Gated recurrent units classify amyloidogenic proteins from sequence alone

Eric Jing Mockler

Department of Biomolecular Engineering
University of California, Santa Cruz

July 1, 2021

Abstract

The evolution of amyloidogenic proteins narrowly balances function with disorder. Amyloid fibrils confer diverse utility across domains of life— curli fibers secreted by enterobacteria such as *Escherichia coli* scaffold cell adhesion and biofilm formation (1, 2), while silk fibers spun by *Bombyx mori* protect the silkworm during metamorphosis. Both proteins aggregate in the spleens of mice upon intravenous injection; a pathology characteristic of amyloidosis (3). Evidence of heterologous proteins cross-seeding amyloid fibrillization (2–5) outlines the need to investigate environmental factors that may contribute to amyloidoses. Here, an artificial neural network model composed of gated recurrent units (GRU) is introduced to the amyloid classification task, rivalling current methods with 95% specificity, 81% sensitivity and 87% accuracy, after 10-fold cross-validation; performance is retained upon external validation. The feature extraction process used to train the model is minimal, emphasizing a data-driven paradigm to mitigate biases in feature selection. Performance, model architectures, and feature sets are compared against other amyloid prediction approaches. The model is applied to predict 310,485 amyloidogenic proteins within the Gut Phage Database, a novel collection of $\sim$7.5 million bacteriophage proteins across 142,809 species (6).

Contents

Introduction

In 1854, Rudolph Virchow first characterized the term "amyloid" in medical literature by describing insoluble, macroscopic tissue abnormalities that stained positive in starch-iodine assays (7–9). The nature of the assay led Virchow to assume these proteins were identical to starch, and named these aggregates *corpora amylacea*— a term rooted in *amylum*, the Latin word for "starch" (8, 9). In 1959, electron microscopy found that amyloids were morphologically diverse proteins, composed of fibrils ranging from 70 to 140 Å in width, and up to 16,000 Å in length (10). X-ray diffraction studies during the 1960s revealed that amyloid fibrils are composed of common cross–β motifs (11), a secondary structure derived from β–sheet proteins. Throughout the 2000s, conformational and thermodynamic studies of amyloid fibrillation found intrinsically disordered proteins and partially unfolded globular proteins driving aggregation (12–17). These proteins leverage mutations, overexpression of amyloid precursors, and free energy disruptions to seed amyloidogenesis. Aggregates remain partially unfolded, yet highly-ordered and extremely stable, arranging into cross–β structures characteristic of amyloid fibrils (12, 13).

Amyloids have long been associated with disease, but many functional roles were outlined in recent years; particularly in bacteria, fungi, and humans (1, 2, 18). In a microbiome environment, functional amyloids may leak from the gut and cross-seed with the host, driving neurodegeneration (16, 17, 19, 20), type 2 diabetes (17), and inflammatory disorder (19). The microbiome (17, 19, 21–23) and central nervous system (24) have established roles in modulating intestinal permeability, and are probable effectors of cross-seeding. Long-term transfer of healthy gut microbiota via faecal transplantation ameliorated Aβ deposition, tau pathology, reactive gliosis and memory impairment in a murine model (25), though etiology is intermingled with the downstream consequences of systemic inflammation. More evidence of cross-seeding via the intestinal barrier is needed to validate this hypothesis (16, 17, 26).

A nucleated growth mechanism drives amyloid aggregation through sigmoidal kinetics, composed of three stages; a lag phase, growth phase, and final plateau (13, 27). This classical nucleation theory offers consensus on macroscopic kinetics, but several models postulate different physicochemical mechanisms for self-catalyzed fibrillization (27–30). Amyloids lie on a spectrum of virulence; while all aggregates may be transmissible (31), readily infectious aggregates, such as those found in transmissible spongiform encephalopathies, are sub-classified as prions (32).

