Generation for Training and Testing**: The generated sub-images are categorized into positive and negative PPIs and are prepared for training and testing the model.



Training DensePPI model by Deep Learning Approach

The sub-images are processed by the DensePPI model, which incorporates the DenseNet 201 architecture. The DenseNet 201 architecture is depicted in next slide, showcasing its dense blocks, transition layers, and output layer.

- **Training Loop**: The training loop iterates over a specified number of epochs (10 in this case). For each epoch, the model is set to training mode and processes the training data in batches. The optimizer updates the model weights based on the computed gradients.

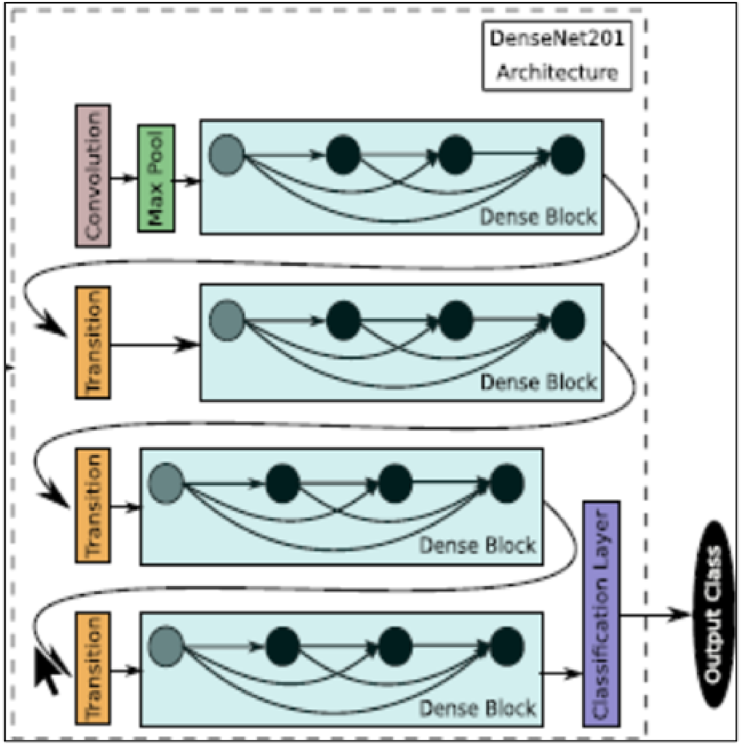


- **Confidence Score and Thresholding**: The classification strategy takes into account the confidence scores for each sub-image. A threshold of 0.5 is applied to determine the final class label for the original PPI. For values >=0.5 are considered 1 hence, positive and <0.5 considered 0, hence negative class.

DenseNet 201 Architecture

DenseNet developed by Huang et al. to overcome deep CNN issues of Vanishing-gradient problem and reduced information flow in deep networks.It has unique feature: each layer connects to every other layer, leading to L(L+1)/2 direct connections in an L-layer network. Configuration setup:

- Average pooling between layers.
- Learning rate: 0.001, Momentum: 0.9.
- Binary classification using categorical crossentropy loss.
- Optimizer: Stochastic Gradient Descent (SGD).
- Mini-batch gradient descent with batch size of 32, over 10 epochs.



Results and Discussions

Training and Testing Accuracy:

- The training accuracy improved steadily over the epochs, starting from 84.83% in the first epoch and reaching up to 98.78% in the tenth epoch.
- The testing accuracy achieved was 95.44%, indicating the model's ability to generalize well to unseen data.

Training Loss:

- The training loss consistently decreased over the epochs, starting from 0.3681 in the first epoch and reducing to 0.0277 by the tenth epoch. This indicates effective learning and convergence of the model.

Prediction Loss:

- This is the loss calculated during the prediction phase of your model. It measures how well your model performs in terms of discrepancy between predicted and actual values. In this case, the prediction loss is 0.1414.

