


# AggNet: Advancing protein aggregation analysis through deep learning and protein language model

Wenjia He<sup>1,2,3</sup>  | Xiaopeng Xu<sup>1,2,3</sup> | Haoyang Li<sup>1,2,3</sup> | Juexiao Zhou<sup>1,2,3</sup> | Xin Gao<sup>1,2,3</sup>

<sup>1</sup>Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

<sup>2</sup>Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

<sup>3</sup>Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

## Correspondence

Xin Gao, Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.  
Email: [xin.gao@kaust.edu.sa](mailto:xin.gao@kaust.edu.sa)

## Funding information

King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA), Grant/Award Numbers: REI/1/5992-01-01, REI/1/5404-01-01, URF/1/4663-01-01, REI/1/5234-01-01, REI/1/5289-01-01, REI/1/5414-01-01; King Abdullah University of Science and Technology (KAUST) Center of Excellence for Smart Health (KCSH), Grant/Award Number: 5932; King Abdullah University of Science and Technology (KAUST) Center of Excellence on Generative AI, Grant/Award Number: 5940

**Review Editor:** Nir Ben-Tal

## Abstract

Protein aggregation is critical to various biological and pathological processes. Besides, it is also an important property in biotherapeutic development. However, experimental methods to profile protein aggregation are costly and labor-intensive, driving the need for more efficient computational alternatives. In this study, we introduce “AggNet,” a novel deep learning framework based on the protein language model ESM2 and AlphaFold2, which utilizes physicochemical, evolutionary, and structural information to discriminate amyloid and non-amyloid peptides and identify aggregation-prone regions (APRs) in diverse proteins. Benchmark comparisons show that AggNet outperforms existing methods and achieves state-of-the-art performance on protein aggregation prediction. Also, the predictive ability of AggNet is stable across proteins with different secondary structures. Feature analysis and visualizations prove that the model effectively captures peptides' physicochemical properties effectively, thereby offering enhanced interpretability. Further validation through a case study on MEDI1912 confirms AggNet's practical utility in analyzing protein aggregation and guiding mutation for aggregation mitigation. This study enhances computational tools for predicting protein aggregation and highlights the potential of AggNet in protein engineering. Finally, to improve the accessibility of AggNet, the source code can be accessed at: <https://github.com/Hill-Wenka/AggNet>.

## KEYWORDS

amyloid, APR, computational biology, machine learning, protein aggregation, protein engineering

## 1 | INTRODUCTION

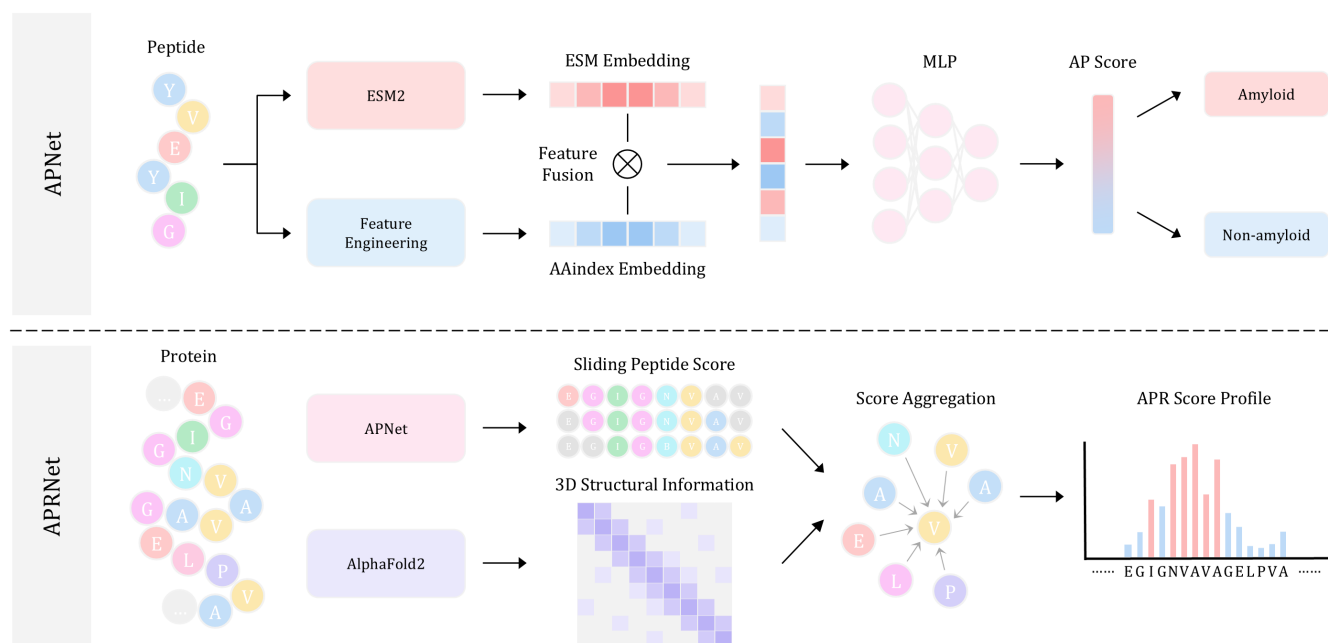
Protein aggregation is a complex biochemical process that often results in the formation of highly ordered amyloid fibrils or amorphous aggregates. This phenomenon is associated with various human diseases, poses challenges in therapeutic protein development, and inspires the creation of novel bio-inspired materials. Neurodegenerative diseases such as Alzheimer's disease (AD), Parkinson's disease (PD), and Huntington's disease (HD) are widely believed to be caused by the misfolding and aggregation of specific peptides or proteins (Koo et al., 1999; Ross & Poirier, 2004). Similarly, type II diabetes is linked to the aggregation of islet amyloid polypeptide (IAPP) (Kahn et al., 1999). With the rapid advancement of protein-based pharmaceuticals like monoclonal antibodies and peptide drugs, protein aggregation has emerged as a significant hurdle in their production, storage, and purification (Lowe et al., 2011; Perchiacca & Tessier, 2012). High concentrations are often required for these biotherapeutics, but increased interactions between protein molecules can lead to severe aggregation, reducing therapeutic efficacy and potentially causing side effects (Lundahl et al., 2021; Rahban et al., 2023). Protein aggregation also inspires the development of self-assembling nanomaterials. Peptides with specifically designed sequences can self-assemble into nanofibers under certain pH conditions and temperatures (Wang et al., 2008). These nanofibers serve as cell scaffolds to promote cell growth and tissue regeneration or function as drug delivery systems to control drug release (Matson & Stupp, 2012). Such applications have spurred considerable interest in understanding and controlling protein aggregation.

