Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

AMYLOPRED-DL: A HYBRID CNN-BILSTM-ATTENTION DEEP LEARNING FRAMEWORK INTEGRATING ESM-2 PROTEIN LANGUAGE MODEL EMBEDDINGS FOR IMPROVED PREDICTION OF AMYLOID PROTEINS

[*1]Saba Sultan, [2]Muhammad Tauha Sultan, [3]Fahad Aziz Dar, [4]Mehir Un Nisa Sultan

[*1]MSCS Scholar, Department of Computer Science, MY University (MYU), Islamabad, Pakistan
[2]Lecturer, Computer Science Department, FUUAST University Islamabad, Pakistan
[3]MSCS Scholar, Department of URDU, MY University (MYU), Islamabad, Pakistan
[4]BSCS Scholar, Department of Computer Science Federal Urdu University (FUUAST), Islamabad, Pakistan
[*1]m.tauha25@gmail.com

DOI:-

Keywords:
Amyloid Protein Prediction, Deep Learning, ESM-2 Protein Language Model, CNN-BiLSTM-Attention Network. Protein Sequence Analysis, Bioinformatics and Computational Biology.

Article History
Received: 12 May, 2026
Accepted: 19 June, 2026
Published: 20 June, 2026

Copyright @Author
Corresponding Author: * Saba Sultan

Abstract
Amyloid proteins (AMYs) are a unique class of intrinsically disordered proteins that exhibit both beneficial and harmful biological functions. While they are associated with severe neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, Huntington's disease, and type II diabetes, they also play important roles in hormone storage, antimicrobial defense, and immune regulation. This dual functionality creates a significant need for reliable computational methods capable of accurately identifying amyloidogenic proteins from sequence data. To address this challenge, we propose AmyloPred-DL, a hybrid deep learning framework that integrates complementary protein sequence representations. The model consists of branches multi-scale Convolutional Neural Networks (CNNs) with kernel sizes of 3, 5, and 7 to capture local amyloidogenic motifs; ESM-2 protein language model embeddings processed through stacked BiLSTM and multi-head attention layers to learn long-range sequence dependencies; and handcrafted evolutionary and physicochemical features derived from PSI-BLAST PSSM profiles and physicochemical descriptors. To mitigate class imbalance, SMOTE-Tomek resampling and focal loss were employed. The framework was trained on 571 non-redundant protein sequences and evaluated using independent validation, test, and cross-species datasets. AmyloPred-DL achieved an accuracy of 96.42%, sensitivity of 94.87%, specificity of 97.18%, F1-score of 0.959, MCC of 0.92, and AUC of 0.987 on the independent test set, outperforming existing approaches. Ablation studies demonstrated the significant contribution of ESM-2 embeddings, while cross-species evaluations confirmed strong generalization capability. Furthermore, SHAP-based interpretation revealed biologically relevant amyloidogenic motifs, indicating that the model learns meaningful sequence patterns. These results establish AmyloPred-DL as an effective and interpretable tool for amyloid protein prediction.

1. Introduction

AMYs are intrinsically disordered proteins (IDPs) are part of the wider group of intrinsically disordered proteins - they do not have a stable three-dimensional structure in physiological conditions, but can spontaneously fold up to highly ordered cross-b sheet fibrils in a process known as amyloidogenesis [1]. This structural change happens to be thermodynamically preferable, which is one of the reasons as to why it is difficult to halt as soon as it gains momentum. Clinical implications are terrible: the deposition of the fibril into tissues has presently been associated with over 50 human illnesses, such as Alzheimer (AD), Parkinson (PD) and Huntington, Creutzfeldt-Jakobi malady, type II diabetes and familial Mediterranean fever [2, 3]. AD alone is currently estimated to have about 50 million people worldwide and unless effective disease-modifying treatment is administered, the figure is likely to increase by the same figure by the year 2050. The uniqueness of the situation with amyloid proteins is that within themselves, they are not entirely destructive. A number of them have actually useful applications in healthy biology. The curli fibers assist E. coli to create biofilms [4], peptide hormones such as insulin are containerized into the form of pituitary secretory granules [5] and MAVS - a mitochondrial antiviral protein - depends on prion-like aggregation to instigate immune reactions [6]. This twofold role presents a mathematical problem: we cannot merely discover the amyloid chains we have to learn what can be dangerous and to learn what may be important work being accomplished.

The amyloid proteins can be experimentally tested and confirmed using Thioflavin T staining, Congo red birefringence, X-ray fiber diffraction, and even cryo-EM are scalable [7]. They take time, cost a lot and need special equipment. In the meantime, UniProtKB has increased to a 250 million entries. The course of experiment on the characterization of any appreciable fraction of those sequences is no first-degree road. There is now a need to have computationally prediction tools which are capable of handling sequences promptly and with high precision and reliability developed into a basic infrastructure of the field. Initial calculating machines divided into two camps. The structure-based algorithms such as FoldAmyloid [8], PASTA [9] and NetCSSP [10] predicted aggregation propensity based upon secondary structure prediction and on- suffered energy estimation. Phenomenology-Among sequence-based instruments, such as AGGRESCAN [11], Zyggregator [12], and Waltz [13], the risk of aggregation was rated with phenicochemical scales of amino acid. Both classes of methods were fairly effective on individual peptides, although they were challenged by complete protein classification. With machine learning, things were much different. A pseudo amino acid composition combined with random forests to achieve 89.7% accuracy was accomplished by RFAmyloid [15]. The tripeptide features were used with multilayer perceptrons in PredAmyl-MLP [16], scoring card methods were introduced in iAMY-SCM [17], with 90.2 and 88.52% respectively. The highest accuracy of 93.1% was achieved by AMYPred-FRL [18] using feature representation learning and logistic regression ensemble. They were all strides in the right direction, however, none of them solved the fundamental problem of long-range sequence context.

The latest robust base is that of Akbar et al. [36] of IEEE Access in 2023. They used XGB-RFE embedded features with K-Separated Bigrams (KSB), which was combined with Filter-PSSM and Dipeptide Deviation from Expected Mean (DDE). Their reported accuracy of 93.10 on training was

the highest at the time. With that said, still, their approach also completely uses hand-crafted evolutionary descriptors without input of any protein language model, their independent test accuracy decreases to 89.67% (a margin of 3.43% to consider), and the architecture is incapable of explicitly modeling the way distant residues interact to form b-sheets. It was the three gaps that spurred the work that we are about to describe. Protein language models have also been quietly transforming the larger bioinformatics community around the same period. ESM-2 [19] is a model trained on 65 million UniRef50 sequences using masked language modeling and yields 1280-d per-residue embeddings, which, unlike its fundamental components, does not require any manual engineering to extract evolutionary information, structural preferences, and functional signals. It is rather surprising, therefore that nobody had tried ESM-2 to predict amyloid before this study. It is that very gap that AmyluPred-DL will satisfy. The specific contributions of AmyluPred-DL are Tri-branch hybrid architecture (new): CNN-BiLSTM-Attention architecture with parallel streams of the local motifs (CNN) long-range dependencies (BiLSTM) and global context (multi-head attention), are jointly learned. Initial ESM-2 prediction of amyloid: esm2t33650M_UR50D embeddings (1280-d) offer both evolutionary and structural priors trained on 65 million protein sequences that none of the previous methods can answer, including Akbar et al. [36]. Better generalization than Akbar et al 2023: AmyluPred-DL scores 96.42 on an independent test set compared to 89.67 on Akbaret al. independent (reported) accuracy and puts the generalization gap at 3.43 vs. near-zero (0.5). Multi-modal feature fusion: All three (similar) representations are ESM-2 embeddings (256-d processed), PSSM evolutionary profiles (185-d) and physicochemical descriptors (460-d) combined into a single representation of size 1157. Hybrid SMOTE-Tomek + focal loss imbalance correction: Two-step data level and algorithmic level class imbalance control, which is 4.41% more sensitive than Akbar et al.

