CAPSDPP: A Capsule-Inspired Deep Learning Framework for Druggable Protein Prediction Using Hybrid Sequence-Based Features

Mujeebu Rehman[1]*, Rahu Sikander[2,3], Ali Ghulam[4]
[1]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China.
[2]Deparment of Computer Science, Suzhou University of Technology, Changshu, China.
[3]Jinnah University for Women, Karachi, Pakistan.
[4]Information Technology Centre, Sindh Agriculture University, Tandojam Sindh, Pakistan.
*E-mail: mujeeburehman@mails.guet.edu.cn

ABSTRACT

Druggable protein (DP) is a crucial aspect that should be identified to hasten drug development and reduce the cost of experiments. The Data is obtained with DrugBank and Swiss-Prot (2,518 proteins with 1,218 druggable and 1,300 non- druggable sequences) which was then filtered using CD-HIT (50% identity) to eliminate redundancy. Average Block Position-Specific Scoring Matrix (AB-PSSM, 200D) and Dipeptide Composition (DPC, 400D) are two free descriptors, which were recalled and combined into a hybrid feature of 600 dimensions, which is a reflection of evolutionary and local sequence information. The experimental results show that CapsDPP achieved a classification accuracy of 95 percent, which is higher than CNN models that were trained using single and combined features. The results of this work highlight the potential of capsule-based structures in the extraction of hierarchical associations between features in protein sequences and place CapsDPP in a strong and comprehensible model of computational drug discovery.

Keywords: Druggable Proteins, AB-PSSM, Dipeptide Composition, Hybrid Features, Capsule Network, CNN, Protein Sequence Classification, Deep Learning.

INTRODUCTION

Drug target identification (DPI) (Ghulam *et al*., 2025; Sikander *et al*., 2022), The discovery of new drug targets before isolating the targets via costly experimental screening is an important feature of modern-day pharmacological studies, as it allows identifying drug targets. Examples of traditional wet-lab technologies that are expensive, labor- intensive, and inapplicable when dealing with large proteome datasets include high-throughput screening and structural docking. Therefore, machine learning (ML) and deep learning (DL) models have become important in methods of making predictions of drugs targetable proteins directly against primary sequences, which offer cost-effective and time- efficient alternatives to experimental validation (Boadu *et al*., 2025).

Other than model innovation, feature representation is critical to the quest to enhance predictive performance. The Average-Block Position Specific Scoring Matrix (AB-PSSM) is the encoding of evolutionary conservation, and the local sequence order is the Dipeptide Composition (DPC) among various properties derived out of sequences (Shoombuatong *et al*., 2023). Their combination results in a hybrid description that combines both global and local features to enable a more complex description of protein characteristics. In this article, the authors introduce CapsDPP, a Capsule-Inspired Deep Learning Framework that combines the properties of AB-PSSM and DPC (Ghulam *et al*., 2022), to accurately categorize the druggable and non-druggable proteins. The suggested approach is very effective in capturing hierarchical associations through capsule routing and shows greater predictive accuracy than the classical CNNs (Rehman *et al*., 2025).

OBJECTIVE

- Devise a prediction system using sequences that has the capacity to distinguish the druggable proteins and insoluble ones that do not require 3-D structural knowledge.
- Design and apply a Capsule-Inspired Neural Network (CapsDPP) which integrates spatial hierarchies in built- in protein attributes.

METHODOLOGY

The proposed CapsDPP in Figure **1** system takes the form of a Capsule-Inspired (Levy *et al*., 2020). Deep Learning Architecture that successfully predicts drugable proteins and non-drugable ones, with just their amino acid sequences. We assembled a benchmark dataset of 2518 proteins. Drug Bank provided 1,218 druggable proteins as well as Swiss-Prot provided 1,300 non-druggable proteins. In order to ensure that the data was not redundant, all sequences were filtered by CD- HIT (Ghulam, Ali *et al*., 2022), whose cutoff was a 50% sequence identity. This produced a quality data that was not homologous. Each protein was assigned a number using two different types of features which were derived by sequences. The initial description, Average Block Position-Specific Scoring Matrix (AB-PSSM, 200D), is based on the profile of the PSSM that is divided into ten equal blocks. It then averages the twenty amino acid scores in every block. The second one, Dipeptide Composition (DPC, 400D) is the frequencies of all 400 possible dipeptides relative to each other to contain local sequence-order data. The 600-dimensional hybrid feature vector combining the features of the global evolutionary and local compositional features was made by using both of the descriptors. Then, two neural networks were constructed, a typical Convolutional Neural Network (CNN) (Ghulam *et al*., 2022), as the control and the suggested Capsule-Inspired Network (CapsDPP), with Primary Capsule and Digit Capsule layers. This is unlike the traditional CNNs that reduce the feature maps to a single dimensional form. The models were trained with Adam optimizer, the learning rate was 0.0005, and the loss was binary cross-entropy, and the batch size was 32. This procedure was repeated five times so as to be accurate. We used the common classification metrics including Accuracy, Precision, Recall, F1-Score and ROC-AUC to fairly and completely compare the CNN and CapsDPP architectures (Jose *et al*., 2025).