Chapter 1

Finding amyloids in the proteome

1.1 Published methods

Amino acid composition strongly indicates amyloidogenic propensity (33–36). Sequence-based prediction is a productive approach; many algorithms trained upon abstract physicochemical features classify amyloids with useful accuracy, but have notable variances in specificity and sensitivity. Several of these are focused on prion classification, such as PAPA (37), pWALTZ (28), PLAAC (38), prionW (39), and pRANK (33). These classifiers are biased towards glutamine/asparagine-enriched prion peptides within yeast, impairing validity in human disease contexts (33, 38). Recent approaches are generalized towards amyloid classification, and will be the focus of comparison in this study.

Various amyloids such as islet amyloid polypeptide (40), silkworm chorion protein (41) and *Sup35* yeast prion (42) depend on small interchangeable motifs to aggregate at a significant rate, ranging from five to seven amino acids in length (40–42). Because of this observation, some amyloid classifiers limit feature extraction to hexapeptides (35, 36, 43, 44). It is important to note that local motifs are not the sole drivers of amyloidogenesis. Global sequence specificity is directly proportional to the amyloid fibrillization rate of hen lysozyme, emphasizing that long-range interactions are relevant (34). Moreover, gatekeeper residues often flank amyloidogenic domains and neutralize electric charge (45, 46) or bind allosteric factors (47), modulating aggregation rate and further defining the spectrum of amyloidogenesis. More comprehensive models can be developed by leveraging elusive features in the protein sequence. The three most performant algorithms to-date, RFAmyloid (48), PredAmyl-NLP (49), and iAMY-SCM (50), extract distal features within the greater protein sequence and will be compared against the current model.

RFAmyloid was developed in 2018, and utilizes random forest architecture (48) based upon decision trees to classify amyloid. Two feature extraction methods are applied; SVMProt-188D extracts amino acid frequencies across the first twenty dimensions, while eight phyisicochemical features fill the remaining 168 dimensions: hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary struc-

ture, and solvent accessibility (51). The 20 amino acids are placed into one of three abstract categories, on a per-feature basis. Three within-category descriptors are applied to each physicochemical feature— composition (C) captures amino acid frequency, transition (T) notes the frequency of changes to physicochemically different amino acids in the sequence, and distribution (D) describes emergent physicochemical properties with respect to the first 25, 50, and 75% of amino acids in the sequence, normalized by peptide length. The complete feature space is described in Equation 1.1 (51).

$$20\ amino\ acids\ +(C+T+D)\times 8\ chemical\ properties = 188\ dimensions. \tag{1.1}$$

Pse-in-one recovers a broader sequence profile by embedding positional features via pseudo components, which represent a set of feature vectors embodying the composition of amino acids within a peptide (48, 52). Short-range pseudo components are calculated using the occurrence frequencies of k nearest residues with respect to a single amino acid, while long-range psuedo components capture autocorrelation across di– and tri–peptides within the greater sequence (52).

The only shared feature between RFAmyloid and the current model is amino acid composition. Despite this, sensitivity, specificity and accuracy are highly comparable, which demonstrates that a neural network architecture can learn the amyloid classification problem without imperatively assigning a feature set. A performance trade-off is observed with the current model; RFAmyloid has over 15% lesser specificity, while sensitivity is 10% greater. Table 1.1 compares these predictors across the amyloid classifiers discussed here. The RFAmyloid training dataset filters out peptides shorter than 50 amino acids, which is a notable departure from previous methods trained exclusively on hexapeptides. The current model does not apply a minimum length restriction; the shortest amyloidogenic protein sequence in training data, physalaemin, is 11 amino acids long.

The group behind RFAmyloid introduced another tool in 2020, PredAmyl-MLP, marking an iteration upon their previous work (49). This tool applies a multilayer-perceptron neural network architecture trained on the same dataset used for RFAmyloid, now filtered using CD-HIT (53) to cluster highly similar sequences; emphasizing canonical features and reducing noise (51, 53). As in RFAmyloid, SVM-Prot188-D is applied to extract physicochemical and higher-order compositional features across the protein into 188 dimensions, but features are also filtered using the maximum relevant maximum distance method (49, 54). Psuedo components are more focused, only utilizing tri–peptide composition, and the binomial distribution method (55) is applied to find optimal tri–peptides. Model sensitivity and specificity increased by 3-5% compared to RFAmyloid, closely competing with the current model.

iAMY-SCM is the most recent amyloid classification algorithm to emerge in 2021, applying a novel score card method to classify amyloid (50). Di–peptide composition is extracted and normalized from the protein sequence, lending a 400-dimensional feature set for training (50). Di–peptide scores range from 0 to 1000, with 1000 indicating residues critical to amy-

loidogenic proteins. Notably, this method exposes specific amyloidogenic regions within a protein via dipeptide arrangements, lending more interpretability than a single classification produced by RFAmyloid (48) and the current model. Tables 1.1 and 1.2 compare accuracy, sensitivity, specificity and Matthews correlation across these models during cross-validation and external validation, respectively.