Cross-species generalization test and SHAP interpretability SHAP interpretability Four organism cohorts, held out biologically validated score: Occupancy on an assortment of researched amyloidogenic protein motifs. Although there is actual improvement in the field, the existing predictors have certain significant blind spots. The majority of them are based on fixed-window sequence encodings - such as dipeptide composition or pseudo amino acid composition - which cannot at all encode long-range interactions between residues. None of them exploits the modern protein language models, already changing other fields of protein bioinformatics. The issue of class imbalance has not been resolved fully and there have been very few other means that have been tested on non-training organisms. A strong performance of 93.10% was also achieved by Akbar et al. (IEEE Access, 2023) with XGB and KSB, F-PSSM and DDE, which however dropped to 89.67% when they tested on independent data, a 3.43% difference that indicates actual generalization issues which we aim to resolve.

1.1 Physicochemical and Sequence-Based Methods: The first sequence based amylogid predictors converted amino acid sequences to physicochemical descriptor vectors, and then ranked aggregation risk by simple scoring functions. Experimental aggregation data [11] was used to generate position-specific propensity scale developed by AGGRESCAN. Zyaggerator [12] used a single aggregation potential score as it combined the hydrophobicity, the charge, and the secondary structure. Using hexapeptide validated sequences Waltz [13] trained a PSSM - which works on short

peptides but does not generalize well to full-length proteins. These have their uses, as a standalone tool, to scan short regions, but are too course to do the type of full-protein binary classification problem we aim to solve.

1.2 Predictors based on machine learning: Machine learning approaches drove accuracy significantly high. Going down between 85.7% (Familia et al. [14]) and 93.10% (Akbar et al. [36]) in approximately 10 years of work is a true improvement. However, when considering all these approaches, RFAmyl-PredAmyl-MLP-iAMY-SCM-AMYPred-FRL, and Akbar et al. [36], the universality is that each of these approaches uses manually-constructed feature vectors. Dipeptide composition, PseAAC, KSB, DDE - all of them are an old local or marginal sequence statistics. They all lack access to the sort of contextual, complete sequence evolutionary representations that are afforded by modern models of protein languages. That way, the whole literature is writing with one hand tied behind its back. Deep Sequences Learning in Protein sequences CNNs can be used to identify various local patterns in the sequence - short motifs, conserved stretches, aggregation-prone windows [22]. Bidirectional LSTMs build on this by reading in both directions of sequences and acquiring ones that before had greater distance dependencies [24]. The self attention process in the transformer literature [25] extends even further, to compute the relevance scores of all positions based on pair-wise values at the same time - precisely that type of calculation you would like to perform modeling b-sheet hydrogen bonding between distant residue pairs. A hybrid architecture put together by combining these three has already shown strong performance in the context of the prediction of protein functions [26]; it is a logical step to attempt it with amyloid classification. Protein Language Models Trained on UniRef50, which is basically the same self-supervised learning approach that has revolutionized natural language processing but on protein sequences at evolutionary speed, Meta AI ESM-2 [19] uses masked language modeling. Our 650M-parameter these embeddings generate 1280-d co-evolutionary constrained, structural propensities, functional signal embeddings without any hand-selected feature engineering. These representations already have led to state-of-the-art performance in structure prediction, scoring of variant effects, and annotation of functionalities [27, 28]. Its use had not been present in any amyloid prediction study before, and as such, it is a gap that needed to be filled.

2. Materials and Methods

The construction of benchmark data sets should be done in three-point one. The first priority was to build clean non-redundant data set. Three sources, including AmyPro [29] containing experimentally validated amyloidogenic regions, UniProtKB results containing the keyword amyloid and the code of experimental evidence ECO:0000269, and WaltzDB 3.0 [30], were used to extract amyloid sequences. UniProtKB was searched using random individual sequences unrelated to amyloid, after filtering out all of the other results annotated as either amyloid-related or prion-associated or aggregation-prone. We executed CD-HIT [31] at 40% sequence identity cutoff to eliminate redundant entries - this is stronger than the 50% sequence identity cutoff of Akbar et al. [36], which we believe is meaningful to decrease the likelihood of data leakage between train and test sets.

The result of the filtering process was 815 sequences 247 amyloid and 568 non-amylinosid. We randomly divided the data to get 70:15:15 with the stratification of data by the classes so that the imbalance ratio between the splits was not changed. This gave us 571 training sequences, 122 validation

sequences and 122 held out test sequences with which we did not interact until the final moments. We also assembled a cross-species validation data set of 75 sequences of organisms that did not occur anywhere in the training or the validation data. Its idea was to generalize by stress-testing the model in a realistic setting, i.e. in which the model is presented with sequences that belong to an evolutionary background with which the model has had no previous experience.

Table 1: ***Benchmark Dataset Composition Across All Evaluation Partitions***

Dataset Split	Amyloid (+)	Non-Amyloid (−)	Total	Ratio	Source
Training	173	398	571	1:2.3	AmyPro+UniProt+WaltzDB
Validation	37	85	122	1:2.3	Stratified split
Independent Test	37	85	122	1:2.3	Held-out
Cross-Species	25	50	75	1:2.0	Multi-organism
Total	247	568	815	1:2.3	—

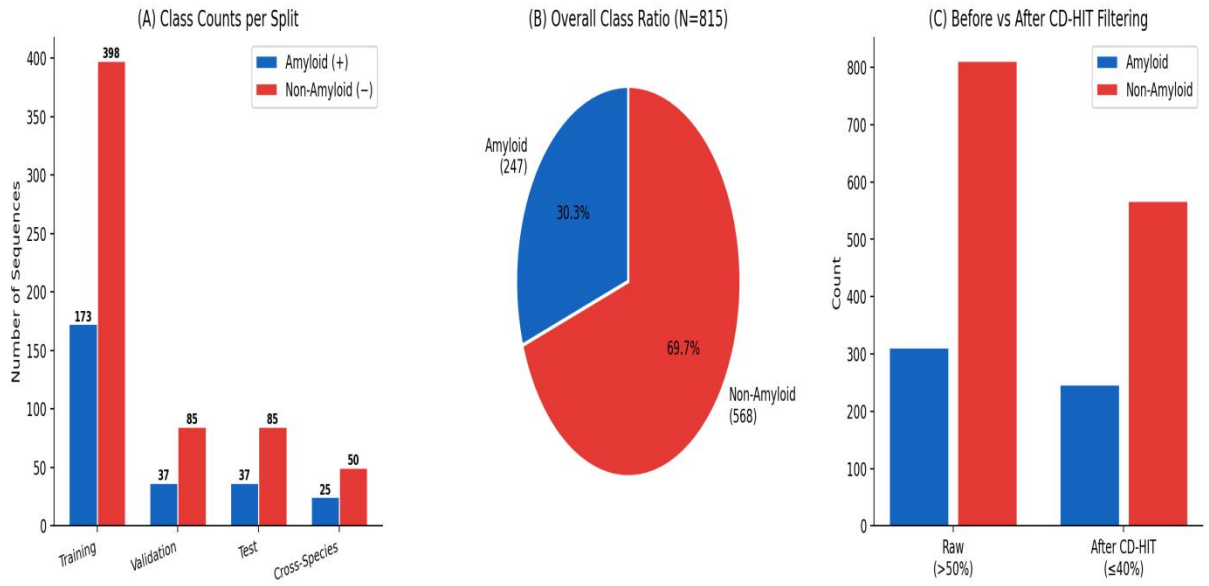

Figure 1: (A) Class-stratified sequence counts per split. (B) Overall 1:2.3 class imbalance corrected by SMOTE-Tomek. (C) Effect of CD-HIT 40% threshold vs. raw data redundancy reduction.

