Published in final edited form as:
J Chem Inf Model. 2023 September 25; 63(18): 5727–5733. doi:10.1021/acs.jcim.3c00817.

AggBERT: Best in Class Prediction of Hexapeptide Amyloidogenesis with a Semi-Supervised ProtBERT Model

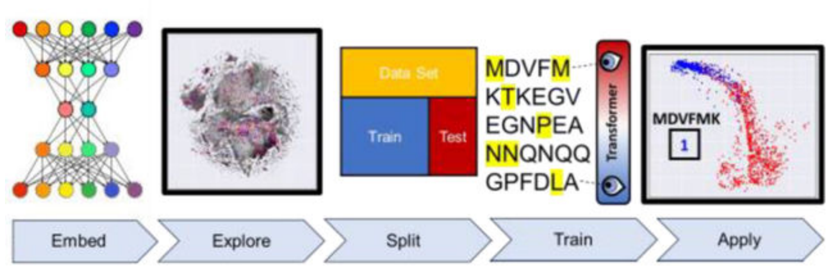
Ryann Perez[a], **Xinning Li**[a], **Sam Giannakoulias**[*,a], **E. James Petersson**[*,a]

[a]Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Abstract

The prediction of peptide amyloidogenesis is a challenging problem in the field of protein folding. Large language models, such as the ProtBERT model, have recently emerged as powerful tools in analyzing protein sequences for applications such as predicting protein structure and function. In this letter, we describe the use of a semi-supervised and fine-tuned ProtBERT model to predict peptide amyloidogenesis from sequence alone. Our approach, which we call AggBERT, achieved state-of-the-art performance, demonstrating the potential for large language models to improve the accuracy and speed of amyloid fibril prediction over simple heuristics or structure-based approaches. This work highlights the transformative potential of machine learning and large language models in the fields of chemical biology and biomedicine.

Graphical Abstract



1. INTRODUCTION

Biologics, namely proteins and peptides, have found numerous applications in biotechnology and medicine, serving as catalysts, antibodies, and signaling molecules.[1–4] Peptides in particular have demonstrated efficacy as therapeutics, serving as hormones or modulators of protein-protein interactions.[5] However, the formation of amyloid-like structures through a process called amyloidogenesis is a persistent challenge in the

[*] **Corresponding Authors**: SGG: gianna1@sas.upenn.edu; EJP: ejpetersson@sas.upenn.edu.

Author Contributions

R.P, X.L. and S.G.G. performed feature generation, model training, and data analysis guided by E.J.P. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

SUPPORTING INFORMATION

Information on how amino acids were grouped, autoencoder training results, clustering heuristic curves, additional embedding plots, clustering motif analysis, and additional links to our Github (PDF) are available free of charge at http://pubs.acs.org.

development of novel peptide therapeutics.[6] Amyloidogenesis leads to a significant decrease in half-life, and therapeutic agents that aggregate when administered *in vivo* have shown significantly more immunogenicity than their nonaggregating counterparts.[7] Furthermore, amyloidogenesis of native polypeptides is a threat to human health, as many prevalent diseases including type II Diabetes Mellitus, Alzheimer's Disease, and Parkinson's Disease are associated with protein or peptide amyloidogenesis.[8]

While amyloidogenesis is often associated with unfolded or misfolded proteins, recent evidence suggests that even natively folded, globular proteins can undergo amyloid-like aggregation.[9, 10] Characteristics such as general hydrophobicity, β-strand propensity, and low net charge have been commonly attributed to amyloid-like sequences and have been used to develop many of the predictive models for amyloid-like aggregation.[11] However, functional amyloids found in bacteria and yeast have challenged this view, as they possess properties that do not conform to the traditional heuristics for predicting amyloidogenic sequences.[8] Further investigation has revealed that globular proteins additionally contain a significant number of aggregation-prone regions (APRs) throughout their polypeptide sequences.[9, 10] These short, buried peptide sequences stabilize protein tertiary structure and can contribute to amyloid formation when exposed to solvent. To better understand amyloid structure and identify APRs, alternative models have investigated residue-specific propensities for native folding versus aggregation.[12]

For instance, Waltz[13] makes predictions of pro-amyloidogenicity simply from a position specific scoring matrix derived from the WaltzDB[14] itself. The Aggrescan model[11] is similarly based on the presence of motifs from sliding windows in experiments involving amyloid-beta mutant fusion proteins. Other models like Tango[15] are simpler and are built upon beta-sheet secondary structure prediction. On the other hand, MetAmyl, Pasta 2.0 and GAP utilized machine learning to predict peptide pro-amyloidogenicity. MetAmyl[16] is a logistic regression model built on sequence and physiochemical property features. Pasta 2.0[17] and GAP[18] are very similar methods to MetAmyl, but both utilize support vector machines for prediction rather than logistic regression. However, a unique approach by Louros *et al*. demonstrated that a logistic regression machine learning model constructed from energetic terms in the FoldX empirical forcefield (Cordax) was most effective in predicting amyloidogenic peptides compared to prior methods.[19] These findings highlight the potential for new approaches to improve the prediction of pro-amyloid sequences to understand their potential implications in human disease and drug design.

Large language models (LLMs), such as the ProtBERT model, have revolutionized the field of natural language processing and have recently been applied to the analysis of protein sequences.[20, 21] ProtBERT is a transformer-based language model that has been pre-trained on a massive corpus of protein sequences (UniRef 90, ~106M sequences) and has shown remarkable performance in predicting protein structure and function.[22] In this work, we utilize a semi-supervised and fine-tuned ProtBERT model to predict peptide amyloidogenesis, taking advantage of its ability to learn complex sequence-phenotype relationships.[23, 24] Our approach, which we call AggBERT, represents a significant improvement in F1 score over traditional heuristics or structure-based methods and demonstrates the potential for LLMs to accelerate the discovery of new insights into protein/peptide amyloid formation.[11, 19]

2. COMPUTATIONAL METHODS

2.1 DATABASE AND MACHINE LEARNING STRATEGIES

To train and evaluate our LLMs for prediction of peptide amyloidogenesis, we used the Waltz-DB 2.0 database, which is the largest publicly compiled database of peptides with curated annotations of amyloidogenesis.[14] The Waltz-DB 2.0 database of 6-mer peptides contains 1399 peptides (when sequences present in both classes are removed), of which 507 are amyloidogenic and 892 are non-amyloidogenic.