Model	Accuracy	Specificity	Sensitivity	MCC
GRU	0.869	**0.952**	0.812	0.751
RFAmyloid	0.892	0.781	0.927	0.739
PredAmyl-MLP	**0.916**	0.836	0.950	**0.798**
iAMY-SCM	0.895	0.757	**0.954**	0.750

Table 1.1: Comparison of accuracy, sensitivity, specificity, and Matthews correlation coefficients between models, on 10-fold cross-validation (48–50). Heuristics for the best model are in bold. "NA" abbreviates "not available", indicating that the heuristic was not published. "MCC" abbreviates "Matthews correlation coefficient".

Model	Accuracy	Specificity	Sensitivity	MCC
GRU	0.889	**1.00**	0.818	**0.797**
RFAmyloid	**0.897**	0.818	**0.932**	0.757
PredAmyl-MLP	NA	NA	NA	NA
iAMY-SCM	0.827	0.606	0.922	0.570

Table 1.2: Comparison of accuracy, sensitivity, specificity, and Matthews correlation coefficients between models, on external validation (48–50). Heuristics for the best model are in bold. "NA" abbreviates "not available", indicating that the heuristic was not published. "MCC" abbreviates "Matthews correlation coefficient".

1.2 Model architecture & dataset collection

The current study is the first known approach using a recurrent neural network to classify amyloids using protein sequence alone. A stacked bi-directional, gated recurrent unit (GRU) architecture was applied to find amyloidogenic likelihood, by traversing tokenized amino acids of a given protein. Two stacked GRUs increase non-linearity, enabling deeper representations of abstract features in the protein sequence. A bi-directional recurrent neural network highlights features dependent on protein sequence direction with respect to N– and C– termini, and reliably predicts secondary structure prediction concerning long-range motifs (56).

Two pre-existing amyloid and prion datasets were constructed into 180 positive examples, from the AmyPro amyloid database (57) and the set of yeast prions used to train PLAAC (38). Protein sequences obtained from AmyPro are limited to broad amyloid-forming regions, as defined in literature (57). The negative dataset was constructed with protein sequences not

known to be amyloidogenic, from UniprotKB (58). A balanced proportion of peptides was collected from each species in the positive dataset, totalling 180 negative examples. The negative dataset has not been validated for amyloidogenic activity, so contamination may be present.

1.3 Feature extraction & training

A minimal set of features was extracted from protein sequences. Proteins were tokenized by amino acid into a predetermined, 1-hot-like embedding layer of 64 dimensions, for each of the 20 amino acids and a wildcard character. No inferences were made upon the protein sequence; all abstract features were left for the model to learn, in favor of a data-driven approach. This is in stark contrast to existing methods, which extract physicochemical properties, position-specific information, and 3-D structural features. Performance compares favorably against these methods using the data-driven paradigm. The model is penalized using smooth L_1-loss, illustrated in Equation 1.2.

$$L_{1;smooth} = \begin{cases} |x| & \text{if } |x| > \alpha \\ \frac{1}{|\alpha|}x^2 & \text{if } |x| \leq \alpha \end{cases} \tag{1.2}$$

The smooth L_1-loss function utilizes hyperparameter α to determine if the output is large or small. This loss function maintains steady gradients over large and small outputs by squaring the loss at a certain threshold, mitigating the effect of outliers in the dataset.