To effectively characterize the aggregation and stability of proteins, many biosensors are proposed to monitor and engineer proteins. For instance, Ebo et al. (2020) developed a tripartite  $\beta$ -lactamase enzyme assay (TPBLA) that correlates protein aggregation propensity with bacterial susceptibility of beta-lactam antibiotics. Similarly, Ren et al. (2021) linked the stability of the protein of interest (POI) with the activity of the bacterial enzyme CysGA, which catalyzes the formation of endogenous fluorescent compounds. Nonetheless, these *in vivo* methods, along with traditional approaches like x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (Housmans et al., 2023), are notably resource-intensive and time-consuming. In contrast, *in silico* methods offer a rapid, cost-effective alternative for identifying aggregation-prone regions (APRs), which are essential for understanding aggregation mechanisms, developing mitigation strategies, and even suggesting

new therapeutic strategies for cancer. A recent innovation by Janssen et al. (2023) involved designing amyloid peptides (Pept-ins™) that target the APRs of KRAS to induce its misfolding and aggregation, offering a novel approach to address previously undruggable targets through protein aggregation.

Several computational tools have been proposed for protein aggregation prediction, including sequence-based and structure-based methods. Sequence-based models such as TANGO (Fernandez-Escamilla et al., 2004), AGGRESCAN (Conchillo-Solé et al., 2007), WALTZ (Maurer-Stroh et al., 2010), ANuPP (Prabakaran et al., 2021), and AggreProt (Planas-Iglesias et al., 2024) estimate aggregation propensity solely based on the intrinsic amino acid properties and sequence patterns. For example, TANGO uses a thermodynamic approach to consider interactions between residues—including hydrophobic interactions, charge effects, polarity, and volume effects—to predict the tendency of specific sequence regions to form cross- $\beta$  structures. ANuPP, a more recent sequence-based method, clusters amyloidogenic hexapeptides based on physicochemical properties like hydrophobicity and charge, then builds individual logistic regression models for each cluster before ensemble prediction. However, due to the lack of spatial context, sequence-based models may produce false positives as they cannot determine whether a predicted APR is buried within the hydrophobic core of a protein. Conversely, structure-based methods like CORDAX (Louros, Orlando, et al., 2020), SAP (Chennamsetty et al., 2009), Aggrescan3D (Zambrano et al., 2015), CamSol (Sormanni et al., 2015), and AggScore (Sankar et al., 2018) incorporate three-dimensional structural information, including atomic coordinates, surface hydrophobic patches, solvent accessibility, and the local microenvironment, to enhance predictive performance. Aggrescan3D, for instance, combines amino acid properties with solvent accessibility and spatial proximity, while CamSol adjusts intrinsic aggregation scores based on the solvent-accessible surface area of residues, facilitating more precise identification of aggregation hotspots, especially on protein surfaces. Despite their enhanced precision, these methods are limited by the availability of high-quality protein structural data, which is often scarce.

Given the significant limitations of current models, necessitated by inadequate and unbalanced training data and the sparse availability of high-resolution protein structures, there is a pressing need for more advanced methodologies. Recent breakthroughs in computational biology, such as the development of large-scale protein language models, like ESM2 (Lin et al., 2023), and



**FIGURE 1** Schematic of the AggNet framework. It comprises two submodules: APNet for amyloid peptide classification and APRNet for protein APR identification.

cutting-edge protein structure prediction models represented by AlphaFold2 (Tunyasuvunakool et al., 2021), have provided us potentially breakthrough solutions to solve the challenging aggregation prediction problem. These tools allow for the extraction of rich, contextual representations from amino acid sequences and have demonstrated remarkable accuracy in predicting three-dimensional protein structures directly from sequences. By integrating these advanced models, we propose AggNet, a comprehensive framework that leverages physicochemical, sequential, and structural information to predict protein aggregation, including amyloid peptide prediction and protein APR identification. For amyloid peptide prediction, AggNet utilizes the pre-trained protein language model ESM2 to extract informative representations and further fuse with traditional AAindex (Kawashima, 2000) features for amino acid sequence modeling. For APR identification, it employs the three-dimensional structures predicted by AlphaFold2 to incorporate spatial information. By combining the strengths of both sequence and structure-based approaches, AggNet aims to overcome their respective limitations and address the issue of data insufficiency. Our benchmark comparisons confirm that AggNet outperforms existing methods across various datasets. Additionally, a case study on MEDI1912 further verifies AggNet's effectiveness for protein engineering to rank different variants. Our findings suggest that AggNet is a useful tool for studying protein aggregation and can contribute significantly to the development of therapeutic proteins and novel biomaterials.

## 2 | RESULTS

### 2.1 | The architecture of AggNet

The architecture of AggNet, illustrated in Figure 1, comprises two primary submodules: APNet and APRNet. APNet is tasked with predicting amyloid peptides, while APRNet focuses on profiling protein APRs. APNet initiates the processing of peptide inputs through one-hot encoding and AAindex feature encoding. These encoded features are subsequently integrated with a feature fusion module that synthesizes an informative fused embedding. This embedding is then input into a multilayer perceptron (MLP) to estimate the aggregation propensity (AP) score of a peptide. In contrast, APRNet processes proteins by employing both a sequential and a structural channel. The sequential channel dissects the primary sequence into hexapeptides, assesses them using a sliding window approach, and scores them via the trained APNet to predict their intrinsic aggregation scores. Concurrently, the structural channel folds the corresponding 3D structure using AlphaFold2 and extracts spatial information, including neighborhood interactions and relative solvent accessible surface areas (RSA). After data preparation, this sequential and structural information is aggregated at the residue level to profile the aggregation characteristics of specific proteins. AggNet alleviates data scarcity by leveraging ESM2 and addresses the lack of structural data through the use of AlphaFold2. Moreover, the novel feature fusion module integrates sequential and

Model	ACC (%)	SE (%)	SP (%)	Q (%)	F1 (%)	MCC	AUC
TANGO	64.8	5.9	97.8	51.8	10.7	0.096	0.597
WALTZ	75.4	39.2	95.6	67.4	53.3	0.446	0.675
GAP	51.4	94.1	27.5	60.8	58.2	0.260	0.721
FishAmyloid	69.0	45.1	82.4	63.8	51.1	0.296	0.798
Pasta2	75.4	37.3	96.7	67.0	52.1	0.450	0.855
Aggrescan	79.6	68.6	85.7	77.2	70.7	0.551	0.855
ANuPP	83.1	82.4	83.5	82.9	77.8	0.645	0.883
AggNet	87.3	80.4	91.2	85.8	82.0	0.723	0.913

**TABLE 1** Benchmark comparison of amyloid peptide classification on the Hex142 dataset.

evolutionary information from ESM embeddings with physicochemical data from AAindex features, thereby enhancing the performance in both amyloid peptide classification and protein APR identification. More details can be found in the method section.

## 2.2 | AggNet's discriminative power in amyloid and non-amyloid peptide classification

Discriminating between amyloid and non-amyloid peptides is a critical task in predicting protein aggregation, essential for understanding the formation of amyloid fibers associated with various neurodegenerative diseases. We first conduct a benchmark comparison using the Hex142 dataset (see the method section) to highlight the advantages of AggNet in classifying amyloid and non-amyloid peptides.