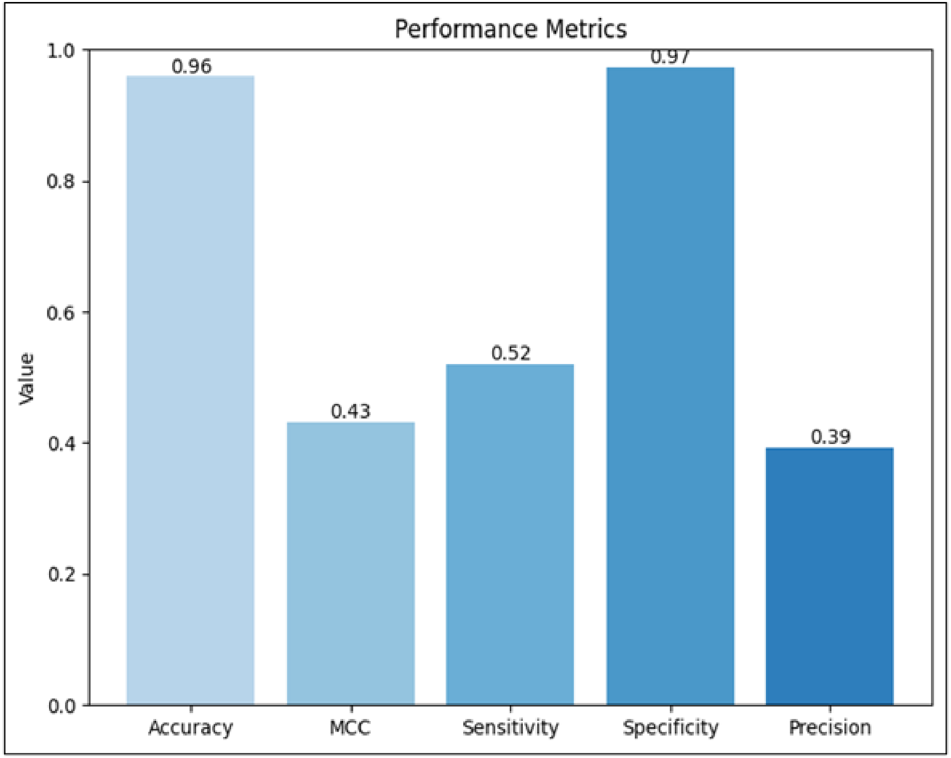
Prediction Accuracy

- This is the accuracy of your model's predictions on a test dataset. It measures the proportion of correctly classified instances out of the total instances. Here, the test accuracy is 0.95445, indicating that about 95.44% of the predictions were correct.

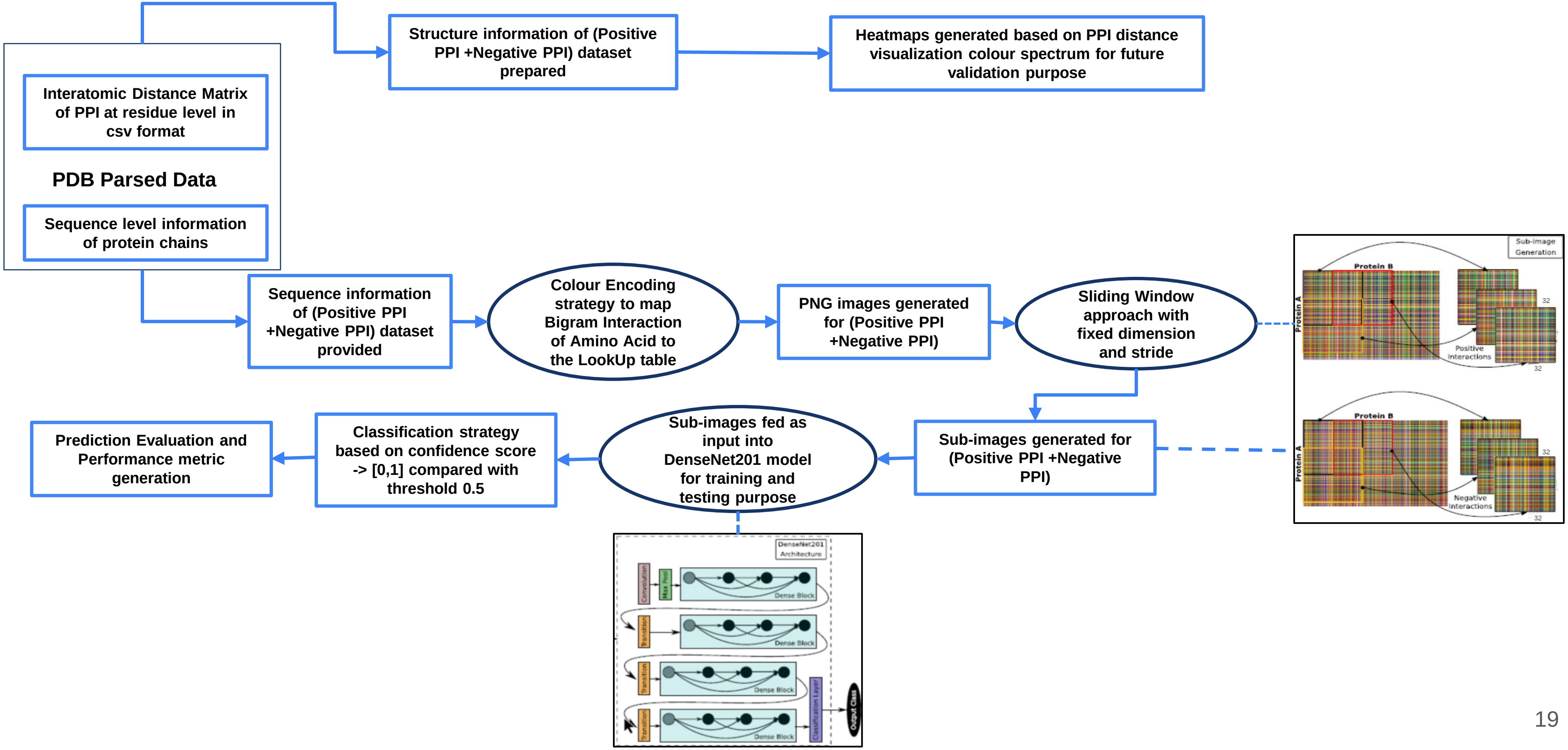
Evaluation Metrics

- Evaluation of DensePPI Model Performance Metrics showing Accuracy, MCC, Sensitivity, Specificity, Precision score.