A 70:30 split was applied to the full dataset; 15% of the data is kept within an independent set for external validation. The model is cross-validated 10-fold during training. Regularization is applied via dropout layers of 30% probability in-between the stacked directional GRUs, and at the final layer as outputs are concatenated across both directions. Model architecture is illustrated in Supplemental Data 1. The Pytorch (59) and SKLearn (60) libraries were applied during implementation, outlined in Supplemental Data 2 as an IPython notebook. Supplemental Data 3 provides a confusion matrix, receiver operating characteristic curve, and loss plots. Tables 1.3 and 1.4 contextualize empirical performance upon 10-fold cross-validation and external validation, respectively.

Fold	Accuracy	Specificity	Sensitivity	MCC
1	0.869	0.952	0.812	0.751
2	0.853	0.909	0.810	0.712
3	0.867	0.938	0.814	0.742
4	0.866	0.931	0.818	0.740
5	0.856	0.943	0.798	0.727
6	0.849	0.950	0.786	0.717
7	0.869	0.952	0.812	0.751
8	0.846	0.949	0.782	0.712
9	0.866	0.938	0.814	0.742
10	0.860	0.882	0.839	0.720

Table 1.3: Comparison of accuracy, sensitivity, specificity, and Matthews correlation coefficient of the current model, on cross-validation. "NA" abbreviates "not available", indicating that the heuristic was not published. "MCC" abbreviates "Matthews correlation coefficient".

Fold	Accuracy	Specificity	Sensitivity	MCC
1	0.870	0.955	0.813	0.753
2	0.852	0.913	0.807	0.712
3	0.870	0.955	0.813	0.754
4	0.870	0.917	0.833	0.745
5	0.889	1.00	0.818	0.798
6	0.870	0.955	0.813	0.754
7	0.870	0.955	0.813	0.753
8	0.870	1.00	0.794	0.767
9	0.870	0.955	0.813	0.754
10	0.796	0.786	0.808	0.592

Table 1.4: Comparison of accuracy, sensitivity, specificity, and Matthew correlation coefficient of the current model, on external validation. "NA" abbreviates "not available", indicating that the heuristic was not published. "MCC" abbreviates "Matthews correlation coefficient".

Chapter 2

Amyloid in gut bacteriophages

2.1 Current evidence

Bacteriophages are the dominant viral constituents of the microbiome, and are concentrated in the gut (61–64). Most phage-derived peptides are not well-represented in reference databases (6). Discussion of bacteriophages in disease etiology has risen in literature, with the gut-brain axis being a prominent concern (63–65). Studies have observed the presence of bacteriophage communities within cerebrospinal fluid (CSF) donated by healthy individuals, challenging prior assumptions of CSF sterility in the absence of disease (66, 67).

Tetz and Tetz (68) marks the first in-depth investigation of amyloidogenic propensity in phagobiota. With a focus on the prion amyloid subclass, Tetz and Tetz (68) examined 370,617 bacteriophage protein sequences available on UniprotKB (58) for prion-like domains using PLAAC (38): a hidden–Markov model that profiles sequence similarity to yeast prions, via the maximum likelihood estimation (38, 68). PLAAC found 5040 putative prions with a log likelihood ratio greater than 0.003; representing 1.35% of phage peptide sequences available on UniprotKB. Tetz and Tetz (68) characterized cell attachment and DNA injection as the most abundant functional groups of proteins matching these amyloid predictions. Several pathogenic amyloid proteins in humans readily interact with nucleic acids (69–71), indicating the possibility of cross-seeding.

Interestingly, filamentous phage capsid protein g3p binds Aβ in a manner inducing disaggregation, through a general amyloid–interaction motif (GAIM) dependent on hydrophobic interactions (72). In a murine model, intraperitoneal injection of a fusion protein containing this GAIM reduced Aβ deposition, phospho-tau pathology, atrophy, and improved cognition in murine models of Alzheimer's (73). While the g3p protein disrupts amyloidogenesis, it stands to reason that other phage peptides have a converse effect on amyloid aggregation, possibly those sharing sequence profile similarity with yeast prions (68).

Camarillo-Guerrero et al. (6) presents the Gut Phage Database (GPD), a massive collection of 7,581,807 phage-derived proteins derived from human gut microbiota, ranging across 142,809 bacteriophage species; of which the majority are uncharacterized to-date. Less than

1% of canonical sequences overlap with bacteriophages represented in the NCBI RefSeq database (6, 74).