Although previously published methods utilized heuristics or Leave-One-Out Cross Validation (LOO-CV) to benchmark their predictions, in this study, we employed a rigorous training and testing set that was constructed by unbiased sampling from a manifold of peptide 6-mer space.[25] This approach allows for a more accurate assessment of the generalizability of our model compared to the LOO-CV strategy which has been commonly used. Enumeration of our datasets can be found on our GitHub at https://github.com/ejp-lab/EJPLab_Computational_Projects/tree/master/AmyloidPrediction.

2.2 MANIFOLD LEARNING AND DATASET CURATION

In order to develop a manifold of 6-mer space useful for rigorous dataset curation, we began by generating a list of all 64 million possible 6-mer sequences. To featurize these sequences, we employed two strategies: one-hot encoding and bioinformatic features from the pySAR library.[26] One-hot encoding was employed using the single letter codes for the standard 20 amino acids as there are no positional ambiguities in the WaltzDB 2.0 database. The resulting feature files were too large for manifold learning directly, as the database files were on the order of tens of gigabytes. Therefore, we utilized the dask python library to create partitions that could be batch processed in memory.[27]

We trained two dense autoencoders using Keras and TensorFlow 2 to learn a low-dimensional manifold for sequence and bioinformatic features.[28, 29] The model architecture and other hyperparameters of these autoencoders were tuned using the Optuna python library to minimize reconstruction error (mean squared error loss, MSE loss) over 25 trials of Bayesian optimization.[30] Model embeddings were extracted from the bottleneck layers of the autoencoders.

To curate the training and testing datasets, we utilized the k-means clustering algorithm from scikit-learn on the combined one-hot encoded sequence features and pySAR manifold.[31, 32] The clustering was performed using the CuDF/CuML python libraries for faster processing on GPUs. The number of clusters was selected by optimizing for the minimum of the Davies-Bouldin score as the number of clusters increased.[33] Once a suitable k was identified, we constructed stratified samples from each cluster to create the training and testing datasets with a testing set size of 20%. Visualization of the manifolds were performed using the Uniform Manifold Approximation (UMAP) decomposition strategy from CuML.[34] The small number of outliers (~5) produced by GPU instability were removed if the point was 3 standard deviations from the manifold centroid and the resulting embeddings were plotted with matplotlib.[35]

2.3 SUPERVISED TRANSFORMER

We utilized a supervised transformer model as our first approach to predicting protein amyloidogenicity. This involved fine-tuning the ProtBERT model from the Hugging Face library on our curated training set.[36] To optimize the model's performance, we employed Optuna to tune its hyperparameters over 25 trials, with the F1 score from 5-fold cross-validation serving as the tuning metric. We also applied regularization to the binary cross-entropy loss function by incorporating the expected calibration error. Calibration was accomplished using a temperature term, which prevented the model's logits from becoming too hard early in the training phase. Once the models were tuned, we applied hard vote bagging with the five cross-validation models to label the curated testing set.[37]

2.4 SEMI-SUPERVISED TRANSFORMER

To further improve the quality of our supervised transformer model, we implemented a semi-supervised training approach. The entire space of 6-mer peptides, consisting of 64 million sequences, were labeled using the supervised transformer model. Highly confident predictions with a probability greater than 0.8 were selected to train a large semi-supervised transformer model. Given the size and training times for this model, we did not perform tuning, and training occurred for only one epoch. Finally, as before, the trained semi-supervised model was used to label the testing set for the final predictions.

RESULTS AND DISCUSSION

In this study, we developed a workflow to predict 6-mer peptide amyloidogenicity using LLMs. As shown in Figure 1, our workflow begins with manifold learning, which maps the high-dimensional representation of the peptides (one-hot encoded sequences plus bioinformatic descriptors) into a lower-dimensional space more suitable for analysis. We then curated training and testing datasets to train and benchmark a supervised transformer classification model on the Waltz-DB 2.0 database as detailed in the methods section. Finally, we improved the performance of the supervised transformer by using a semi-supervised training procedure that learned from both the labeled and unlabeled 6-mer peptide possibilities. We call this final model AggBERT.

To evaluate the characteristics of the Waltz-DB 2.0 database, we performed exploratory data analysis and investigated the relationship between sequence composition and amino acid type effects on amyloidogenicity. As shown in Figure 2, we observed that peptides containing valine and isoleucine were more likely to be pro-amyloidogenic whereas peptides including glutamine, serine, asparagine, arginine, glycine, and proline were more likely to be nonamyloidogenic. This is further supported from the by-group analysis demonstrating that nonpolar amino acids (Asp, Ilu, Leu, Met, Val) are more likely to be involved in amyloid aggregation whereas polar amino acids (Cys, Asn, Gln, Ser, Thr) are not. While these simple heuristics are consistent with previous studies, they incompletely describe amyloid space.

An important hypothesis that we wanted to test was that the Waltz-DB 2.0 database is representative of the greater 6-mer space and that models trained on this database would generalize well to novel amyloid sequences. Therefore, we employed a strategy of learning low-dimensional manifolds for 6-mers. In total, we learned three different manifolds for the 64 million 6-mer peptides according to one-hot sequence features (Supporting Information, SI, Figure S4), pySAR features (SI Figure S5), and the concatenation of the two embeddings (Figure 3). We observed that the one-hot sequence manifold was circular and uniform, which is unsurprising given that every possible combination of 6-mer was represented. However, the manifold for the pySAR features, which includes chemical, physiochemical, and autocorrelation features, exhibited unique and structured characteristics. Nevertheless, we found that the Waltz-DB 2.0 database had strong coverage across the combined manifold, and the response class labels were relatively separated by the Davies-Bouldin metric for 22 clusters. Additionally, we noted enrichment of class labels within the two-dimensional representation of embedding space. This indicates that our manifold learning approach effectively represents the high-dimensional sequence and bioinformatic space in a lower-dimensional space. This dataset curation method enabled us to robustly assess model generalization and learn more about amyloid space.