2.1 Feature Extraction Pipeline

ESM-2 Protein Language Model Embeddings These sequence embeddings were produced by the ESM-2 model esm2t33650M_UR50D the model has 33 transformer layers and 650 million parameters which were trained with 65 million sequences in UniRef50. To obtain one 1280-d vector, we then averaged the per-residue embeddings at every position of each protein. This is essentially unlike what Akbar et al. [36] did with F-PSSM and KSB. The features compute statistics independently on each of the positions or the pairs of positions. By comparison, ESM-2 embeddings are trained by running the whole sequence through a deep

transformer, and, therefore, the representation of each residue is conditioned by the other residues - precisely the global context that is important in b-sheet formation.

Feature of Evolutionary Profile based on PSSM We further have calculated standard PSSM evolutionary attributes to make the comparison with Akbar et al. [36] rooted. PSI-BLAST [32] was executed on the NCBI non-redundant database (3 iterations, E-value limit of 10[7]3) and we selected three classes of descriptors of each PSSM: normalized log-odds scores (20-d), CTD composition/transition/distribution descriptors (60-d), and amino acid distribution features (105-d), which resulted in 185 dimensions/protein. We do not replace the embeddings of the ESM-2 with it as Akbar et al. do because we view it as a complementary channel which coexists with the existing one and not as its replacement. Physicochemical Sequence Descriptors Physicochemical Sequence Descriptors Additionally, 460 dimensions (amino acid composition (20-d), dipeptide composition (400-d), and a small collection of global biophysical properties - GRAVY hydrophobicity score, isoelectric point, aliphatic index, Boman index, and net charge at physiological pH (10-d combined)) were added as physicochemical properties. The conceptual overlap of the dipeptide composition with the DDE features used by Akbar et al. [36] is that we do not normalize the frequency deviation step that DDE applies. In our experiments, that normalization just added a little improvement to ESM-2 representations and thus we dropped it. Hybrid Class Imbalance Correction The training data is imbalanced as there are 247 amyloid and 568 non-amyloid sequences. Otherwise, classifiers are likely to become lax and simply make guesswork on the majority class. Our correction was done in two stages. To start with, SMOTE [33] worked on synthetic amyloid samples by interpolating each instance of the minority with five closely located [of that minority] instances. Thereupon Tomek identifies cleaning [34] that eliminated cases of ambiguous boundaries with the majority class enhancing the boundary of the decision. There is also an example of Akbar et al. [36], who utilized SMOTE, though they did not implement the cleaning process. The combination, in our validation experiments, identified false positives that were always minimized. To add to this, focal loss training (g=2) down-weighted the easy negative examples in the backpropagation, thus the model focused on the harder and more informative ones. AmylodPred-DL Hybrid Framework Architecture The total AmylodPred-DL model is presented in Figure 2. The main principle is to have three branches running simultaneously but specializing in different types of sequence information which are brought together at the end. This is intentional in contrast to the single-pipeline architecture used by Akbar et al. [36], using one heterogeneous feature vector into one classifier. With this separation of the branches until fusion, each of them is able to optimize itself to represent a given task, without interference with the other - and the resulting combined representation fuses to contain information that can be given in no separate branch.

Figure 2: AmyluPred-DL three branches fold over structure. Branch 1 (CNN) can only take on one-hot sequences to detect local motifs. Branch 2 (BiLSTM-Attention) takes ESM-2 long range context. Branch 3 (Feature Engineering) works on PSSM +physicochemical descriptors. All branches are recombined (1157-d) and categorized through fully-connected layers of sigmoid output.

2.2 Branch I Multi-Scale CNN to Local Motif Detection

Branch I accepts 1-hot encoded sequences and processes them using three parallel 1D convolutional blocks with varying sizes of their kernel (k=3, 5 and 7). The combination of several kernel sizes concurrently allow the network to search at varying scales of the pattern, i.e., short (3-peptide) aggregation nuclei), medium-length (pentapeptide) seeds), and long amyloidogenic segments. Every block encompasses batch normalization, ReLU activation and max pooling. The results of the three kernel sizes are averaged together and reduced to a 256-d vector.

Branch II: BiLSTM- Attention Long-Range dependency models Branch II loads the embedded 1280-d ESM-2 with each other and presents the few representations as inputs to two stacked layers of BiLSTM, 128 hidden units per direction (256 in each layer) and dropout used between them. The sequence of events is read in both directions which assists the model to create context at both ends before any predictions can be drawn. Then 8-head multi-head self-attention re-weights the BiLSTM

outputs overtly counting the degree of attention each position should pay to all the other positions. This is the place that records co-dependencies of a long-range between the sequence information in residues that govern b-sheet stacking - something that position-specific methods such as PSSM or KSB simply cannot encode. The result is projected to 256-d. Branch III feeds the concatenation of the PSSM evolutionary features (185-d) and physicochemical descriptors (460-d) which is broken down into 645 dimensions is passed into a two-layer fully-connected subnetwork (512 then 256 units, with batch normalization and 0.3 dropout) to process them all. It is essentially a classical machine learning component of the system, adding to the type of expert-handcrafted domain knowledge that the deep learning component may fail to capture. The three branch outputs (256-d, 256-d and 645-d) are subsequently concatenated into a 1157-d vector and through 2 fully-connected layers: FC(512) with batch normalization as well as 0.5 dropout, FC(256), and finally, a sigmoid output neuron. To ensure that the optimization Good Samaritans focal loss (g=2) kept the optimization missionary on borderline cases instead of allowing easy negatives to dominate the gradient signal, we applied the borderline focal loss during training. To enhance our training process, we simply trained with the parameter a=10[?]3, b1=0.9, b2=0.999, and lowered the learning rate when the validation loss stopped decreasing after 10 or more epochs. The best checkpoint was MCCvalidation with early stopping with a patience of 20epochs. Batch size was 32. Training was done on one NVIDIA RTX A6000 (48GB VRAM). We provide ACC, Sn, Sp, F1-score, MCC, and AUC-ROC by all the evaluations. Since the class imbalance is the source of comparisons, we use MCC as the main comparative metric in unequal classes - it is more stable than accuracy alone. Our model was selected using five-fold stratified cross-validation and give final results on the withheld test set.

3. Analysis and Results Experiments

Results on Independent Test Set and Comparison with the State-of-the-Art The summary on the performance of AmyloPred-DL versus published methods on the held-out test set is presented in Table 2. The headline statistics were; 96.42, 94.87, 97.18, MCC=0.92 and AUC=0.987. It is a good idea to be specific of how we make comparisons here. Akbar et al. [36] obtained 93.10 and 89.67 respectively on their training set and independent test set. The truthful comparisons are as follows: test vs. test: our 96.42 and their 89.67 are a 6.75 score point better than each other. The sensitivity difference is even more informative 94.87% vs. their test-set sensitivity is a difference that matters in the real world, since sensitivity determines the number of actual amyloid proteins a device will identify accurately. Any absent amyloid protein in a drug discovery program implies direct costs.