First, we compare the AggNet with typical machine learning methods and AggNet exhibits superior performance over conventional machine learning algorithms including k-nearest neighbor (KNN), logistic regression (LR), Naive Bayes (NB), support vector machine (SVM), multilayer perceptron (MLP), in discriminating between amyloid and non-amyloid peptides. The comparative analysis, detailed in Tables S1 and S2, highlights AggNet's advancement over both basic one-hot feature-based models and more complex AAindex feature-based models, demonstrating a significant improvement in both F1 score and Matthews correlation coefficient (MCC). Importantly, while the best performing traditional models like MLP (using one-hot features) and Logistic Regression (using AAindex features) achieve an AUC of up to 0.902, AggNet surpasses these with notably higher F1 scores and MCC values, exceeding 80% in F1 score, indicative of its balanced prediction capability.

Furthermore, we compare the performance between AggNet and existing methods to show its superiority. The benchmark comparisons outlined in Table 1

confirm AggNet's state-of-the-art performance in amyloid peptide classification across various metrics including accuracy (ACC), sensitivity (SP), specificity (Q), F1 score, MCC, and area under the curve (AUC). AggNet achieves an accuracy of 87.3%, sensitivity of 80.4%, specificity of 91.2%, F1 score of 82.0%, MCC of 0.723, and AUC of 0.913, which is notably higher than the next best method, ANuPP, by about 3.0% in AUC and 4.2% in F1 score. This superior performance, particularly in balancing precision and recall in an unbalanced training dataset, underscores the potential of protein language models in enhancing peptide representation for improved amyloid prediction. Additionally, we conducted a fivefold cross-validation using the widely recognized WALTZ-DB 2.0 dataset to further assess AggNet's performance relative to existing methods. Results detailed in Table S3 reaffirm AggNet's superiority over contemporary approaches.

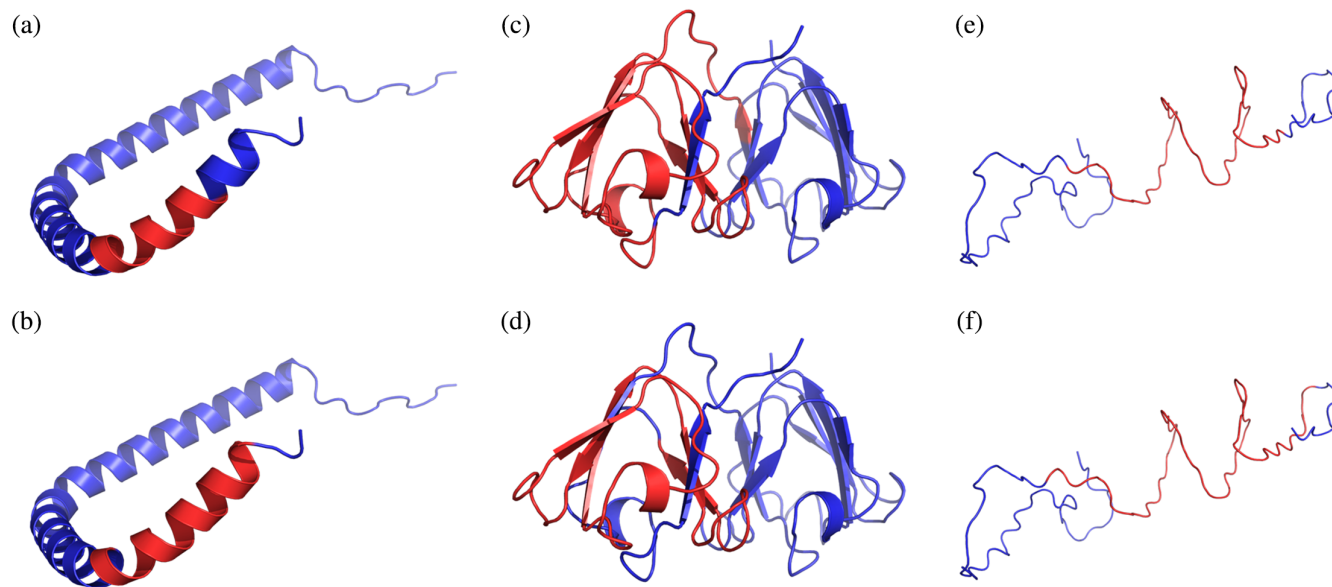
## 2.3 | AggNet's proficiency in identifying APRs

Identifying APRs helps pinpoint the aggregation hotspots within specific proteins, enhances understanding of aggregation mechanisms, and informs strategies to mitigate aggregation. Similarly, we conduct a benchmark study to validate the superiority of AggNet in performing this task across diverse proteins.

AggNet sets a new standard in identifying APRs of various proteins, outperforming existing models as shown in Table 2. It demonstrates substantial gains across all performance metrics compared with the second-best model, ANuPP. For instance, AggNet achieves a Segment Overlap (SOV) Overall score of 54.6 and an SOV Average score of 51.4, surpassing ANuPP by 8.8% and 5.5%, respectively. The significant performance drop when excluding structural information from AggNet's inputs highlights the critical role of spatial data in achieving these results.

**TABLE 2** Benchmark comparison of protein APR identification on the Amy37 dataset.

Model	SOV APR	SOV non-APR	SOV overall	SOV average	Total score <sup>a</sup>
Pasta2	13.2	24.9	23.2	19.1	42.3
WALTZ	44.4	28.9	28.7	36.6	65.3
FishAmyloid	14.5	45.2	37.5	29.9	67.4
Aggrescan	34.3	36.5	32.4	35.4	67.8
TANGO	19.1	57.8	48.1	38.5	86.6
ANuPP	45.2	52.3	50.2	48.7	98.9
AggNet (w/o structure)	27.2	59.2	51.0	43.2	94.3
AggNet	48.1	54.6	54.6	51.4	106.0

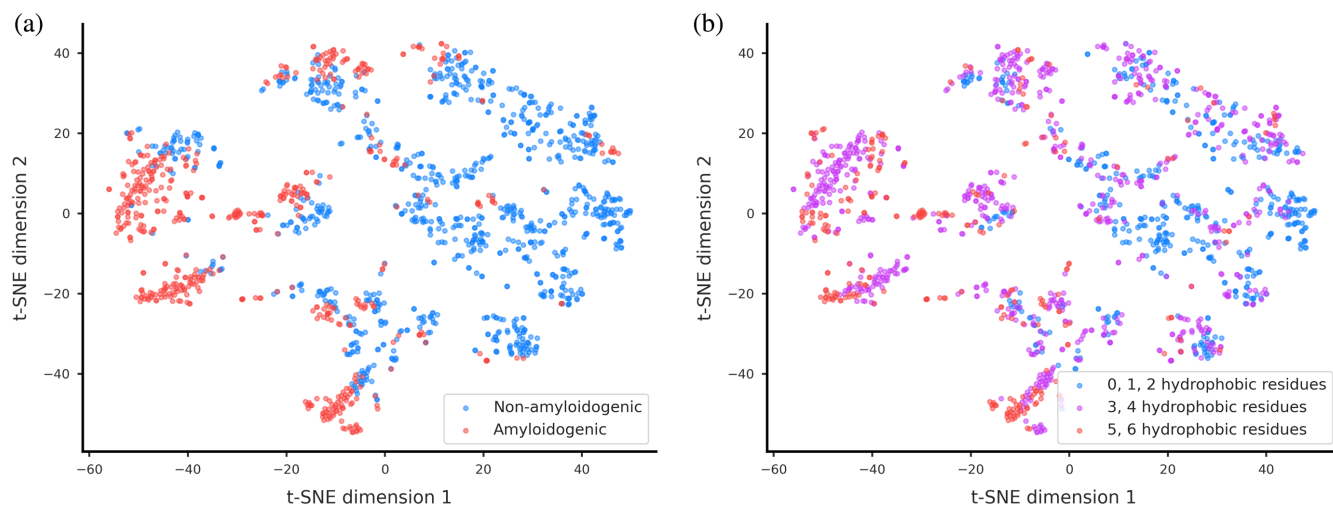
<sup>a</sup>Total score = SOV overall + SOV average.