We performed model training and benchmarking after curating our datasets, which included a comparison of both the final AggBERT model and our supervised transformer model’s performance against other literature benchmarks. Our models outperformed the benchmarks for the overwhelming majority of classification metrics, as shown in Table 1. Particularly noteworthy was the large improvement in F1 score for the positive class, which is a significant step forward as it enables the model to better identify novel amyloid sequences from the greater 6-mer space. It is important to note that our models were evaluated using a held-out testing set, while the other literature benchmarks used a leave-one-out validation strategy. Therefore, the difference in model metrics between our approach and other literature benchmarks is more pronounced, providing a strong argument for our LLM method. The AggBERT semi-supervised approach further improved the benchmarking metrics over the supervised transformer model, indicating that the learned representations from Waltz-DB 2.0 were capable of labeling meaningful patterns in the greater 6-mer space, leading to further improvements in generalization.

In order to better understand the behavior of our two LLMs, we performed the following three analyses: classification histograms, receiver operating characteristic (ROC) plots, and visualized model embeddings. As shown in Figure 4, the classification histograms for the supervised transformer and the AggBERT semi-supervised transformer show good separation between the positive and negative classes, with probabilities near 0 and 1. The AggBERT semi-supervised model shows a slight decrease in model confidence which may serve as targets for future dataset expansion in an active learning approach. Additionally, as shown in Figure 5, the ROC plots for both models show strong performance, with area under the curve (AUC) values of 0.90. Finally, we visualized the embeddings of the supervised and semi-supervised transformer models. Displayed in Figure 6, the embeddings of the AggBERT semi-supervised transformer appear more continuous than those of the supervised transformer. This potentially suggests that larger scale training on more diverse sequences of the AggBERT semi-supervised transformer may be better at capturing a presumed gradient between amyloidogenic and non-amyloidogenic sequences (Figure S8). This gradient could hint that an accumulation of amylogenic features that increases the likelihood of amyloidogenicity. These model properties may be more useful in downstream analyses such as identifying novel amyloid sequences.

CONCLUSION

In this letter, we present a novel approach to predict peptide amyloidogenesis using AggBERT, a semi-supervised and fine-tuned ProtBERT model benchmarked on diligently generated datasets from the Waltz-DB 2.0 database. Our method achieves state-of-the-art performance and demonstrates the potential of LLMs to improve the accuracy and speed of amyloid aggregation prediction over structure-based approaches. This work highlights the transformative potential of machine learning and LLMs in the fields of chemical biology and biomedicine.

Furthermore, our approach is the first to use a rigorous testing set generated by unsupervised autoencoders, rather than simple leave-one-out cross-validation, to evaluate the performance of amyloidogenesis predictive models. This ensures that our results are robust and reliable, providing a basis for future work in this area. Additionally, the AggBERT model has the potential to be used for identifying new amyloid sequences within proteins, which could lead to important discoveries in disease diagnosis and treatment. We have made available the predictions of AggBERT for all possible hexamer sequences on our GitHub at https://github.com/ejp-lab/EJPLab_Computational_Projects/tree/master/AmyloidPrediction/ProspectivePredictions.

Overall, this work represents a significant step forward in the field of amyloidogenesis prediction, providing a powerful tool for academics and those in biologics drug discovery. In the future, we hope to combine the model presented here with models based on other experimental data sets for specific proteins, such as the collection of *in vitro* aggregation data for α-synuclein that we collected in Pancoe *et al.*[38] or the α-synuclein yeast screening data from Newberry *et al.*[39] More broadly, we are working to combine LLMs with molecular simulations in order to make further improvements on the customized score functions that we have previously developed for predicting the effects of mutations on the stability of protein interfaces,[40] identifying tolerated sites for unnatural amino acid incorporation,[41] and designing peptide probes based on modest-sized data sets.[42] We believe that there is a tremendous opportunity to develop highly effective predictive models by combining LLMs with simpler machine learning methods. By utilizing these techniques at their respective scales, LLMs can extract a vast amount of information from large unlabeled corpuses, while simpler physics-based ML (often involving *de novo* simulation) approaches can produce highly relevant features for describing smaller local datasets. With these tools at our disposal, we can improve our approaches in situations where we have previously only used one of these tools, such as identifying small molecule binding sites on α-synuclein fibrils[43] or designing protease inhibitors.[44] Of course, these represent just a small subset of the many biological applications of LLMs in combination with other machine learning methods and molecular simulations which can revolutionize our ability to interpret and predict biochemical phenomena.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

This work was supported by the University of Pennsylvania and the National Institutes of Health (R01-NS103873 to E.J.P.). S.G.G. thanks the National Science Foundation (NSF) for funding through the NSF Graduate Research Fellowship Program (DGE-1845298). R.M.P. thanks the NIH for funding through the Chemistry Biology Interface Training Program (T32 GM133398).

SOFTWARE AND DATA AVAILABILITY

All software, datasets, and scripts utilized in this publication can be found at our Github repository:

https://github.com/ejplab/EJPLab_Computational_Projects/tree/master/AmyloidPrediction/

Uncategorized References

(1). Tang L; Persky AM; Hochhaus G; Meibohm B Pharmacokinetic aspects of biotechnology products. J. Pharm. Sci. 2004, 93 (9), 2184–2204, Review. DOI: 10.1002/jps.20125. [PubMed: 15295780]

(2). Milletti F Cell-penetrating peptides: classes, origin, and current landscape. Drug Discov. Today 2012, 17 (15–16), 850–860, Review. DOI: 10.1016/j.drudis.2012.03.002. [PubMed: 22465171]

(3). Nelson AL; Dhimolea E; Reichert JM Development trends for human monoclonal antibody therapeutics. Nat. Rev. Drug Discov. 2010, 9 (10), 767–774, Article. DOI: 10.1038/nrd3229. [PubMed: 20811384]

(4). Husted AS; Trauelsen M; Rudenko O; Hjorth SA; Schwartz TW GPCR-Mediated Signaling of Metabolites. Cell Metab. 2017, 25 (4), 777–796, Review. DOI: 10.1016/j.cmet.2017.03.008. [PubMed: 28380372]

(5). Souroujon MC; Mochly-Rosen D Peptide modulators of protein-protein interactions in intracellular signaling. Nat. Biotechnol. 1998, 16 (10), 919–924, Review. DOI: 10.1038/nbt1098-919. [PubMed: 9788346]

(6). Zapadka KL; Becher FJ; dos Santos ALG; Jackson SE Factors affecting the physical stability (aggregation) of peptide therapeutics. Interface Focus 2017, 7 (6), 18, Article; Proceedings Paper. DOI: 10.1098/rsfs.2017.0030.