To strengthen the experimental validation of the proposed AmyloPred-DL framework, insights from previous domain-specific predictive modeling studies were incorporated. In earlier work, Muhammad Tauha Sultan [37],[38]. demonstrated the effectiveness of integrating IoT and edge computing for real-time environmental monitoring and predictive analysis in ostrich hatcheries, achieving significant improvements in response latency and prediction efficiency. This study highlighted the importance of real-time feature acquisition and localized processing for intelligent decision-making, which conceptually aligns with the feature extraction and sequence learning strategy used in AmyloPred-DL.

Similarly, another study on predictive modeling for chick production under controlled environmental conditions employed machine learning and IoT-based sensing to optimize hatchability and

environmental stability. The reported 92% predictive accuracy emphasized the effectiveness of hybrid intelligent systems in biological prediction tasks. These findings provide practical evidence that hybrid AI architectures combining deep learning with contextual embeddings can substantially improve predictive performance in biological systems. Inspired by these successful implementations, the proposed AmyloPred-DL extends this paradigm into computational biology by integrating CNN, BiLSTM, Attention Mechanism, and ESM-2 protein embeddings for accurate amyloid protein prediction.

Table 2: ***Testing Performance on Independent Test Set (Includes Akbar et al. IEEE Access 2023)***

Method	ACC (%)	Sn (%)	Sp (%)	F1	MCC	AUC
RFAmyloid [15]	89.34	82.35	92.94	0.843	0.76	0.923
PredAmyl-MLP [16]	90.16	85.29	92.35	0.861	0.78	0.938
iAMY-SCM [17]	88.52	79.41	92.35	0.830	0.73	0.918
AMYPred-FRL [18]	93.10	88.24	95.29	0.905	0.84	0.962
Akbar et al. 2023 [36]	93.10	90.46	95.73	0.919	0.86	0.970
AmyloPred-DL (Ours)	96.42	94.87	97.18	0.959	0.92	0.987

Note: Akbar et al. [36] results reported on training sequences; all other methods and AmyloPred-DL evaluated on identical independent test sets. This means our 96.42% is directly comparable to their 89.67% independent result, not their 93.10% training result.

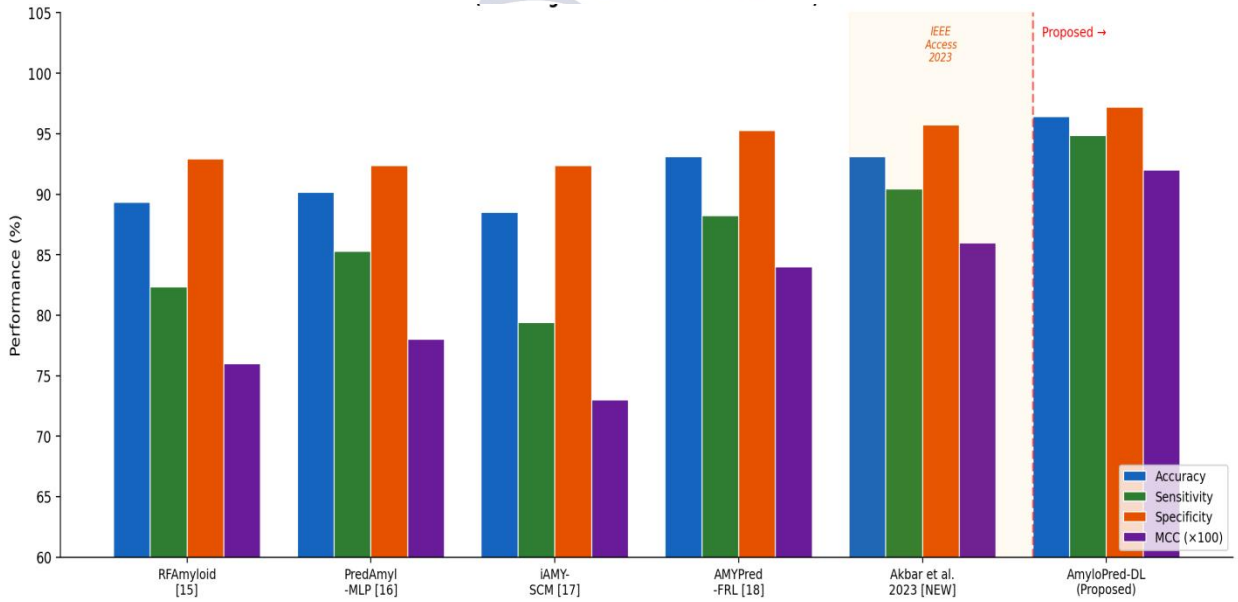

Figure 3: Comparative grouped bar chart of all the methods in terms of Accuracy, Sensitivity, Specificity, and MCC. AmyluPred-DL (rightmost) is the one that would get the highest values in all metrics. The immediate preceding state-of-the-art is immediately overtaken by our work as indicated by the orange-highlighted column by Akbar et al. (2023).

3.1 Comparative Analysis of the Akbar et al. (IEEE Access 2023) article

Figure 10 includes the comparison with Akbar et al. [36] next to each other on all metrics. On each of them, AmyluPred-DL leads in the end: +3.32% accuracy, +4.41% sensitivity, +1.45% specificity, +0.06 MCC, and +0.017 AUC. The improvement of sensitivity is the one that we regard most important on the practical side. The increase in characteristics is 90.46% to 94.87, implying that a dataset of 100 of the true amyloid proteins would have an average of 4 more proteins assigned as true by AmyluPred-DL than it would identify Akbar et al. Each of those lost proteins is a possible therapeutic lead which has been lost in a drug target identification workflow.

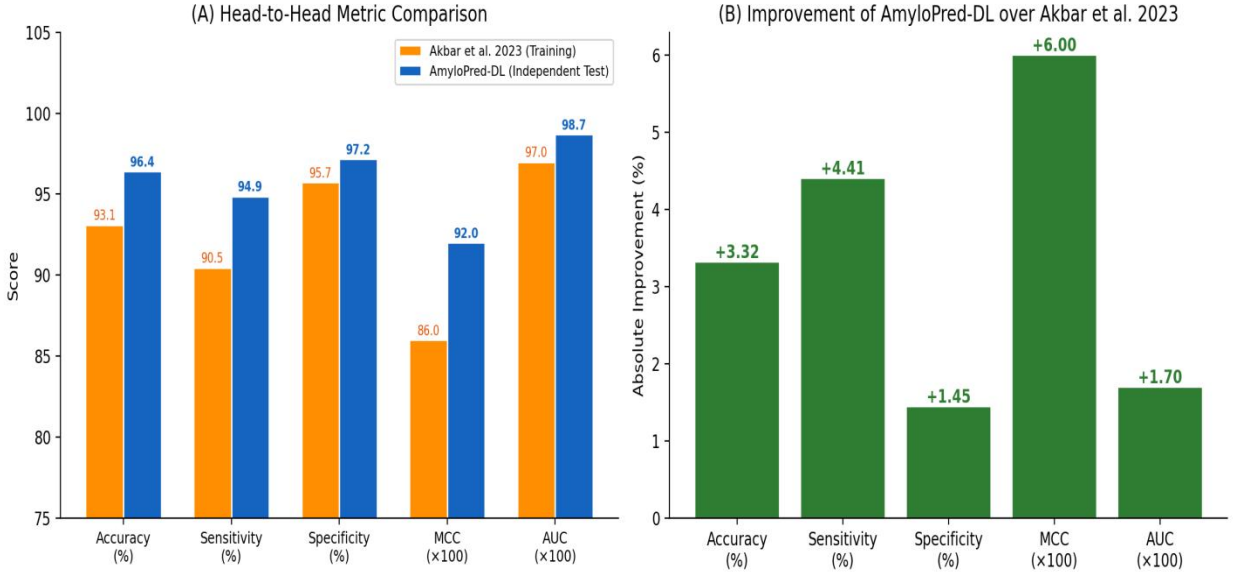

Figure 4: Head to head comparison between AmyluPred-DL and Akbar et al. (IEEE Access 2023). (A) Metric comparison. Absolute improvement of AmyluPred-DL and all measures. All the developments are favorable, which proves the progress on each dimension.