**FIGURE 2** Comparison of predicted and experimental APRs. APRs are indicated in red. (a) Experimental APR of Apolipoprotein C-II. (b) Predicted APR of Apolipoprotein C-II. (c) Experimental APR of Gamma-crystallin D. (d) Predicted APR of Gamma-crystallin D. (e) Experimental APR of Chorion class A protein PC292. (f) Predicted APR of Chorion class A protein PC292.

Moreover, AggNet's ability to identify APRs across diverse secondary structures—helices, strands, or coils—is visually depicted in Figure 2. Notable examples include the experimental APRs of Apolipoprotein C-II, predominantly  $\alpha$ -helical, and Gamma-crystallin D, primarily  $\beta$ -strand, alongside Chorion class A protein PC292, which is mostly coiled. AggNet's predictions align closely with experimental data, albeit with slight deviations at the residue level, exemplifying its precision in localizing APRs across different protein structures. Specifically, the predicted APR of Apolipoprotein C-II is located in residues 61–76 while the experimental one is in residues 60–70 (Wilson et al., 2007). The predicted APR of Gamma-crystallin D is located in residues 91–146 while the experimental one is in residues 80–163 (Moran, Decatur, & Zanni, 2012; Moran, Woys, et al., 2012). The predicted

APR of Chorion class A protein PC292 is located in residues 46–103 while the experimental one is in residues 48–96 (Iconomidou et al., 2006).

## 2.4 | Ablation study and feature analysis

The input of AggNet is one-hot encodings of original peptide sequences and the corresponding AAindex features. AggNet's sophisticated integration of ESM embeddings and AAindex features results in an informative fused embedding that transcends traditional one-hot encoding and simple feature-based approaches. This fusion captures both sequence evolution information from ESM embeddings and physiochemical data from AAindex features, significantly enhancing the model's discriminatory



**FIGURE 3** The t-SNE visualization of peptide embeddings in AggNet. (a) Peptides are depicted as scatter points, color-coded by their labels: Amyloid or non-amyloid. (b) Peptides are color-coded based on the count of hydrophobic residues they contain.

capabilities. To enhance our understanding of AggNet's prediction mechanisms and improve its interpretability, we conduct an ablation study and feature analysis on the learned fused embeddings.

The results of the ablation study presented in Table S4 illustrate a significant decline in performance when the model is deprived of either ESM2 embeddings or AAindex features. Similarly, the removal of element-wise multiplication without a skip connection in the feature fusion process also leads to a comparable decrease in performance. These findings underscore the crucial impact of the feature fusion strategy on the final classification outcomes. By integrating our specifically designed feature strategy for ESM2 embeddings and AAindex features, the model achieves a more balanced predictive performance, particularly in the discrimination of amyloid peptides. This balance is vital for enhancing the model's accuracy and reliability in practical applications.

Dimensionality reduction and visualization using t-SNE (Hinton & Roweis, 2002), illustrated in Figure 3, reveal clear distinctions between amyloid and non-amyloid peptides. This separation is especially evident among peptides with varying counts of hydrophobic residues, demonstrating AggNet's ability to accurately reflect the physicochemical properties of peptides. It is evident that peptides with a lower number of hydrophobic residues (0–2) tend to cluster together, in contrast to peptides with a higher count (5–6), which are grouped separately. Peptides with a moderate number of hydrophobic residues (3–4) are distributed between these two extremes. This pattern underscores the model's capability to discern and represent the physicochemical properties of peptides effectively. Such clear stratification enhances the interpretability of how AggNet discriminates between amyloid

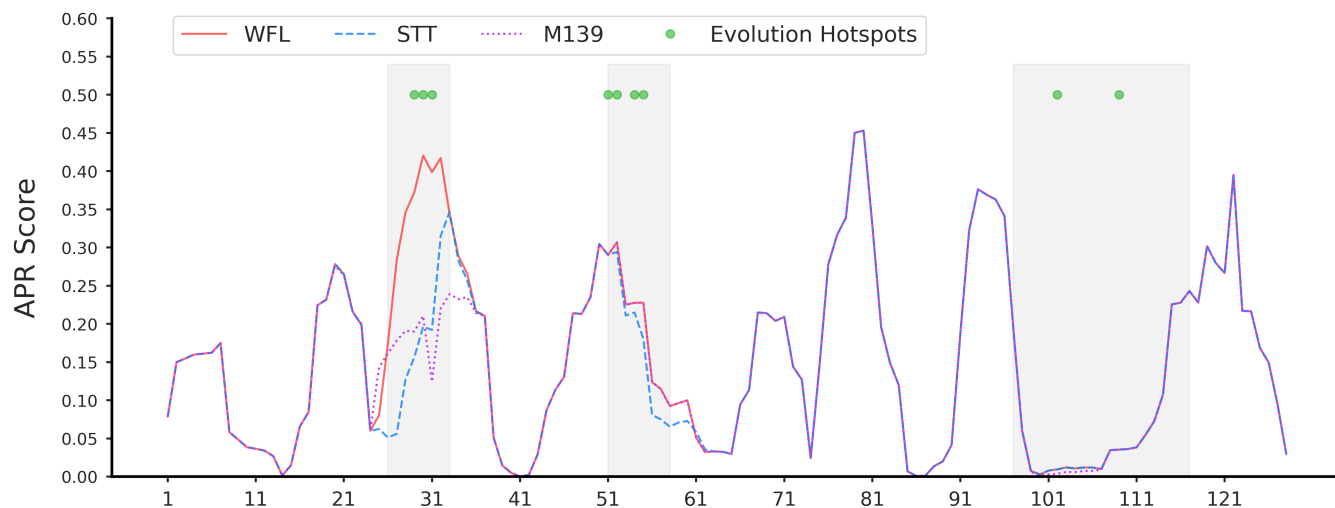
and non-amyloid peptides, showcasing its analytical strength.