(7). Fernandez L; Bustos RH; Zapata C; Garcia J; Jauregui E; Ashraf GM Immunogenicity in Protein and Peptide Based-Therapeutics: An Overview. Curr. Protein Pept. Sci. 2018, 19 (10), 958–971, Review. DOI: 10.2174/1389203718666170828123449. [PubMed: 28847291]

(8). Chiti F; Dobson CM Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. 2006, 75, 333–366, Review; Book Chapter. DOI: 10.1146/annurev.biochem.75.101304.123901. [PubMed: 16756495]

(9). Pawar AP; DuBay KF; Zurdo J; Chiti F; Vendruscolo M; Dobson CM Prediction of “aggregation-prone” and “aggregation-susceptible” regions in proteins associated with neurodegenerative diseases. J. Mol. Biol. 2005, 350 (2), 379–392, Article. DOI: 10.1016/j.jmb.2005.04.016. [PubMed: 15925383]

(10). Tartaglia GG; Pawar AP; Campioni S; Dobson CM; Chiti F; Vendruscolo M Prediction of aggregation-prone regions in structured proteins. J. Mol. Biol. 2008, 380 (2), 425–436, Article. DOI: 10.1016/j.jmb.2008.05.013. [PubMed: 18514226]

(11). Conchillo-Sole O; de Groot NS; Aviles FX; Vendrell J; Daura X; Ventura S AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. BMC Bioinformatics 2007, 8, 17, Article. DOI: 10.1186/1471-2105-8-65. [PubMed: 17239245]

(12). de la Paz ML; Serrano L Sequence determinants of amyloid fibril formation. Proc. Natl. Acad. Sci. U. S. A. 2004, 101 (1), 87–92, Article. DOI: 10.1073/pnas.2634884100. [PubMed: 14691246]

(13). Oliveberg M Waltz, an exciting new move in amyloid prediction. Nat. Methods 2010, 7 (3), 187–188, Editorial Material. DOI: 10.1038/nmeth0310-187. [PubMed: 20195250]

(14). 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–D393, Article. DOI: 10.1093/nar/gkz758. [PubMed: 31504823]

(15). Lu XM; Brickson CR; Murphy RM TANGO-Inspired Design of Anti-Amyloid Cyclic Peptides. Acs Chemical Neuroscience 2016, 7 (9), 1264–1274, Article. DOI: 10.1021/acschemneuro.6b00150. [PubMed: 27347598]

(16). Emily M; Talvas A; Delamarche C MetAmyl: A METa-Predictor for AMYLoid Proteins. PLoS One 2013, 8 (11), 9, Article. DOI: 10.1371/journal.pone.0079722.

(17). Walsh I; Seno F; Tosatto SCE; Trovato A PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res. 2014, 42 (W1), W301–W307, Article. DOI: 10.1093/nar/gku399. [PubMed: 24848016]

(18). Thangakani AM; Kumar S; Nagarajan R; Velmurugan D; Gromiha MM GAP: towards almost 100 percent prediction for beta-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics 2014, 30 (14), 1983–1990, Article. DOI: 10.1093/bioinformatics/btu167. [PubMed: 24681906]

(19). 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), 13, Article. DOI: 10.1038/s41467-020-17207-3. [PubMed: 31911625]

(20). Young T; Hazarika D; Poria S; Cambria E Recent Trends in Deep Learning Based Natural Language Processing. IEEE Comput. Intell. Mag. 2018, 13 (3), 55–75, Review. DOI: 10.1109/mci.2018.2840738.

(21). Elnaggar A; Heinzinger M; Dallago C; Rehawi G; Wang Y; Jones L; Gibbs T; Feher T; Angerer C; Steinegger M; Bhowmik D; Rost B ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44 (10), 7112–7127, Article. DOI: 10.1109/tpami.2021.3095381. [PubMed: 34232869]

(22). Suzek BE; Huang HZ; McGarvey P; Mazumder R; Wu CH UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007, 23 (10), 1282–1288, Article. DOI: 10.1093/bioinformatics/btm098. [PubMed: 17379688]

(23). Vaswani. Attention is all you need. Adv Neural Inf Process Syst 2017, 5998–6008.

(24). Schwenker F; Trentin E Pattern classification and clustering: A review of partially supervised learning approaches. Pattern Recognit. Lett. 2014, 37, 4–14, Review. DOI: 10.1016/j.patrec.2013.10.017.

(25). Golbraikh A; Tropsha A Beware of q(2)! Journal of Molecular Graphics & Modelling 2002, 20 (4), 269–276, Article. DOI: 10.1016/s1093-3263(01)00123-1. [PubMed: 11858635]

(26). McKenna A; Dubey S Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors. J. Biomed. Inform. 2022, 128, 17, Article. DOI: 10.1016/j.jbi.2022.104016.

(27). Rocklin M Dask: Parallel Computation with Blocked algorithms and Task Scheduling. In Proceedings of the 14th Python in Science Conference, 2015; SciPy: pp 130–136.

(28). Chollet F Keras Year: 2015 URL: https://keras.io.

(29). Abadi M; Agarwal A; Barham P; Brevdo E; Chen Z; Citro C; Corrado GS; Davis A; Dean J; Devin M; Ghemawat S; Goodfellow I; Harp A; Irving G; Isard M; Jia Y; Jozefowicz R; Kaiser L; Kudlur M; Levenberg J; Mane D; Monga R; Moore S; Murray D; Olah C; Schuster M; Shlens J; Steiner B; Sutskever I; Talwar K; Tucker P; Vanhoucke V; Vasudevan V; Viegas F; Vinyals O; Warden P; Wattenberg M; Wicke M; Yu Y; Zheng X TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015.

(30). Srinivas P; Katarya R hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. Biomed. Signal Process. Control 2022, 73, 10, Article. DOI: 10.1016/j.bspc.2021.103456.