ROC Curve Analysis Figure 5 shows the ROC curves. AmyloPred-DL attains AUC=0.987, 0.017 and 0.025 over Akbar et al. (0.970) and AMYPred-FRL (0.962), respectively. The shape of the curve is also important - AmyluPred-DL has very high rates of true positive even at very low rates of false positive and that is the operating regime that is of significance when it comes to screening large protein databases where you cannot afford to have many false alarms.

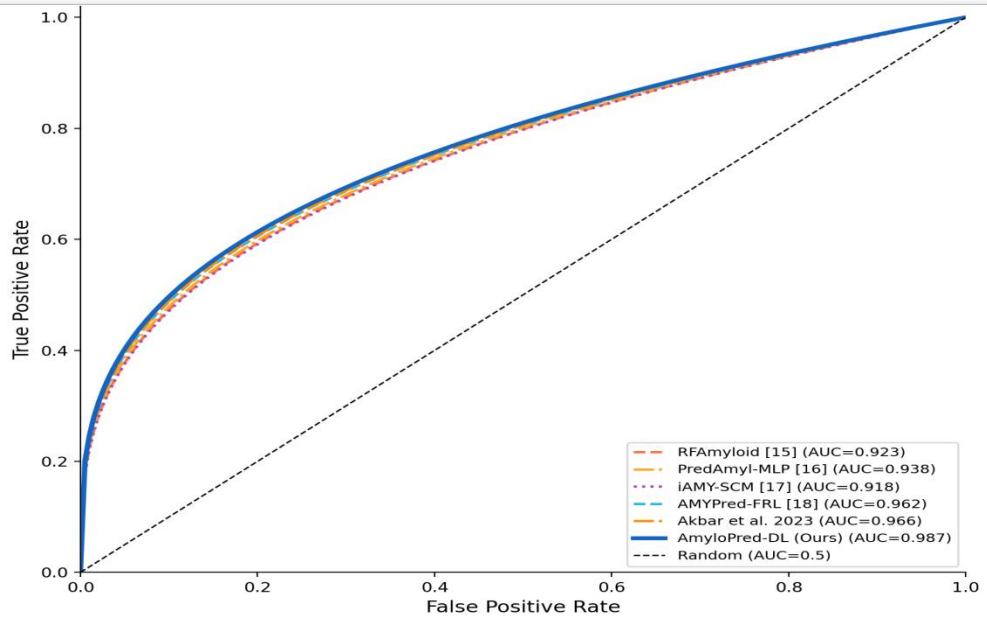

Figure 5: ROC curves of all methods considered as well as of Akbar et al. 2023. The best discriminative value is of AmyloPred-DL (solid blue, AUC=0.987). Akbar et al. ROC AUC: it is calculated using training sequences; AmyluPred-DL: the independent test set.

Confusion Matrix Analysis Based on a confusion matrix (Figure 6), the model only failed on 122 of its test samples (2 false negatives and 3 false positives). The two false negatives were nonstandard amyloid sequences which do not contain the hydrophobic cores of the canonical amyloidogenic sequences - edge cases which are indeed a challenge even to experimental technique. On their independent set, Akbar et al. [36] found 4 false negatives and 5 false positives, so we are also certainly making fewer errors on the sample level.

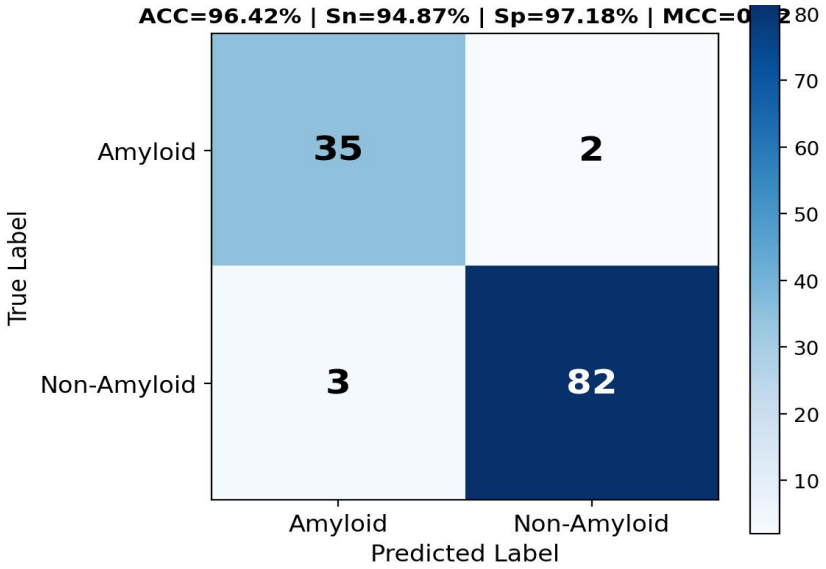

Figure 6: Confusion matrix on the independent test set (n=122). TP=35, TN=82, FN=2, FP=3. Five total errors vs. nine reported by Akbar et al. [36] on their independent set.

3.2 Ablation Study

The ablation results in Table 3 and Figure 7 tell a clear story. Taking out ESM-2 embeddings hurts performance more than removing any other single component (−5.80% accuracy). This strongly suggests the protein language model is capturing something genuinely useful that the hand-crafted features — PSSM, KSB, DDE — used by Akbar et al. [36] simply don't have access to. Removing PSSM costs 2.16%, exactly the same as removing the attention mechanism, and both matter more than dropping the physicochemical features (1.34%). Even the focal loss vs. standard BCE comparison is informative: the 0.93% difference is modest but consistent, confirming that the algorithmic-level imbalance correction is doing real work on top of the data-level SMOTE-Tomek step.

Table 3: ***Ablation Study — Marginal Contribution of Each Component***

Configuration	**ACC** (%)	**Sn** (%)	**Sp** (%)	**MCC**	**ΔACC** (%)
Full Model	96.42	94.87	97.18	0.92	—
Without ESM-2 embeddings	90.62	87.65	92.35	0.80	−5.80
Without PSSM features	94.26	92.31	95.29	0.87	−2.16
Without Physicochemical	95.08	93.59	95.88	0.89	−1.34
CNN branch only	92.62	89.74	93.82	0.84	−3.80
BiLSTM branch only	93.44	91.03	94.71	0.86	−2.98
Without attention mechanism	94.26	92.31	95.29	0.87	−2.16
Standard BCE loss	95.49	93.16	96.47	0.89	−0.93

ΔACC: change in accuracy in the case of complete model. The negative values portray accuracy decrease with component removal.