## 2.5 | Case study: MEDI1912 antibody

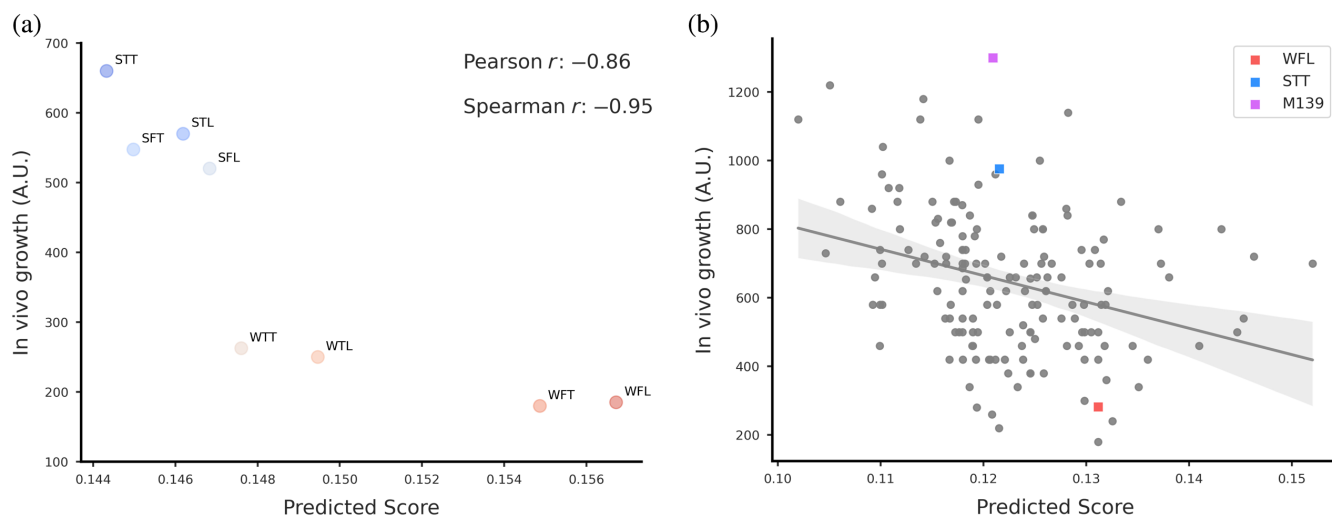
To further validate AggNet's utility in protein engineering aimed at reducing aggregation, we analyzed variants of the MEDI1912 antibody using the tripartite  $\beta$ -lactamase enzyme assay (TPBLA). This novel *in vivo* method correlates with traditional aggregation assays like HP-SEC and AC-SINS, as evidenced by the area under the bacterial growth curve measurements (Ebo et al., 2020).

Analysis of MEDI1912 and its variants—specifically, STT and M139—using AggNet revealed distinct APR score distributions in their VH domains, shown in Figure 4. Notably, the STT variant, known for its reduced aggregation propensity (Dobson et al., 2017), and M139, the most effective variant evolved by TPBLA (Ebo et al., 2020), showed significant differences in the CDRH1 region compared with wild-type MEDI1912. Specifically, the predicted scores by AggNet show that WFL has a high peak in the CDR1, STT has a lower peak, while the M139 eliminates the peak. These findings align well with experimental outcomes, demonstrating AggNet's potential to predict how specific mutations influence aggregation tendencies.

Additionally, AggNet's ability to rank the aggregation propensity of MEDI1912 variants was rigorously tested. The model's predictions for the seven specific variants at specific sites correlate strongly with experimental *in vivo* growth (A.U.) values, with Pearson and Spearman correlation coefficients of  $-0.86$  and  $-0.95$ , respectively. Consistent with the experimental results, WFL exhibits the



**FIGURE 4** Aggregation profile of the VH Domain in WFL, STT, and M139 as predicted by AggNet. The CDR regions are highlighted in gray.



**FIGURE 5** Performance of AggNet on the MEDI1912 dataset from TPBLA in vivo experiments. (a) Scatter plot of different WFL variants with mutations in specific sites. (b) Scatter plot of WFL variants evolved by TPBLA.

highest aggregation propensity and receives the highest predicted score, while STT shows the converse. Hence, we verify the efficacy of AggNet to guide the mutation design for aggregation mitigation by scoring different protein variants of MEDI1912.

Furthermore, we evaluated whether AggNet remains effective when applied to WFL variants harboring multiple mutations. A total of 162 variants evolved via TPBLA were analyzed, 115 of which contain more than three mutations (Ebo et al., 2020). The results are presented in Figure 5b, omitting the seven variants of WFL previously discussed. The scatter plot delineates a negative correlation between predicted aggregation scores and in vivo growth rates (A.U.), suggesting an intuitive inverse relationship. Although the specific predicted values lack

**TABLE 3** Performance comparison on MEDI1912 dataset.

Model	Pearson $r$	Spearman $r$
Aggrescan3D	$-0.07$	$-0.11$
CamSol	$0.29$	$0.29$
AggNet	$-0.32$	$-0.34$

precision, they generally mirror the trend observed in experimental data: higher aggregation propensities correlate with reduced experimental growth rates. Notably, AggNet demonstrates superior performance with a Pearson correlation coefficient of  $-0.32$  and a Spearman correlation coefficient of  $-0.34$ , surpassing state-of-the-art methods Aggrescan3D and CamSol, as shown in Table 3.

This comparison further underscores AggNet's advantage in ranking protein variants, efficiently capturing the impact of multiple mutations on the aggregation landscape without necessitating recalibration or additional data. Collectively, these findings corroborate the utility of AggNet in protein aggregation studies and its potential in guiding mutation design to reduce aggregation in therapeutic proteins.

### 3 | DISCUSSION AND CONCLUSION

In this study, we introduced AggNet, a deep learning framework that leverages physicochemical, sequential, and structural information to address the challenge of protein aggregation, including amyloid peptide prediction and protein APR identification. To mitigate the issue of insufficient data and to derive more informative amino acid sequence embeddings, we utilized the protein language model ESM2 (Lin et al., 2023) for feature extraction. We further enhanced the integration of physicochemical features and evolutionary information from ESM2 through a feature fusion layer, facilitating the learning of fused embeddings.

Performance comparisons in amyloid peptide classification and analyses of the fused embedding landscape affirm the efficacy of our feature fusion approach. Moreover, to augment the prediction accuracy in APR identification, we employed AlphaFold2 (Mirdita et al., 2022; Tunyasuvunakool et al., 2021) to model the corresponding structures and extract spatial information. By leveraging aggregation scores from neighboring residues and their RSA, AggNet achieved state-of-the-art performance, significantly outperforming existing methods. Visual comparisons of predicted and experimental APRs across various secondary structures—helices, strands, and coils—illustrate AggNet's precision in locating APRs.

AggNet addresses several limitations of existing models in predicting APRs in proteins. Sequence-based models such as ANuPP often struggle to identify the hydrophobic core of proteins, leading to a high rate of false positive predictions. In contrast, AggNet not only incorporates sequential context to score each residue but also leverages spatial neighbor information, resulting in more reliable APR predictions and a reduced false positive rate. Similarly, while Aggrescan3D relies on a scoring formula based on intrinsic aggregation scores derived from prior experimental data, these scores are static for a given residue regardless of its context within the protein structure, thereby constraining performance. AggNet overcomes this limitation by dynamically computing the intrinsic aggregation score for each residue based on its

sequential context, utilizing APNet, a deep learning-based framework. This approach enhances the flexibility and accuracy of the scoring process.