(31). Lloyd SP LEAST-SQUARES QUANTIZATION IN PCM. IEEE Trans. Inf. Theory 1982, 28 (2), 129–137, Article. DOI: 10.1109/tit.1982.1056489.

(32). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V; Vanderplas J; Passos A; Cournapeau D; Brucher M; Perrot M; Duchesnay E Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830, Article.

(33). Davies DL; Bouldin DW A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. 1979, PAMI-1 (2), 224–227.

(34). Becht E; McInnes L; Healy J; Dutertre CA; Kwok IWH; Ng LG; Ginhoux F; Newell EW Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. 2019, 37 (1), 38–+, Article. DOI: 10.1038/nbt.4314.

(35). Hunter JD Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 2007, 9 (3), 90–95, Editorial Material. DOI: 10.1109/mcse.2007.55.

(36). Wolf T; Debut L; Sanh V; Chaumond J; Delangue C; Moi A; Cistac P; Rault T; Louf R; Funtowicz M; Davison J; Shleifer S; von Platen P; Ma C; Jernite Y; Plu J; Xu C; Le Scao T; Gugger S; Drame M; Lhoest Q; Rush A Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.

(37). Rajaraman S; Ganesan P; Antani S Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One 2022, 17 (1), 23, Article. DOI: 10.1371/journal.pone.0262838.

(38). Pancoe SX; Wang YJ; Shimogawa M; Perez RM; Giannakoulias S; Petersson EJ Effects of Mutations and Post-Translational Modifications on α-Synuclein In Vitro Aggregation. J. Mol. Biol. 2022, 434 (23), 167859. DOI: 10.1016/j.jmb.2022.167859. [PubMed: 36270580]

(39). Newberry RW; Leong JT; Chow ED; Kampmann M; DeGrado WF Deep mutational scanning reveals the structural basis for α-synuclein activity. Nature Chemical Biology 2020, 16 (6), 653–659. DOI: 10.1038/s41589-020-0480-6. [PubMed: 32152544]

(40). Shringari SR; Giannakoulias S; Ferrie JJ; Petersson EJ Rosetta custom score functions accurately predict ΔΔG of mutations at protein–protein interfaces using machine learning. Chem. Commun. 2020, 56 (50), 6774–6777, 10.1039/D0CC01959C. DOI: 10.1039/D0CC01959C.

(41). Giannakoulias S; Shringari SR; Ferrie JJ; Petersson EJ Biomolecular simulation based machine learning models accurately predict sites of tolerability to the unnatural amino acid acridonylalanine. Scientific Reports 2021, 11 (1), 18406. DOI: URL: 10.1038/s41598-021-97965-2. [PubMed: 34526629]

(42). Giannakoulias S; Shringari SR; Liu C; Phan HAT; Barrett TM; Ferrie JJ; Petersson EJ Rosetta Machine Learning Models Accurately Classify Positional Effects of Thioamides on Proteolysis. The Journal of Physical Chemistry B 2020, 124 (37), 8032–8041. DOI: URL: 10.1021/acs.jpcb.0c05981. [PubMed: 32869996]

(43). Ferrie JJ; Lengyel-Zhand Z; Janssen B; Lougee MG; Giannakoulias S; Hsieh C-J; Pagar VV; Weng C-C; Xu H; Graham TJA; Lee VMY; Mach RH; Petersson EJ Identification of a nanomolar affinity α-synuclein fibril imaging probe by ultra-high throughput in silico screening. Chemical Science 2020. DOI: URL: 10.1039/D0SC02159H.

(44). Phan HAT; Giannakoulias SG; Barrett TM; Liu C; Petersson EJ Rational design of thioamide peptides as selective inhibitors of cysteine protease cathepsin L. Chemical Science 2021, 12 (32), 10825–10835, 10.1039/D1SC00785H. DOI: URL: 10.1039/D1SC00785H. [PubMed: 35355937]

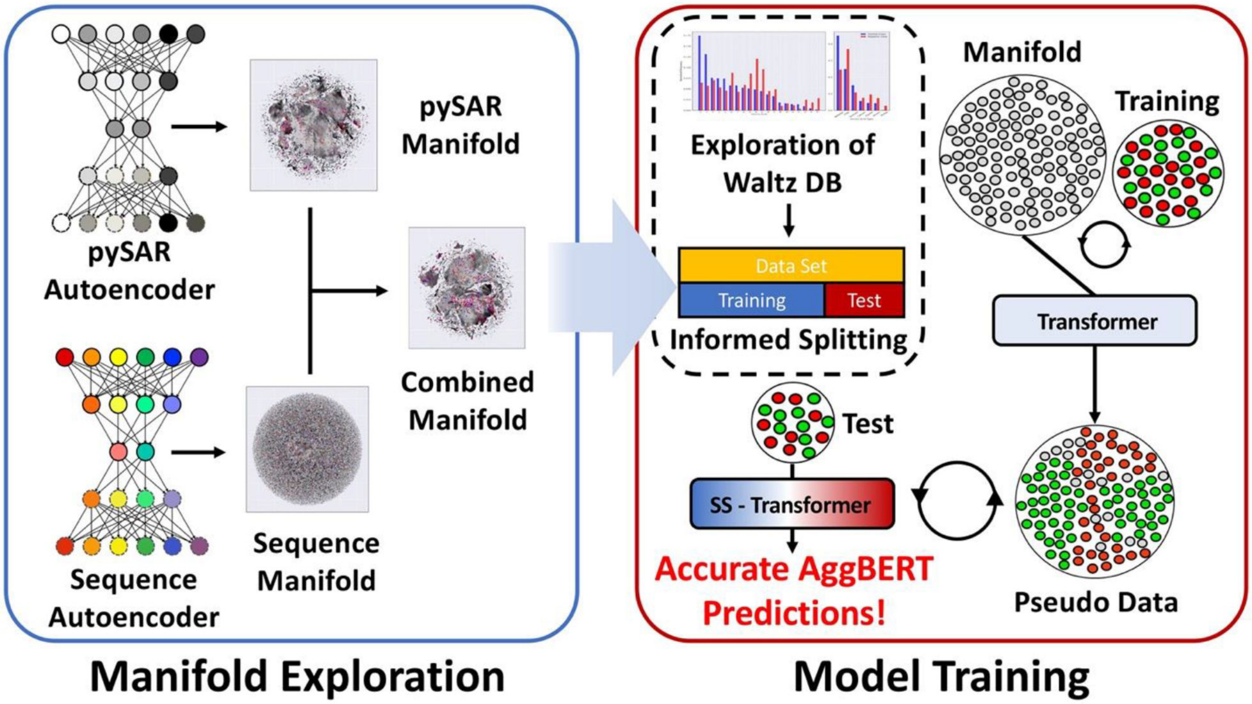


Figure 1.