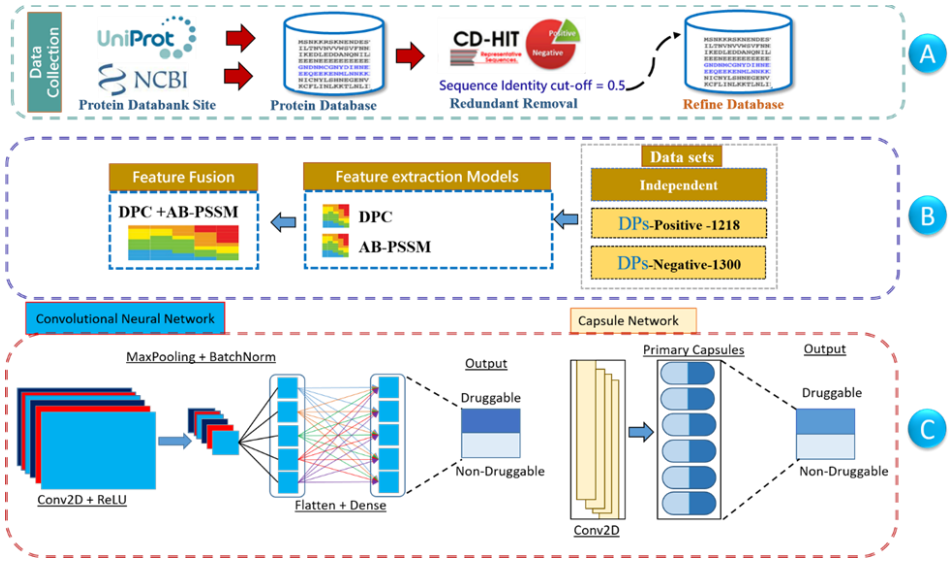

Figure 1. The proposed CapsDPP framework's workflow shows how to obtain data from UniProt/NCBI, get rid of duplicates with CD-HIT, extract hybrid features (AB-PSSM + DPC), and classify using CNN and Capsule-Inspired networks.

RESULT

The analysis of the CapsDPP framework experimental performance indicates that the classification performance is enhanced to a considerable extent relative to the conventional convolutional models. It was evident that the capability of the model to differentiate between entities is enhanced with the inclusion of both characteristics of AB-PSSM and DPC. In the case of the ABC-PSSM features, the CNN was 62 percent accurate and Capsule Network was 80 percent accurate. This indicates that when capsule routing is used; it is possible to easily capture spatial correlations of evolutionary characteristics. However, hybrid 600-dimensional features (AB-PSSM + DPC) significantly improved the performance: the CNN achieved the accuracy of 87% and the suggested CapsDPP achieved the highest possible accuracy of 95%. This 15 percent savings over CNN and 35 percent savings over the traditional 2D CNN on AB-PSSM shows that a hybrid of the capsule design and the complementary descriptors work. This was also evident with training and validation curves converging with little or no overfitting as a result of the existence of the batch normalization and dropout layers. Additionally, it was discovered that the CapsDPP model was more accurate and recalled a higher rate in all folds which validates the stability and improved generalization of the finding of drugs targets as shown in Figure **2**.