Beyond benchmark comparisons, we employed the MEDI1912 antibody as a case study to simulate a real-world protein engineering task. This included an examination of how AggNet's predicted APR profiles influence experimental outcomes by comparing the APR score landscapes in the VH domains of WFL, STT, and M139 variants. We also assessed AggNet's ranking performance on WFL variants with three mutations and on TPBLA-evolved WFL variants with more than three mutations. The results confirmed that, even without recalibration or fine-tuning with additional data, AggNet provides reasonable predictions, underscoring its utility in protein engineering for aggregation mitigation.

AggNet represents a valuable tool for protein aggregation research and has potential applications in protein engineering. The integration of deep learning technology and protein language models offers promising insights for this field. However, there are limitations to AggNet, such as its suboptimal blind test performance in downstream tasks, like protein engineering, where it does not directly offer mutational recommendations. Besides, predicting aggregation effects of multiple mutations is still challenging for AggNet. Advancing the development of more effective aggregation prediction models remains both challenging and essential, particularly for variants that exhibit multiple mutations. Continued innovation from the computer science and deep learning communities is expected to further advance the field, enhancing the predictive accuracy and utility of protein aggregation prediction.

### 4 | MATERIALS AND METHODS

#### 4.1 | Dataset preparation

For amyloid peptide identification, we have two schemes including train-test and cross-validation. For the train-test split, we used the Hex1421 dataset, which comprises 1421 experimentally validated amyloidogenic and non-amyloidogenic hexapeptides sourced from the CPAD 2.0 database (Rawat et al., 2020). This dataset is non-redundant, ensuring a broad representation of peptide varieties. We adhered to a 90%:10% division for training (Hex1279) and testing sets (Hex142), respectively, resulting in 1279 peptides for training and 142 for validation, maintaining consistency with the train-test split used in ANuPP. Detailed dataset characteristics are provided in Table S5. For the additional fivefold cross-validation, a dataset comprising 1416 non-redundant peptides was



curated from WALTZ-DB 2.0 (Louros, Konstantoulea, et al., 2020), which stands as one of the largest publicly accessible repositories of experimentally validated amyloidogenic peptides. This dataset includes 901 amyloid hexapeptides and 515 non-amyloid hexapeptides. Each fold is based on a 4:1 stratified sampling of amyloid and non-amyloid categories to avoid data bias.

For protein APR identification, we adopted ANuPP's methodology, selecting 162 proteins from the AmyPro database (Varadi et al., 2018). Proteins with unclear or ambiguous APR annotations or those with APR residue fractions outside the 10%–95% range were excluded. We applied CD-HIT (Huang et al., 2010) to cluster these sequences at 40% sequence identity to reduce redundancy, culminating in a distilled set of 54 non-redundant proteins. These were further divided based on the presence of hexapeptides into two subsets: Amy17, containing proteins with more than one hexapeptide overlapping with Hex1279, and Amy37, comprising proteins with none or one hexapeptide from Hex1279, serving respectively for calibration and assessment.

In the context of a case study on MEDI1912, a targeted dataset was compiled, consisting of seven scFv variants of WFL, each displaying mutations at specified sites. The specific variants, derived via TPBLA (Ebo et al., 2020), include WFT, WTL, WTT, SFL, STL, SFT, and STT, where “WFL” denotes W35/F36/L64, “STT” signifies S35/T36/T64 mutations, and so on. Additionally, a more extensive dataset of scFv variants evolved through TPBLA, was acquired from the corresponding author for comprehensive analysis. The detailed statistical information of these mutants is summarized in Table S6.

## 4.2 | Model architecture

### 4.2.1 | APNet architecture

APNet integrates the ESM2 module, a feature engineering module, a feature fusion module (illustrated in Figure S1), and a prediction module. Based on existing research and insights from biophysics, the amyloidogenic properties of peptides are primarily influenced by their intrinsic physicochemical characteristics. To capture these properties, we employ features derived from the AAindex database to represent each peptide. In addition to traditional feature representations like AAindex, recent studies have demonstrated that protein language models, such as ESM2, are capable of effectively learning sequential, evolutionary, and physicochemical features of amino acid sequences, achieving notable performance across various tasks. Therefore, we hypothesize that integrating these two complementary and diverse feature

sets—AAindex-derived features and representations from protein language models—can enhance the model's ability to discern the discriminative boundary between amyloidogenic and non-amyloidogenic peptides. ESM2 extracts embeddings for each residue, producing a 1280-dimensional vector encapsulating evolutionary and sequential information. These are dimensionally reduced to 256 dimensions via a multilayer perceptron (MLP) (Taud & Mas, 2018) and aggregated through mean pooling to form a peptide-level representation, denoted as  $x$ . In parallel, the feature engineering module initially narrows down the set of 3396 AAindex features to the 600 most informative ones using ANOVA. Subsequently, this refined dataset is processed through an additional MLP to transform the non-redundant physicochemical AAindex features into a 256-dimensional embedding, denoted as  $y$ . These embeddings, denoted as  $x$  and  $y$ , are then combined in the feature fusion module to form a more expressive representation  $z = f(x, y) + x = x * y + x$ . Here, we utilize element-wise multiplication and skip connection (He et al., 2016) for feature expression capabilities enhancement and efficient learning, respectively. This fused representation is further processed by the final MLP in the aggregation prediction head to yield the final aggregation propensity score, with higher scores indicating a greater intrinsic propensity for aggregation.

### 4.2.2 | APRNet architecture

APRNet comprises modules for sequence information extraction, structure prediction, and score aggregation. For sequence data, protein sequences are dissected into hexapeptides using sliding windows, each scored via APNet. The intrinsic aggregation score for each residue is calculated as the average score from all hexapeptides encompassing that residue. For instance, the aggregation score for residue with index 9 is the average of the predicted scores for hexapeptides spanning residues 4–9, 5–10, 6–11, 7–12, 8–13, and 9–14. For the structure part, AlphaFold2 predicts the protein's 3D structure, facilitating the computation of an adjacency matrix and relative solvent accessible surface area (RSA) for each residue to identify spatial neighborhoods within an 8 Å radius and determine exposure levels. The aggregation score for each residue is adjusted based on its neighbors' scores, their distances, and RSA, following the formula:  $E_i = \sum_j s_j \cdot e^{\alpha \cdot r_j} \cdot e^{\beta \cdot d_j}$ , where  $\alpha$  and  $\beta$  are tunable parameters reflecting the increased aggregation contribution from surface-exposed residues and proximal neighbors (Bárceñas et al., 2024). APRs are then delineated by identifying peak scores above a set threshold  $t_{peak}$  and extending these peaks in both directions to include neighboring

residues scoring above a secondary threshold  $t_{expand}$ . These parameters  $\{\alpha, \beta, t_{peak}, t_{expand}\}$  are optimized using the Amy17 dataset (Prabakaran et al., 2021).