Workflow schematic for the development of AggBERT, a semi-supervised ProtBERT-based model for predicting peptide amyloidogenicity.

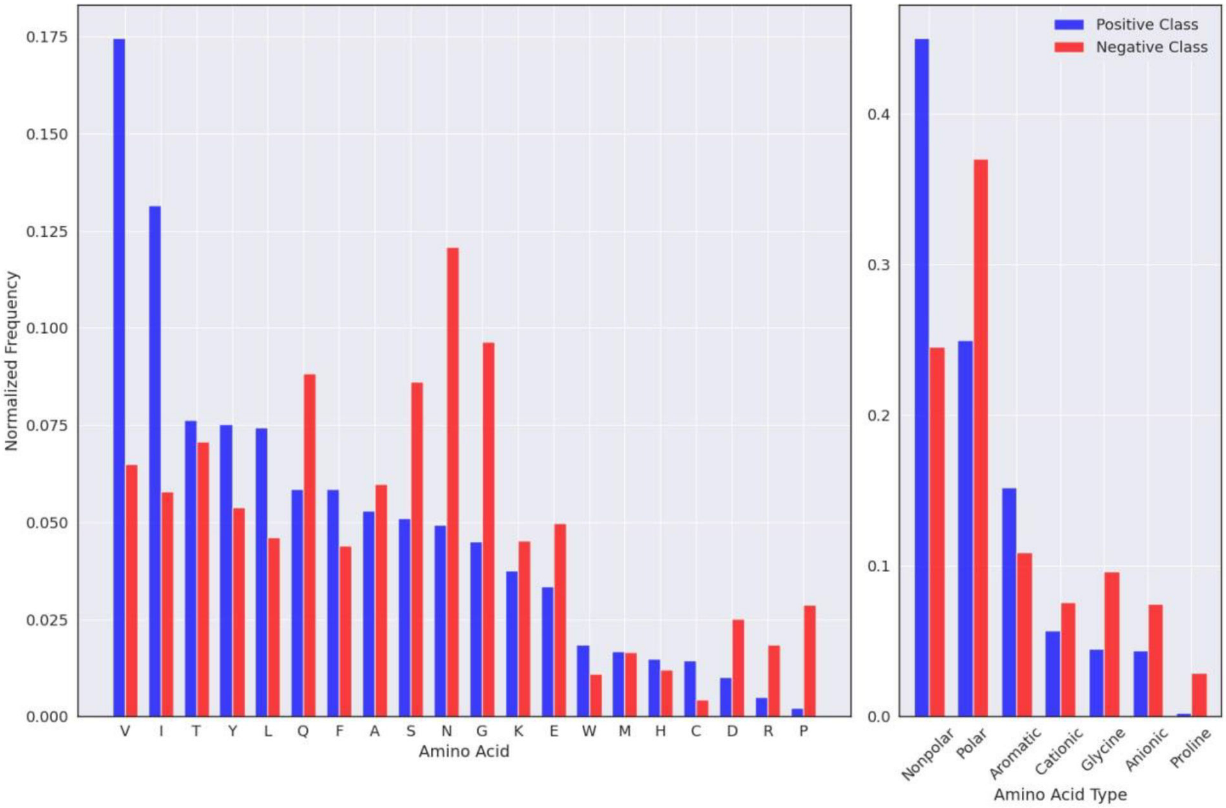


Figure 2.

Sequence composition of the WALTZ DB 2.0. The left panel displays amino acid frequency in the positive (aggregating and in blue) vs. negative (nonaggregating and in red) class. The right panel displays the frequency of amino acid types. Specific amino acid groupings are listed Table S1.