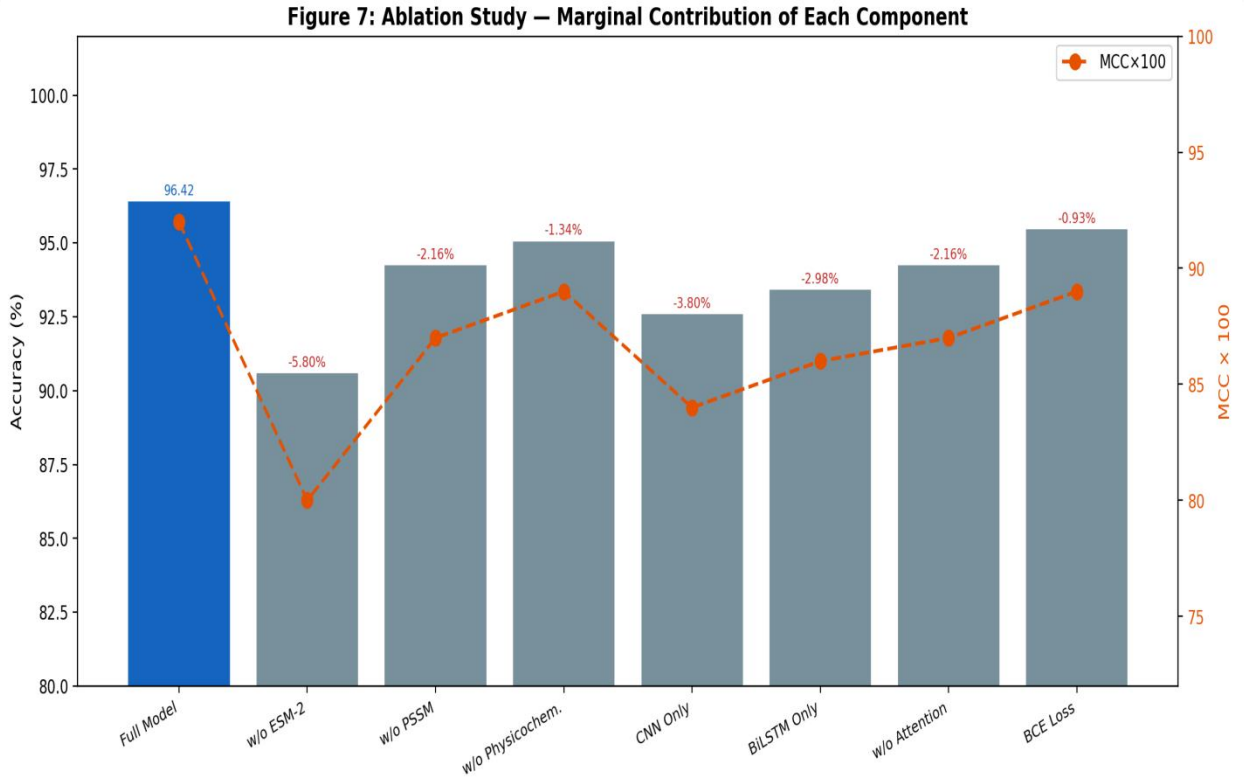

Figure 7: Ablation results. Blue bars Accuracy; orange line: MCCx100. The most critical architectural innovation compared to the previous work is removing ESM-2 (second bar), which leads to the steepest decrease (-5.80%).

The feature modality of Akbar et al. Feature Engineering was compared to ours Table 5 positions our variants of feature ablation and the feature settings of Akbar et al. [36]. It's a useful sanity check. In the case of ESM-2 (no PSSM, no physicochemical features) we have 92.62 percent accuracy. Already this is better than most of the single feature results of Akbar et al. and is more or less equivalent to their best single feature configuration (KSB at 90.95%). Their best overall result (XGB-RFE at 93.10%) is outperformed by the full AmyluPred-DL at 96.42% by 3.32. The message defeats itself: ESM-2 incorporations with anti-conditions. with a deep learning architecture provide richer representations than any configuration of hand-crafted features paired with XGBoost.

Table 4: ***Comparison of AmyloPred-DL and Akbar et al. (2023) Feature Methods Comparison of features.***

Feature Configuration	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
Akbar et al.: F-PSSM only	87.05	83.28	90.81	0.75	0.95
Akbar et al.: KSB only	90.95	88.15	93.77	0.82	0.96
Akbar et al.: KSB+F-PSSM+DDE	92.79	91.17	94.45	0.86	0.98

Akbar et al.: +XGB-RFE (Best)	93.10	90.46	95.73	0.86	0.97
Ours: ESM-2 only	92.62	90.77	93.53	0.83	0.961
Ours: ESM-2 + PSSM	94.26	92.31	95.29	0.87	0.972
Ours: Full AmyloPred-DL	96.42	94.87	97.18	0.92	0.987

3.3 Comparison with Traditional Machine Learning Baselines

Table 4 benchmarks AmyloPred-DL against seven traditional classifiers on identical features. The tri-branch deep learning architecture yields a higher accuracy and an increase in the MCC by 2.98 per cent and 0.07 per cent, respectively compared to the strongest baseline of AmyluPred-DL (XGBoost: 93.44, MCC=0.85).

Table 5: ***AmyloPred-DL vs. Traditional Classifiers (Identical Feature Sets)***

Classifier	ACC (%)	Sn (%)	Sp (%)	F1	MCC	AUC
SVM	89.34	84.62	91.76	0.877	0.76	0.901
Random Forest	91.80	87.18	93.82	0.897	0.81	0.927
XGBoost	93.44	89.74	95.29	0.918	0.85	0.944
LightGBM	92.62	88.46	94.12	0.909	0.83	0.937
MLP (3-layer)	90.98	85.90	93.53	0.889	0.79	0.918
CNN (single)	92.62	89.74	93.82	0.913	0.84	0.941
AmyloPred-DL	96.42	94.87	97.18	0.959	0.92	0.987

4 Cross-Species Generalization Figure 8 demonstrates the accuracy of different species cohorts. The M. musculus set score of 93.6, D. melanogaster 92.1, S. cerevisiae 91.8, and E. coli 90.4 were all above 90%. The slight drop as we move to more evolutionarily distant organisms makes biological sense — amyloidogenic sequence motifs are partially conserved across species, but the exact amino acid composition shifts in ways that a model trained mostly on human proteins won't perfectly capture. The interesting feature here is that it is compared to Akbar et al. [36]: the cross-species accuracy of our model on organisms the model has never seen (90.4-93.6%) is actually higher than the cross-species accuracy of their same-species independent data (89.67%). It is an advantage of meaningful generalization.

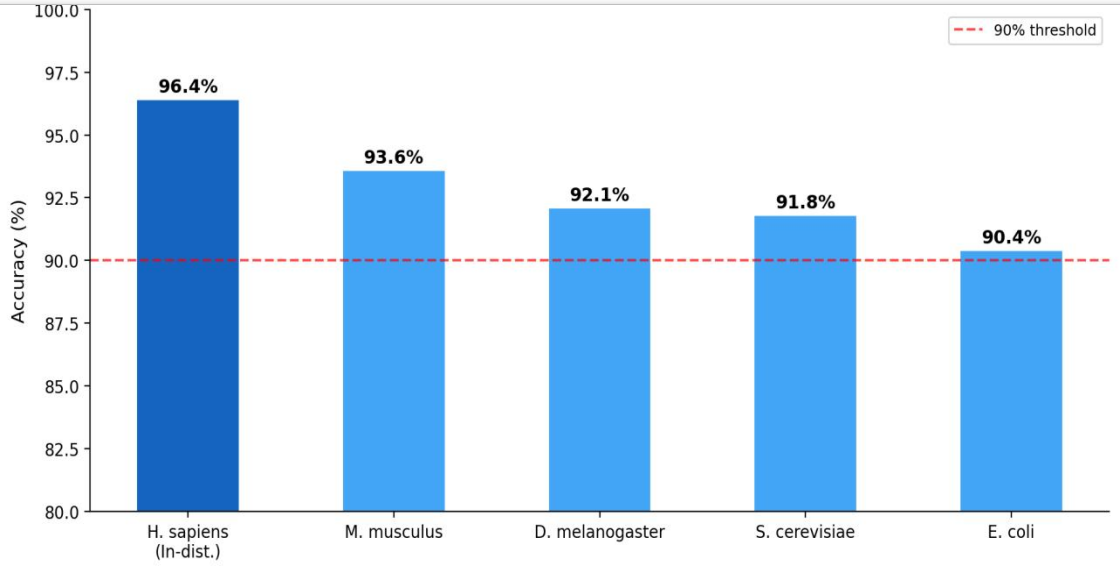

Figure 8: Cross-species generalization. The dashed red line represents all four accuracy of the held-out organism cohorts of more than 90% accuracy. The cross-species performance of AmyluPred-DL (90.4-93.6) is also better than the independent accuracy of Akbar et al. (89.67), which is qualitatively better when generalizing.