- The model achieved an impressive overall accuracy of 95.84%, demonstrating its strong capability in correctly identifying interactions. The high accuracy is complemented by a high specificity (true negative rate), indicating that the model excels at correctly identifying non-interacting pairs. This is crucial for applications where false positives can be particularly detrimental.



Overall Workflow diagram for Active site prediction in PPI at residue level



PDB Parsed Data

Interatomic Distance Matrix of PPI at residue level in csv format

Sequence level information of protein chains

Structure information of (Positive PPI +Negative PPI) dataset prepared

Heatmaps generated based on PPI distance visualization colour spectrum for future validation purpose

Sequence information of (Positive PPI +Negative PPI) dataset provided

Colour Encoding strategy to map Bigram Interaction of Amino Acid to the LookUp table

PNG images generated for (Positive PPI +Negative PPI)

Sliding Window approach with fixed dimension and stride

Sub-images generated for (Positive PPI +Negative PPI)

Sub-images fed as input into DenseNet201 model for training and testing purpose

Classification strategy based on confidence score -> [0,1] compared with threshold 0.5

Prediction Evaluation and Performance metric generation

Conclusions

- This study presents a comprehensive approach utilizing deep learning techniques, specifically the DenseNet 201 architecture, to identify residue-level interactions in protein complexes using a sample of PDB data for *Homo sapiens* species.

- The methodology encompasses key processes such as novel dataset computation, alignment of multiple protein dataset parameters, model training, evaluation, and metrics generation. The model demonstrated robust performance, achieving an impressive overall accuracy of 95.84%. It excelled in accurately identifying non-interacting pairs, as evidenced by its high specificity, which is particularly crucial for applications where minimizing false positives is essential.

- However, this deep learning approach for model training and prediction remains in the experimental phase and requires further research and experimentation, which is currently ongoing. Future improvements will be addressed in the upcoming slide.

Scope for Future Improvement

Learning Rate Scheduling

- Implement dynamic learning rate techniques like annealing, warm-up, or cyclic rates to improve convergence and performance

Data Augmentation

- Use advanced techniques (rotation, flipping, scaling) and synthetic data generation (GANs) to enhance model robustness.

Optimizer Selection

- Experiment with various optimizers (Adam, RMSprop, AdaGrad) and their hyperparameters to find the best strategy for PPI prediction

Regularization Techniques

- Integrate dropout layers, weight decay, and batch normalization to improve model generalization

Model Tuning

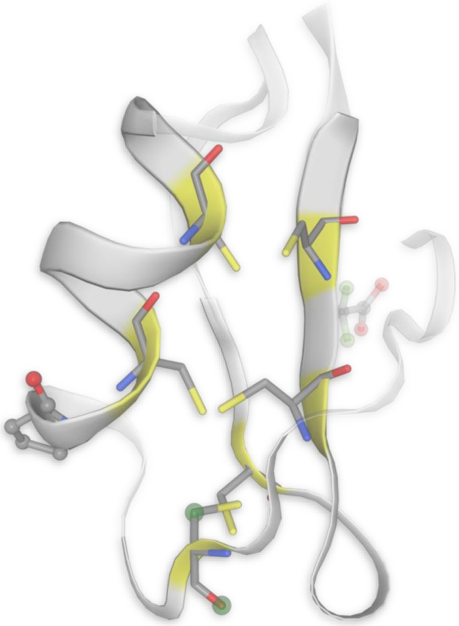
- Optimize hyperparameters using grid search, random search, or Bayesian optimization to reduce false negatives and improve accuracy

Architecture Exploration

- Investigate advanced architectures (Transformers, GNNs, CNNs) and ensemble methods for better robustness and accuracy

References

- Halsana, A.A., Chakroborty, T., Halder, A.K. and Basu, S., 2023. DensePPI: A Novel Image-based Deep Learning method for Prediction of Protein-Protein Interactions. IEEE Transactions on NanoBioscience.
- Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z. and Green, R.K., 2016. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic acids research, p.gkw1000.
- Hamelryck, T. and Manderick, B., 2003. PDB file parser and structure class implemented in Python. Bioinformatics, 19(17), pp.2308-2310.
- Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H., 2007. Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences, 104(11), pp.4337-4341.
- Ballester, P.J. and Mitchell, J.B., 2010. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics, 26(9), pp.1169-1175.
- Zhao, J., Cao, Y. and Zhang, L., 2020. Exploring the computational methods for protein-ligand binding site prediction. Computational and structural biotechnology journal, 18, pp.417-426.



Thank you !

Any Questions ?