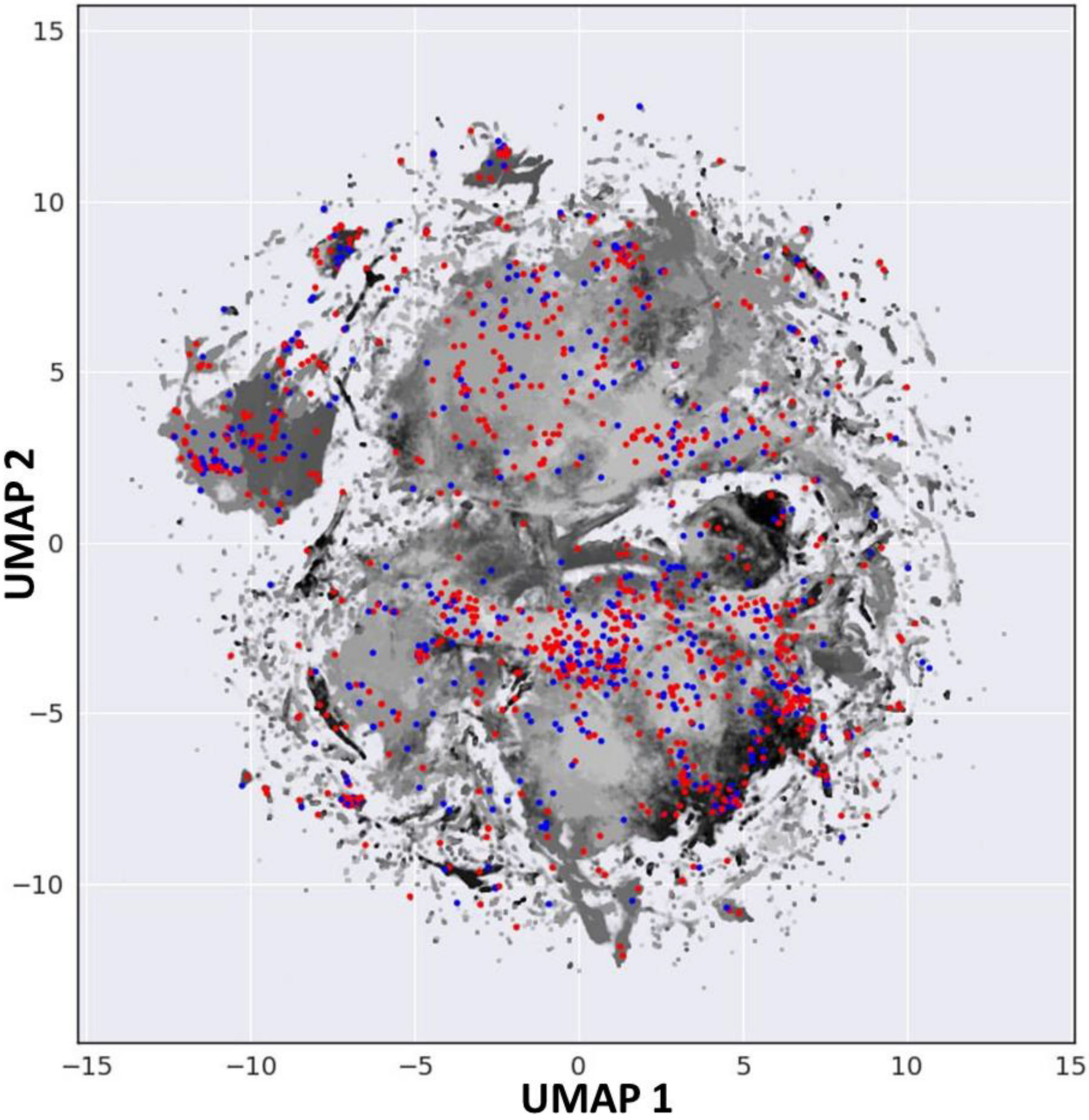


Figure 3.

Visualization of the 6-mer peptide space manifold using a combination of one-hot encoded sequence and pySAR features. The unlabeled data are shown in greyscale while the Waltz DB 2.0 points are displayed in red and blue according to their class label.

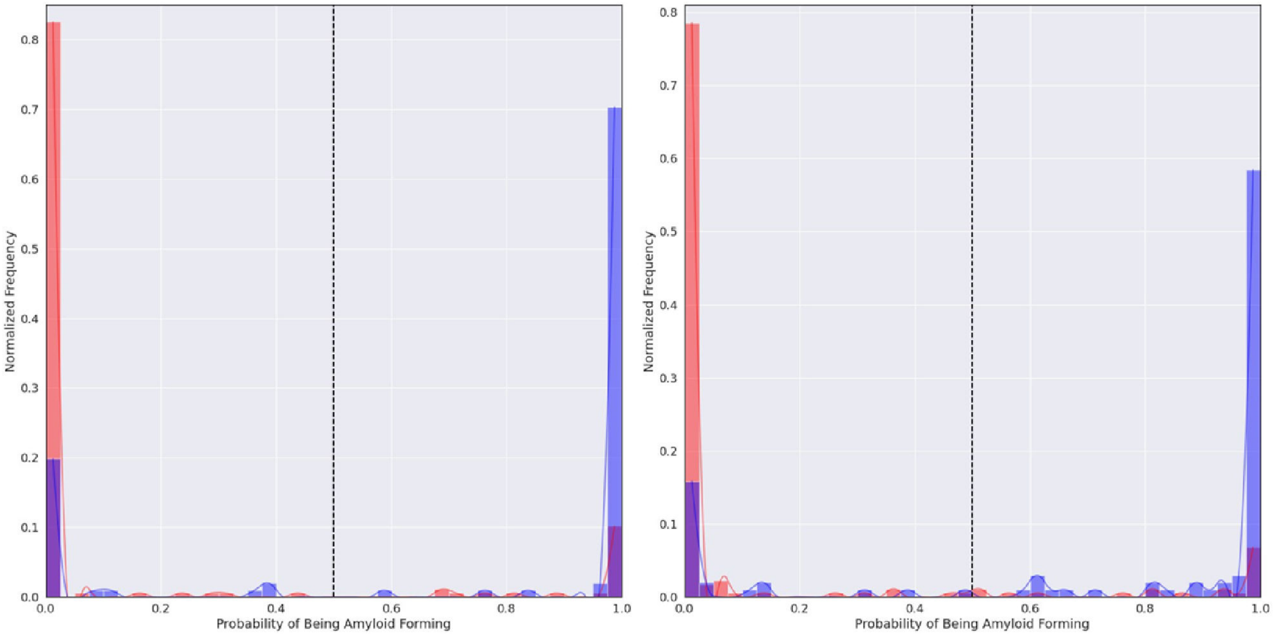


Figure 4.

Histograms representing the probabilistic confidence versus normalized frequency for the supervised transformer (Left) and AggBERT semi-supervised transformer (Right). Bars colored blue are pro-amyloidogenic, while those colored red are non-amyloidogenic. The vertical dashed line represents the probabilistic threshold of 0.5 which separates the classification of a sequence as pro-amyloid or not.