2.2 Results

This study iterates upon the prior work of Tetz and Tetz (68) by applying a highly-specific model to classify aggregation-prone proteins, within the greater amyloid class. The current model predicts 310,947 phage proteins have amyloidogenic propensity, totalling 4.1% of the GPD.

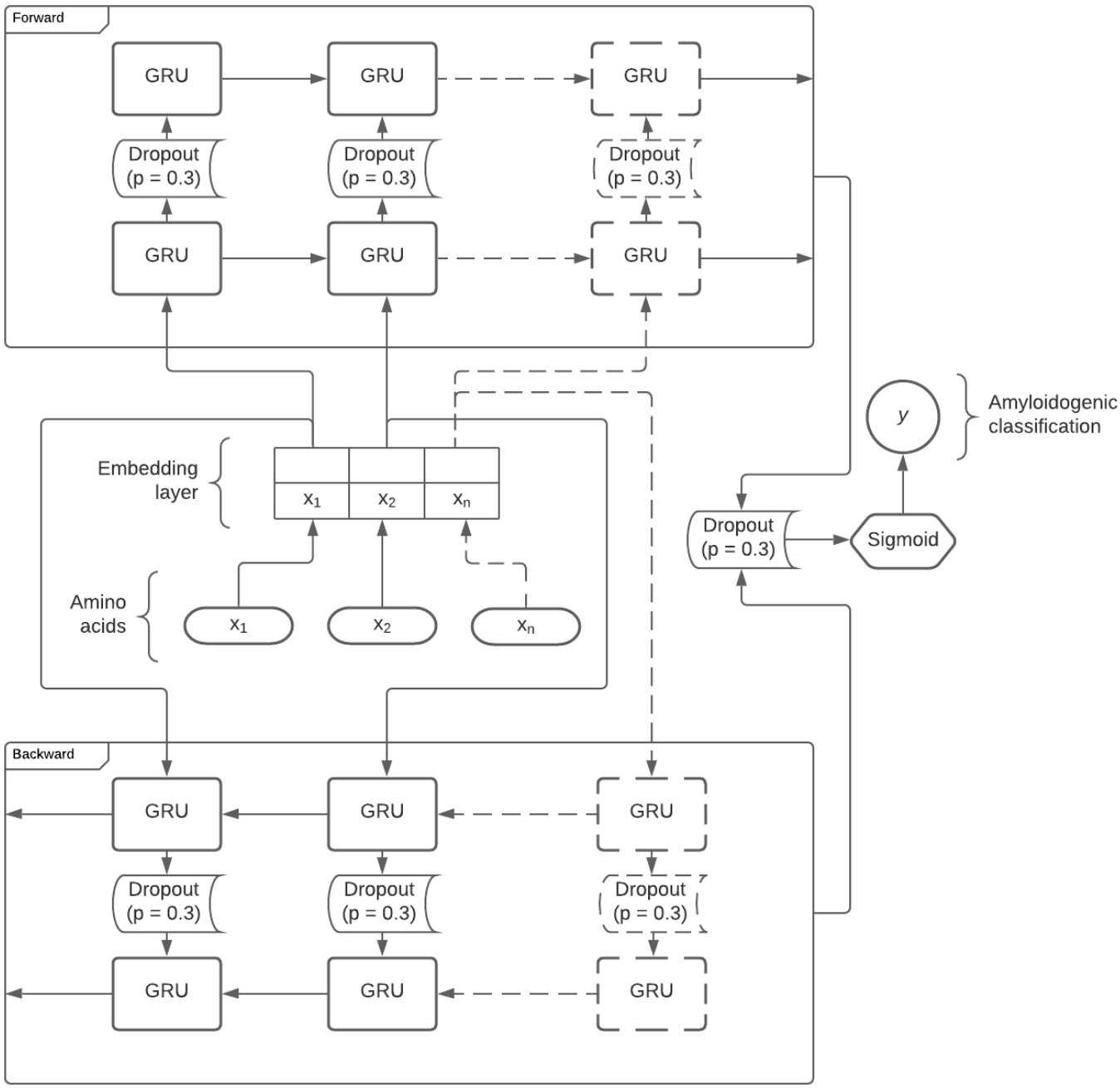
Chapter 3

Conclusions and Future Work

A gated recurrent unit architecture was introduced to classify amyloid protein sequences with high specificity, offering competitive results despite minimal feature extraction. Inclusion of higher-order features, such as tri–peptide composition and physicochemical properties, may close the gap in sensitivity with existing amyloid classifiers. These features can also improve biological interpretability by outlining regions-of-interest within the predicted amyloid sequence.

The current model was applied to predict novel amyloidogenic proteins in phagobiota; these predictions cover 4.10% of the Gut Phage Database. Analysis is ongoing to characterize the broader functions of these proteins, and possible etiologies in human disease.