### 4.2.3 | Training settings

The ESM2 module was kept frozen when training to prevent catastrophic forgetting caused by limited training data. This precaution ensures that the pre-trained model retains its learned features without adverse modifications. Alongside, we utilized the AdamW optimizer, favored for its adept management of weight decay and enhanced convergence characteristics, with an initial learning rate of  $2e-4$ . This rate was carefully chosen to accelerate training while maintaining the stability of the learning process. To address potential exploding gradient problems, we implemented gradient clipping with a norm threshold set at 1.0. APNet was trained for 25 epochs using batch sizes of 50 because we found that extending the training period led to negligible performance improvements. This batch size effectively balances computational resource utilization and the accuracy of gradient estimates, optimizing overall training efficiency. For hyper-parameters searching, we use the AutoML technique to determine the best one.

### 4.3 | Evaluation metrics

For the amyloid peptide identification, common binary classification metrics are used to evaluate the performance of the proposed methods, including accuracy (ACC), sensitivity (SE), specificity (SP), Q score, F1 score, and Matthew correlation coefficient (MCC). The formulas of these metrics are described as follows:

$$\left\{ \begin{array}{l} \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Q} = \frac{\text{SE} + \text{SP}}{2} \\ \text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \\ \text{MCC} = \frac{\text{TP} \times \text{TN} + \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \end{array} \right.$$

where TP is the number of true-positive samples, FP is the number of false-positive samples, TN is the number of true-negative samples, and FN is the number of false-negative samples. Besides, another metric called the area under the receiver operating characteristic curve (AUC) is used for the overall performance evaluation.

For the protein APR identification, Segment Overlap (SOV) (Zemla et al., 1999) scores evaluate the prediction accuracy based on the overlap between the predicted and actual segments instead of residues, which is more appropriate for segment prediction similar to secondary structure prediction. Four different SOV scores, SOV APR, SOV non-APR, SOV Overall, and SOV Average are used to evaluate the performance of predicting APRs in proteins. The formulation of these metrics is described below:

$$\left\{ \begin{array}{l} S(i) = \{(s_1^i, s_2^i) : \exists s_2^i s_1^i \cap s_2^i \neq \emptyset\} \\ S'(i) = \{(s_1^i, s_2^i) : \forall s_2^i s_1^i \cap s_2^i = \emptyset\} \\ \delta(s_1^i, s_2^i) = \min \left\{ (\maxov(s_1^i, s_2^i) - \minov(s_1^i, s_2^i)), \right. \\ \left. \minov(s_1^i, s_2^i), \left\lfloor \frac{L(s_1^i)}{2} \right\rfloor, \left\lfloor \frac{L(s_2^i)}{2} \right\rfloor \right\} \\ S_i = \sum_{S(i)} \frac{\minov(s_1^i, s_2^i) + \delta(s_1^i, s_2^i)}{\maxov(s_1^i, s_2^i)} \cdot L(s_1^i) \\ N_i = \sum_{S(i)} L(s_1^i) + \sum_{S'(i)} L(s_1^i) \\ \text{SOV APR} = \text{SOV}(1) = \frac{S_1}{N_1} \times 100 \\ \text{SOV nonAPR} = \text{SOV}(0) = \frac{S_0}{N_0} \times 100 \\ \text{SOV AVG} = \frac{\text{SOV}(1) + \text{SOV}(0)}{2} \\ \text{SOV Overall} = \frac{S_0 + S_1}{N_0 + N_1} \times 100. \end{array} \right.$$

where  $s_1^i$  is the segment of the ground truth label with state  $i$  and  $s_2^i$  is the segment of the predicted label with state  $i$ . State  $i=1$  and  $i=0$  denote the APR region and non-APR region, respectively.  $\maxov(s_1^i, s_2^i)$  and  $\minov(s_1^i, s_2^i)$  are the maximal and minimal overlap the length between  $s_1^i$  and  $s_2^i$ .  $L(s_1^i)$  and  $L(s_2^i)$  are the length of  $s_1^i$  and  $s_2^i$ .

### AUTHOR CONTRIBUTIONS

**Wenjia He:** Conceptualization; methodology; writing – original draft; visualization; investigation; validation; data curation. **Xiaopeng Xu:** Validation; writing – review and editing; visualization; supervision. **Haoyang Li:** Writing – review and editing; visualization;

validation. **Juexiao Zhou**: Writing – review and editing; visualization; validation. **Xin Gao**: Resources; project administration; writing – review and editing; funding acquisition; supervision.

## ACKNOWLEDGMENTS

The authors would like to thank Prof. David Brockwell for providing the experimental data of MEDI1912 from TPBLA. This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01, REI/1/5992-01-01, URF/1/4663-01-01, Center of Excellence for Smart Health (KCSH), under award number 5932, and Center of Excellence on Generative AI, under award number 5940.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## ORCID