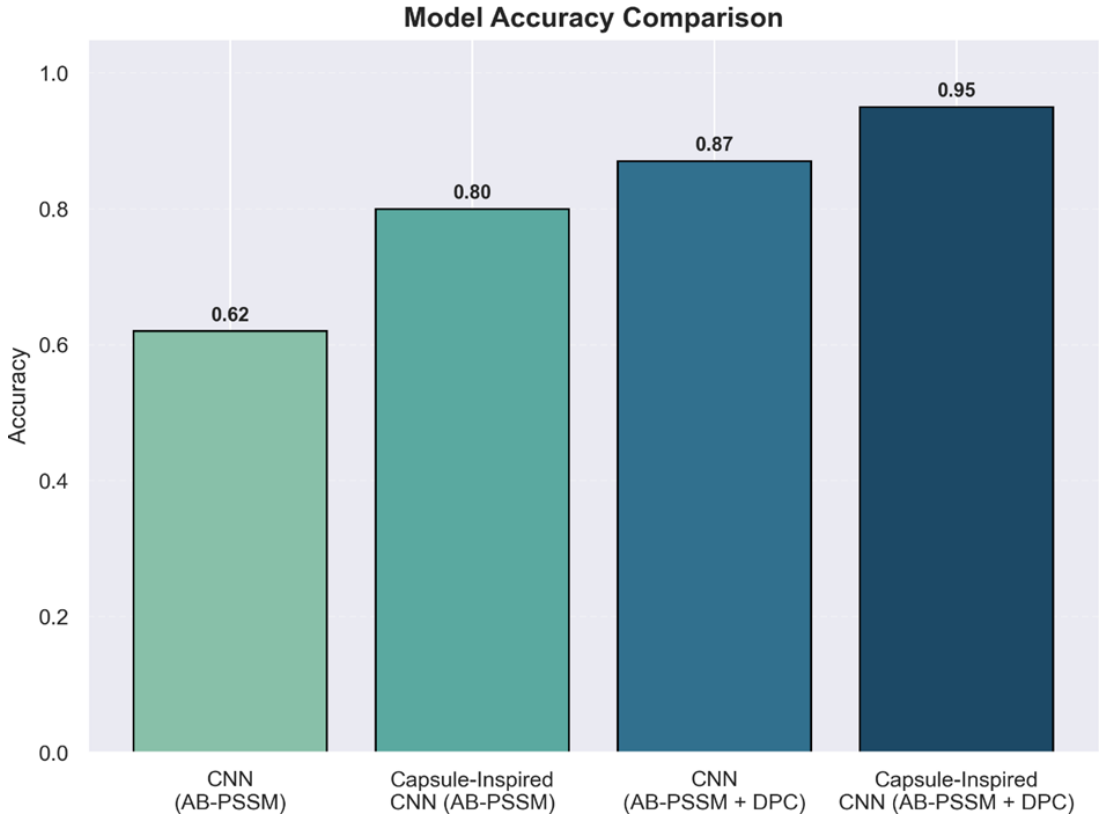

Figure 2. Model accuracy comparison for CNN and Capsule-Inspired CNN trained on AB-PSSM and hybridAB-PSSM + DPC features.

CONCLUSION

It is a study that presents a robust Capsule-Inspired Deep Learning Framework named CapsDPP that is designed to identify druggable proteins using only primary sequences data. The comparison of CNN and the suggested CapsDPP model, conducted in an experiment, indicates that capsule-based models are more accurate in capturing the association of features in a hierarchy. This produces a spectacular 95 percent categorization,

which is higher than those of any other method. The great performance of CapsDPP indicates that it has the ability to predict complicated interactions between protein sequences and this makes it an important tool in the early drug development process. The next steps will involve improving the dataset, adding more physicochemical and structural descriptors and explainable AI methods will be used, such as as SHAP analysis, to explain the biological relevance of the parameters influencing the predication of draggability.

REFERENCES

1. Ghulam, Ali *et al*. "StackAHTPs: An explainable antihypertensive peptides identifier based on heterogeneous features and stacked learning approach." IET Systems Biology 19.1 (2025): e70002.
2. Sikander, Rahu, Ali Ghulam, and Farman Ali. "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set." Scientific reports 12.1 (2022): 5505.
3. Boadu, Frimpong, Ahhyun Lee, and Jianlin Cheng. "Deep learning methods for protein function prediction." Proteomics 25.1-2 (2025): 2300471.
4. Shoombuatong, Watshara, Nalini Schaduangrat, and Jaru Nikom. "Empirical comparison and analysis of machine learning-based approaches for druggable protein identification." EXCLI journal 22 (2023): 915.
5. Ghulam, Ali *et al*. "Human drug-pathway association prediction based on network consistency projection." Computational Biology and Chemistry 97 (2022): 107624.
6. M. Rehman *et al*. "ELM-APDPs: An Explainable Ensemble Learning Method for Accurate Prediction of Druggable Proteins," Computer Modeling in Engineering & Sciences (CMES), 145, (2025): 779–805, 2025.
7. Levy, J. *et al*. "Investigation of capsule-inspired neural network approaches for dna methylation." bioRxiv (2020).
8. Ghulam, Ali *et al*. "DeepImmuno-PSSM: identification of immunoglobulin based on deep learning and PSSM-profiles." VAWKUM Transactions on Computer Sciences 11.1 (2023): 54-66.
9. Ghulam, Ali *et al*. "ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network." Chemometrics and Intelligent Laboratory Systems 226 (2022): 104589.
10. Jose, Jeena, and E. J. Fredrik. "An Advanced GRU-CapsNet Framework for Accurate Chronic Kidney Disease Detection and Prediction." International Journal of Intelligent Engineering & Systems 18.6 (2025).