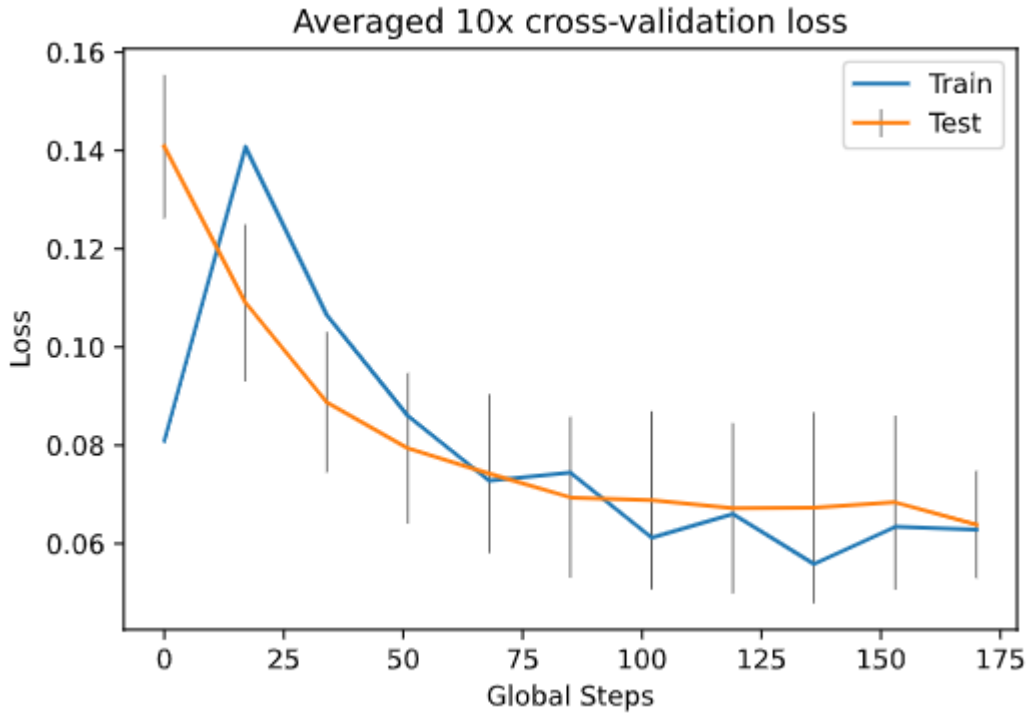
Supplemental Data 1

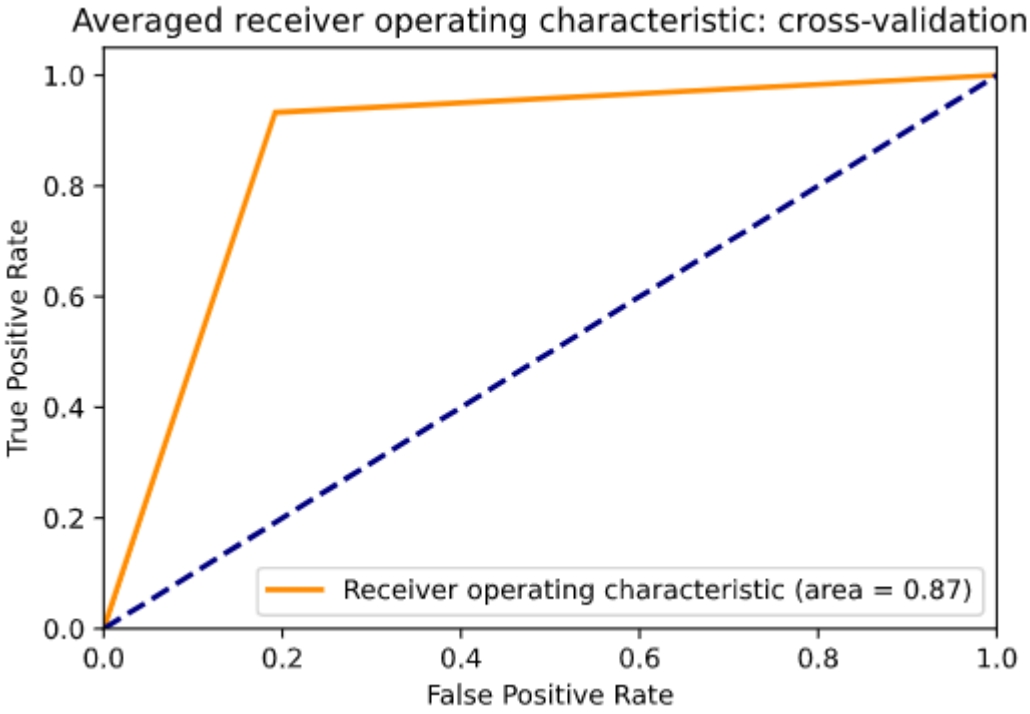