Wenjia He  <https://orcid.org/0000-0001-8161-4642>

## REFERENCES

- Bárceñas O, Kuriata A, Zalewski M, Iglesias V, Pintado-Grima C, Firlik G, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. *Nucleic Acids Res.* 2024; 52(W1):W170–5. <https://doi.org/10.1093/nar/gkae382>
- Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci.* 2009;106(29):11937–42. <https://doi.org/10.1073/pnas.0904191106>
- Conchillo-Solé O, De Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESKAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinform.* 2007;8(1):65. <https://doi.org/10.1186/1471-2105-8-65>
- Dobson J, Kumar A, Willis LF, Tuma R, Higazi DR, Turner R, et al. Inducing protein aggregation by extensional flow. *Proc Natl Acad Sci.* 2017;114(18):4673–8. <https://doi.org/10.1073/pnas.1702724114>
- Ebo JS, Saunders JC, Devine PWA, Gordon AM, Warwick AS, Schiffrin B, et al. An in vivo platform to select and evolve aggregation-resistant proteins. *Nat Commun.* 2020;11(1):1816. <https://doi.org/10.1038/s41467-020-15667-1>
- Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol.* 2004; 22(10):1302–6. <https://doi.org/10.1038/nbt1012>
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV: IEEE; 2016. p. 770–8. <https://doi.org/10.1109/CVPR.2016.90>
- Hinton G, Roweis S. Stochastic neighbor embedding. *Advances in neural information processing systems* 15. Cambridge: MIT Press; 2002.
- Housmans JAJ, Wu G, Schymkowitz J, Rousseau F. A guide to studying protein aggregation. *FEBS J.* 2023;290(3):554–83. <https://doi.org/10.1111/febs.16312>
- Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* 2010;26(5):680–2. <https://doi.org/10.1093/bioinformatics/btq003>
- Iconomidou VA, Chryssikos GD, Gionis V, Galanis AS, Cordopatis P, Hoenger A, et al. Amyloid fibril formation propensity is inherent into the hexapeptide tandemly repeating sequence of the central domain of silkworm chorion proteins of the A-family. *J Struct Biol.* 2006;156(3):480–8. <https://doi.org/10.1016/j.jsb.2006.08.011>
- Janssen K, Claes F, Van De Velde D, Wehbi VL, Houben B, Lampi Y, et al. Exploiting the intrinsic misfolding propensity of the KRAS oncoprotein. *Proc Natl Acad Sci.* 2023;120(9): e2214921120. <https://doi.org/10.1073/pnas.2214921120>
- Kahn SE, Andrikopoulos S, Verchere CB. Islet amyloid: a long-recognized but underappreciated pathological feature of type 2 diabetes. *Diabetes.* 1999;48(2):241–53. <https://doi.org/10.2337/diabetes.48.2.241>
- Kawashima S. AAindex: amino acid index database. *Nucleic Acids Res.* 2000;28(1):374. <https://doi.org/10.1093/nar/28.1.374>
- Koo EH, Lansbury PT, Kelly JW. Amyloid diseases: abnormal protein aggregation in neurodegeneration. *Proc Natl Acad Sci.* 1999;96(18):9989–90. <https://doi.org/10.1073/pnas.96.18.9989>
- Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;379(6637):1123–30. <https://doi.org/10.1126/science.ade2574>
- Louros N, Konstantoulea K, De Vleeschouwer M, Ramakers M, Schymkowitz J, Rousseau F. WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.* 2020; 48(D1):D389–93. <https://doi.org/10.1093/nar/gkz758>
- Louros N, Orlando G, De Vleeschouwer M, Rousseau F, Schymkowitz J. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat Commun.* 2020;11(1):3314. <https://doi.org/10.1038/s41467-020-17207-3>
- Lowe D, Dudgeon K, Rouet R, Schofield P, Jermutus L, Christ D. Aggregation, stability, and formulation of human antibody therapeutics. *Adv Protein Chem Struct Biol.* 2011;84:41–61. <https://doi.org/10.1016/B978-0-12-386483-3.00004-5>
- Lundahl MLE, Fogli S, Colavita PE, Scanlan EM. Aggregation of protein therapeutics enhances their immunogenicity: causes and mitigation strategies. *RSC Chem Biol.* 2021;2(4):1004–20. <https://doi.org/10.1039/D1CB00067E>
- Matson JB, Stupp SI. Self-assembling peptide scaffolds for regenerative medicine. *Chem Commun.* 2012;48(1):26–33. <https://doi.org/10.1039/C1CC15551B>
- Maurer-Stroh S, Debulpaep M, Kuemmerer N, De La Paz ML, Martins IC, Reumers J, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods.* 2010;7(3):237–42. <https://doi.org/10.1038/nmeth.1432>
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to

- all. *Nat Methods*. 2022;19(6):679–82. <https://doi.org/10.1038/s41592-022-01488-1>
- Moran SD, Decatur SM, Zanni MT. Structural and sequence analysis of the human  $\gamma$ D-crystallin amyloid fibril core using 2D IR spectroscopy, segmental  $^{13}\text{C}$  labeling, and mass spectrometry. *J Am Chem Soc*. 2012;134(44):18410–6. <https://doi.org/10.1021/ja307898g>
- Moran SD, Woys AM, Buchanan LE, Bixby E, Decatur SM, Zanni MT. Two-dimensional IR spectroscopy and segmental  $^{13}\text{C}$  labeling reveals the domain structure of human  $\gamma$ D-crystallin amyloid fibrils. *Proc Natl Acad Sci*. 2012;109(9):3329–34. <https://doi.org/10.1073/pnas.1117704109>
- Perchiacca JM, Tessier PM. Engineering aggregation-resistant antibodies. *Annu Rev Chem Biomol Eng*. 2012;3(1):263–86. <https://doi.org/10.1146/annurev-chembioeng-062011-081052>
- Planas-Iglesias J, Borko S, Swiatkowski J, Elias M, Havlasek M, Salamon O, et al. AggreProt: a web server for predicting and engineering aggregation prone regions in proteins. *Nucleic Acids Res*. 2024;52(W1):W159–69. <https://doi.org/10.1093/nar/gkae420>
- Prabakaran R, Rawat P, Kumar S, Michael Gromiha M. ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. *J Mol Biol*. 2021;433(11):166707. <https://doi.org/10.1016/j.jmb.2020.11.006>
- Rahban M, Ahmad F, Piatyszek MA, Haertlé T, Saso L, Saboury AA. Stabilization challenges and aggregation in protein-based therapeutics in the pharmaceutical industry. *RSC Adv*. 2023;13(51):35947–63. <https://doi.org/10.1039/D3RA06476J>
- Rawat P, Prabakaran R, Sakthivel R, Mary Thangakani A, Kumar S, Gromiha MM. CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid*. 2020;27(2):128–33. <https://doi.org/10.1080/13506129.2020.1715363>
- Ren C, Wen X, Mencius J, Quan S. An enzyme-based biosensor for monitoring and engineering protein stability in vivo. *Proc Natl Acad Sci*. 2021;118(13):e2101618118. <https://doi.org/10.1073/pnas.2101618118>
- Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. *Nat Med*. 2004;10(S7):S10–7. <https://doi.org/10.1038/nm1066>
- Sankar K, Krystek SR, Carl SM, Day T, Maier JKX. AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins: Struct, Funct, Bioinf*. 2018;86(11):1147–56. <https://doi.org/10.1002/prot.25594>
- Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol*. 2015;427(2):478–90. <https://doi.org/10.1016/j.jmb.2014.09.026>
- Taud H, Mas JF. Multilayer perceptron (MLP). In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F, editors. Geomatic approaches for modeling land change scenarios. Cham: Springer International Publishing; 2018. p. 451–5. [https://doi.org/10.1007/978-3-319-60801-3\\_27](https://doi.org/10.1007/978-3-319-60801-3_27)
- Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Židek A, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596(7873):590–6. <https://doi.org/10.1038/s41586-021-03828-1>
- Varadi M, De Baets G, Vranken WF, Tompa P, Pancsa R. AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res*. 2018;46(D1):D387–92. <https://doi.org/10.1093/nar/gkx950>
- Wang X, Horii A, Zhang S. Designer functionalized self-assembling peptide nanofiber scaffolds for growth, migration, and tubulogenesis of human umbilical vein endothelial cells. *Soft Matter*. 2008;4(12):2388. <https://doi.org/10.1039/b807155a>
- Wilson LM, Mok Y-F, Binger KJ, Griffin MDW, Mertens HDT, Lin F, et al. A structural Core within apolipoprotein C-II amyloid fibrils identified using hydrogen exchange and proteolysis. *J Mol Biol*. 2007;366(5):1639–51. <https://doi.org/10.1016/j.jmb.2006.12.040>
- Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res*. 2015;43(W1):W306–13. <https://doi.org/10.1093/nar/gkv359>
- Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Struct, Funct, Bioinf*. 1999; 34(2):220–3. [https://doi.org/10.1002/\(SICI\)1097-0134\(19990201\)34:2<220::AID-PROT7>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1097-0134(19990201)34:2<220::AID-PROT7>3.0.CO;2-K)

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** He W, Xu X, Li H, Zhou J, Gao X. AggNet: Advancing protein aggregation analysis through deep learning and protein language model. *Protein Science*. 2025; 34(2):e70031. <https://doi.org/10.1002/pro.70031>