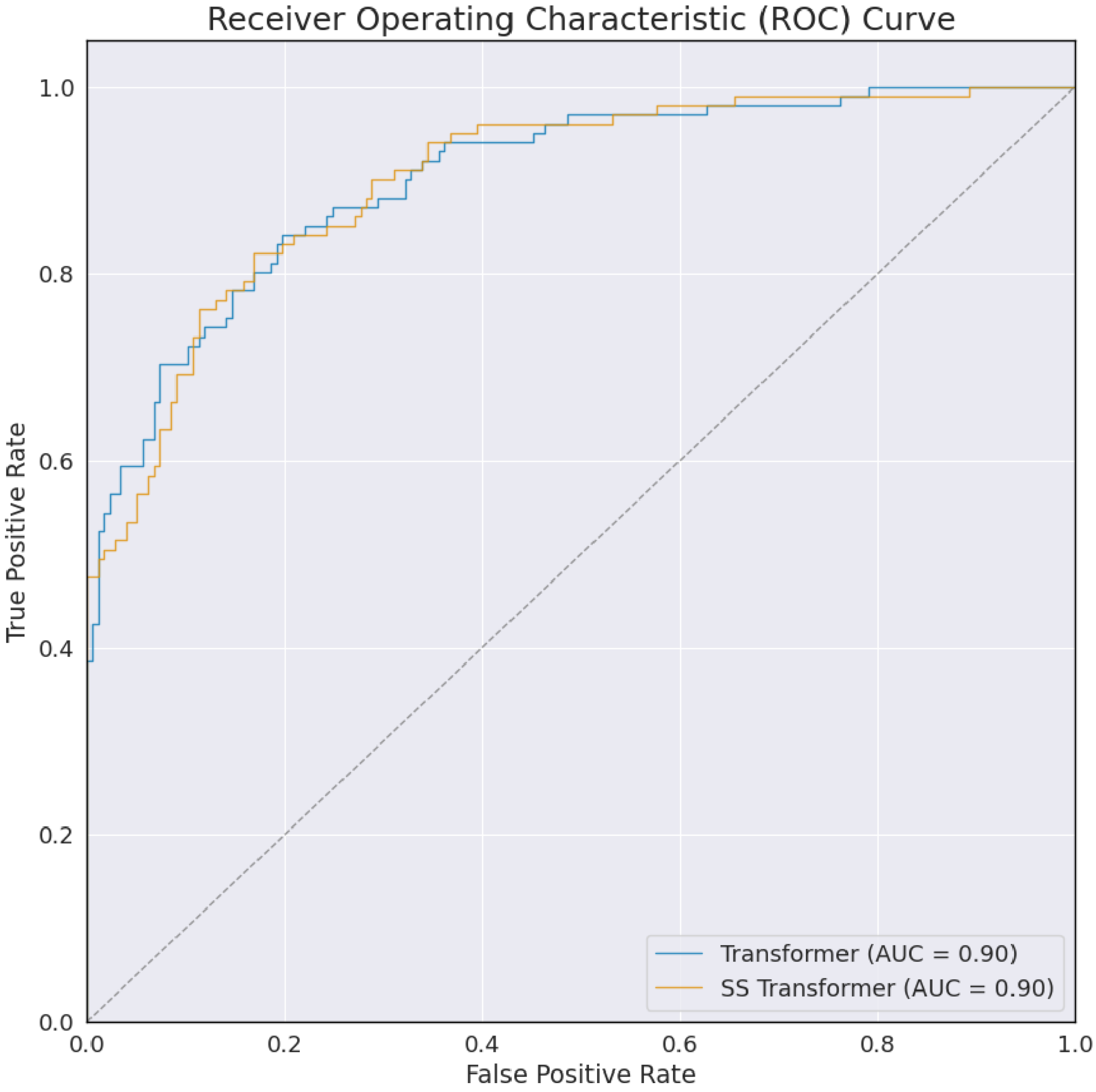


Figure 5.

Receiver Operating Characteristic curves for Waltz DB 2.0 models for the supervised transformer (Transformer) and AggBERT semi-supervised transformer (SS-Transformer). The dashed line in the center represents a random classifier.

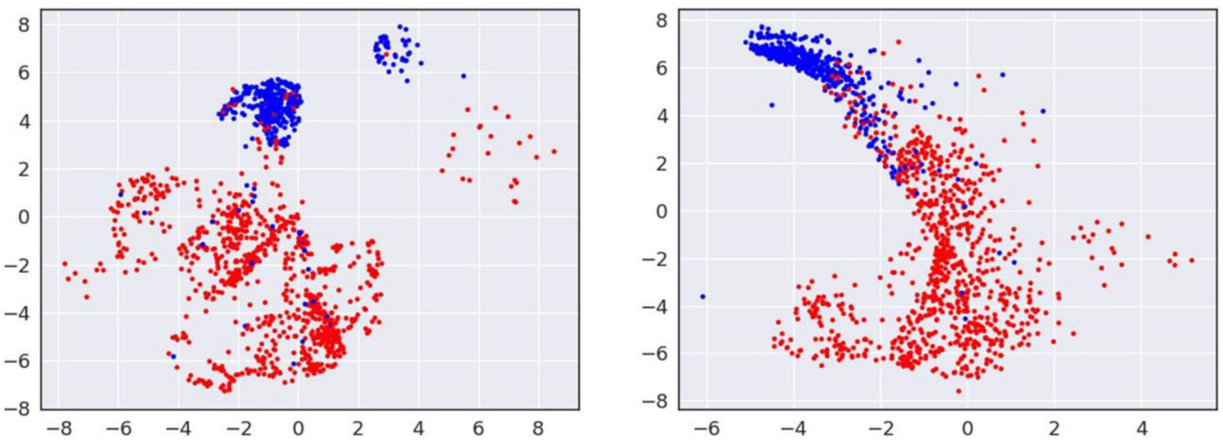


Figure 6.

UMAP visualizations of the model embeddings of the Waltz DB 2.0 datapoints from the fine-tuned ProtBERT model (Left) and the semi-supervised ProtBERT model (AggBERT, Right). The points colored in blue display the pro-amyloidogenic sequences while the red points display the non-amyloidogenic sequences.

Table 1.

Classification metrics from Leave-One-Out cross validation of literature models vs the methods trained in this study (bolded). Models are ordered by F1 score.

Algorithm	Accuracy	Precision	Recall	FPR	MCC	F1	AUC
AggBERT: Semi-supervised transformer	0.83	0.78	0.74	0.12	0.63	0.76	0.90
Supervised transformer	0.83	0.78	0.72	0.11	0.63	0.75	0.90
Cordax[19]	0.81	0.74	0.72	0.14	0.57	0.73	0.87
MetAmyl[16]	0.79	0.72	0.69	0.15	0.54	0.70	0.78
Pasta2.0[17]	0.76	0.74	0.52	0.10	0.46	0.61	0.84
Tango[15]	0.73	0.82	0.31	0.04	0.38	0.45	0.64
Waltz[14]	0.67	0.54	0.65	0.33	0.34	0.60	0.73
AGGRESCAN[11]	0.57	0.60	0.85	0.54	0.29	0.71	0.84
Dummy	0.56	0.40	0.38	0.33	0.04	0.38	0.52
GAP[18]	0.51	0.42	0.94	0.74	0.25	0.58	0.70