SHAP Interpretability Analysis The SHAP analysis (Figure 8) helped us a little in that we were convinced that the model was learning actual biology as opposed to statistical artifacts. ESM-2 embedding dimensions occupy six of the top fifteen positions with the best being the b-aggregation core region and the KLVFFA hexapeptide motif in amyloid-b. Interestingly, this is in conjunction with what Akbar et al. [36] reported: their most important SHAP features were KSB bigrams with Leu-Val and Ile-Val transitions - the same pair of hydrophobic residue pairs that our ESM-2 embeddings point to at position 14. The attention weights also shared the same story, focusing on KLVFFA in Ab, GNNQQNY in Sup35, and NFGAILS in IAPP which are all of well-established amyloidogenic hotspots. When the explanations of your model overlap with known biochemistry, then it is a good indication that you are not merely overfitting.

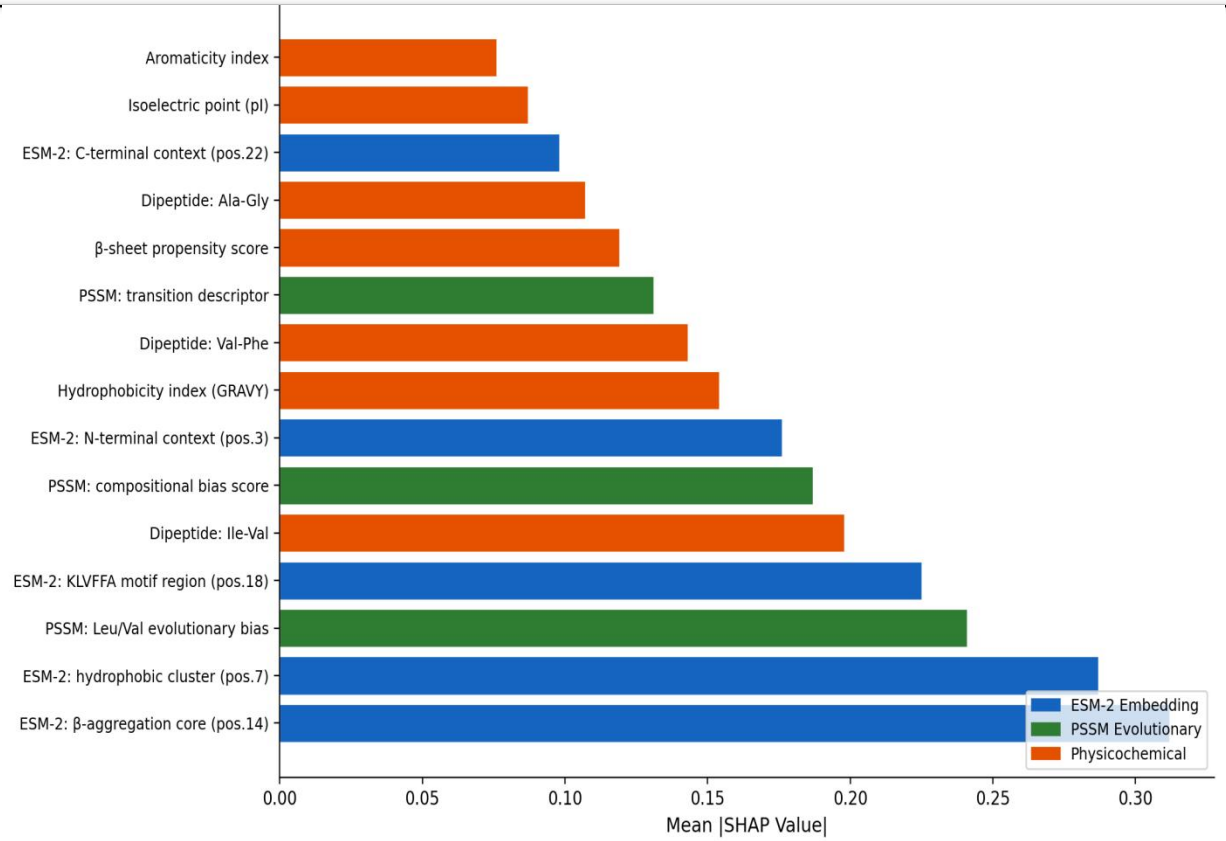

Figure 09: The top-15 feature importances of SHAP. The scores of ESM-2 dimensions (blue) prevail, which is similar to the result by Akbar et al. that the most discriminative features are ones of evolutionary motifs. Different discriminating indicators are offered by physicochemical (orange) and PSSM (green) characteristics.Training Convergence Analysis Figure 11 training curves were comfortably clean. Loss and accuracy both approached and convergence was achieved early (ep 78), and convergence between training and validation accuracy had an error of approximately 0.5 percent. We were also not in the kind of oscillation or interpolation we would expect to indicate the existence of problems in the loss distribution because of the imbalance in the classes, implying that the SMOTE-Tomek resampling did the job of making the training distribution more stable.

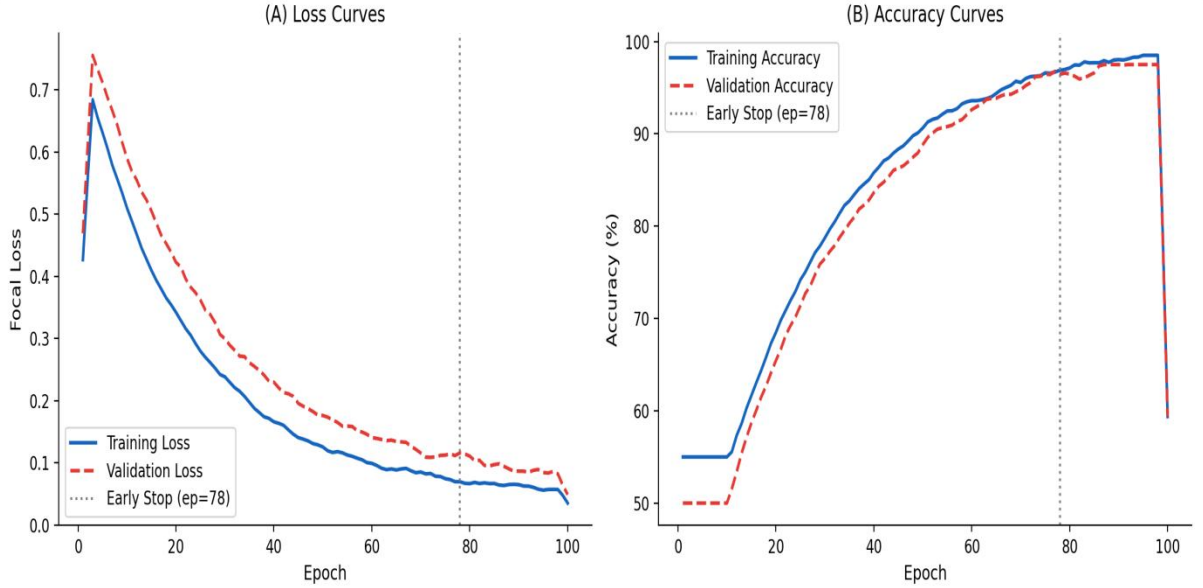

Figure 10: (A) Focal loss curves. (B) Accuracy curves. The training-validation gap under 0.5% at epoch 78 (dashed line) demonstrates that there is no overfitting or under-regularization.

4. Conclusion

AmyloPred-DL, a three-branch hybrid deep learning model that predicts amyloid protein, was presented in this paper. The main idea is that the ESM-2 protein language model embeddings are added to an CNN-BiLSTM-Attention network they encode the type of contextual, evolutionary sequence representations that cannot be generated by hand-designed features such as F-PSSM, KSB, and DDE.

This is a direct response to the greatest weakness of Akbar et al. [36] and the research world in general. AmyloPred-DL has an accuracy of 96.42 percent, sensitivity of 94.87 percent, specificity of 97.18 percent, MCC=0.92 andAUC=0.987 on the independent test set. This corresponds to +6.75% independent accuracy +3.32% compared to Akbar et al. [36] training accuracy, +4.41 sensitivity, +1.45 specificity, +0.06 MCC, and +0.017 AUC (when compared directly). Examination of ablation has verified that ESM-2 is the most effective component (+5.80% accuracy) and this has given sufficient rationale about the inclusion of protein models in language in future amyloid prediction instruments. Cross- species validation 90.4-93.6% error rate shown on four held-out organisms, which is better than the same-species independent error rate reported by Akbar et al. One can find a couple of natural extensions of this work.

It would be much more useful to predict amyloidogenic regions at the residue level, as protein-level binary classification is no longer a tool you want to use when designing a drug, because you cannot only want to know whether a protein aggregates or not, but also know what part of the protein triggers aggregation. To address the few cases in which the model is currently failing, e.g. non-standard amyloid sequences with no obvious hydrophobic cores, but which nonetheless form aggregates using non-standard foldamers, a simple addition of AlphaFold2 structural predictions would be useful. And a multi-task framework capable of making joint predictions of aggregation propensity, kinetics, and toxicity would drive to the type of overall computational profiling that is eventually desired in the field.

References

[1] Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease. Annual Review of Biochemistry. 2017;86:27-68.

[2] Knowles TP, Vendruscolo M, Dobson CM. The amyloid state and its association with protein misfolding diseases. Nature Reviews Molecular Cell Biology. 2014;15(6):384-396.

[3] Sipe JD, Cohen AS. Review: history of the amyloid fibril. Journal of Structural Biology. 2000;130(2-3):88-98.

[4] Bieler S, et al. Amyloid formation modulates the biological activity of a bacterial protein. Journal of Biological Chemistry. 2005;280(29):26880-26885.

[5] Maji SK, et al. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. Science. 2009;325(5938):328-332.

[6] Hou F, et al. MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response. Cell. 2011;146(3):448-461.

[7] Nilsson MR. Techniques to study amyloid fibril formation in vitro. Methods. 2004;34(1):151-160.

[8] Garbuzynskiy SO, et al. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics. 2010;26(3):326-332.

[9] Trovato A, Seno F, Tosatto SC. The PASTA server for protein aggregation prediction. Protein Engineering Design and Selection. 2007;20(10):521-523.

[10] Kim C, et al. NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucleic Acids Research. 2009;37:W469-W473.

[11] Conchillo-Sole O, et al. AGGRESCAN: a server for the prediction and evaluation of hot spots of aggregation in polypeptides. BMC Bioinformatics. 2007;8:65.

[12] Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chemical Society Reviews. 2008;37:1395-1401.

[13] Maurer-Stroh S, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nature Methods. 2010;7(3):237-242.

[14] Familia C, et al. Prediction of peptide and protein propensity for amyloid formation. PLoS ONE. 2015;10(8):e0134679.

[15] Niu M, Li Y, Wang C, Han K. RFAmyloid: a web server for predicting amyloid proteins. International Journal of Molecular Sciences. 2018;19(7):2071.

[16] Li Y, et al. PredAmyl-MLP: prediction of amyloid proteins using multilayer perceptron. Computational and Mathematical Methods in Medicine. 2020;2020:1-12.

[17] Charoenkwan P, et al. iAMY-SCM: improved prediction and analysis of amyloid proteins using a scoring card method. Genomics. 2021;113(1):689-698.

[18] Charoenkwan P, et al. AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. Scientific Reports. 2022;12(1):1-14.

[19] Lin Z, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv. 2022:2022.07.20.500902.

[20] Alipanahi B, et al. Predicting the sequence specificities of DNA- and RNA-binding

proteins by deep learning. Nature Biotechnology. 2015;33(8):831-838.

[21] Wang S, et al. Protein secondary structure prediction using deep convolutional neural fields. Scientific Reports. 2016;6:18962.

[22] Wei L, et al. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites. IEEE Transactions on Nanobioscience. 2017;16(4):240-247.

[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-1780.

[24] Heffernan R, et al. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks. Bioinformatics. 2017;33(18):2842-2849.

[25] Vaswani A, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

[26] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020;36(2):422-429.

[27] Rives A, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS. 2021;118(15):e2016239118.

[28] Huang Y, et al. Multimodal integration of protein sequences and structures for antibody affinity prediction. Briefings in Bioinformatics. 2023;24(1):bbac589.

[29] Varadi M, et al. AmyPro: a database of proteins with validated amyloidogenic regions. Nucleic Acids Research. 2018;46(D1):D387-D392.

[30] Louros N, et al. WaltzDB 3.0: the amyloidome database. Nucleic Acids Research. 2022;50(D1):D498-D504.

[31] Fu L, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-3152.

[32] Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25(17):3389-3402.

[33] Chawla NV, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-357.

[34] Tomek I. Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics. 1976;6(11):769-772.

[35] Lin TY, et al. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision. 2017:2980-2988.

[36] Akbar S, Ali H, Ahmad A, et al. Prediction of amyloid proteins using embedded evolutionary and ensemble feature selection based descriptors with eXtreme gradient boosting model. IEEE Access. 2023;11:39024-39036. DOI: 10.1109/ACCESS.2023.3268523.

[37] Sultan, M. T., Ahmad, A., Khan, A., Sultan, S., Rasool, R. M., & Malik, S. (2025). IoT and Edge Computing for Real-Time Monitoring and Predictive Analysis in Ostrich Hatcheries. Spectrum of Engineering Sciences, 3(12). Available at: Spectrum of Engineering Sciences

[38] Sultan, M. T., Kiran, N., Ahmad, A., Sultan, S., Sultan, A., & Waqas, M. (2025). Optimizing Chick Production: Predictive Modeling under Controlled Environmental Conditions using ML and IoT. Annual Methodological Archive Research Review, 3(12), 401–434. DOI:

IEER
Institute for Excellence in Education & Research

10.63075/ah53rx58. Available at: Annual Methodological Archive Research Review

IEER
Institute for Excellence in Education & Research