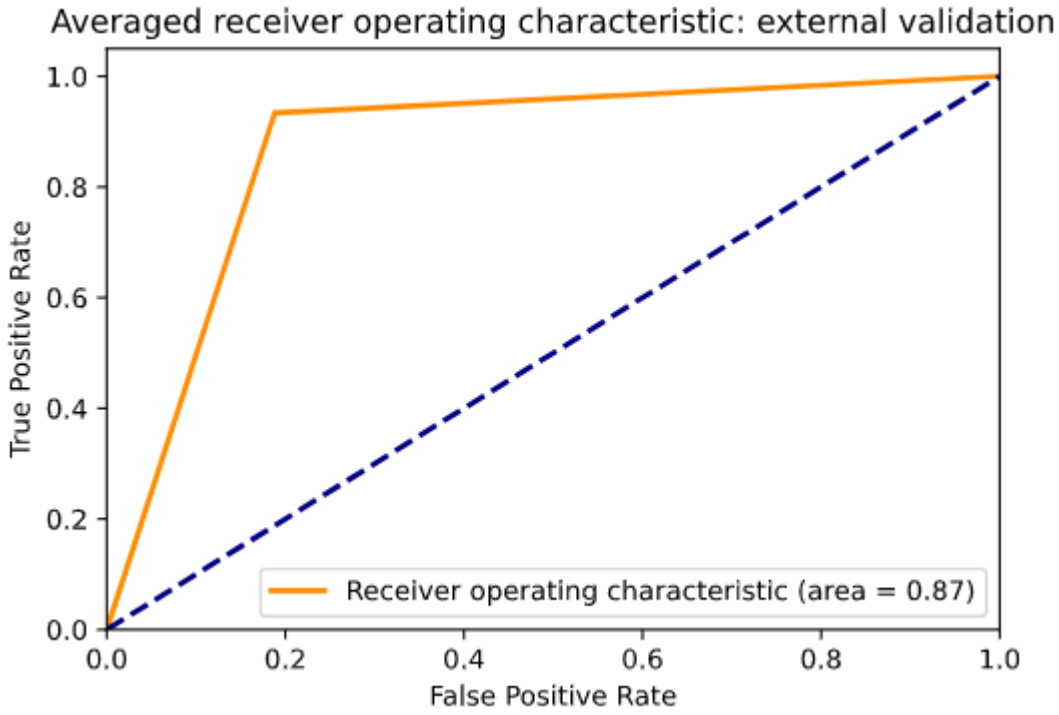
Supplemental Data 2

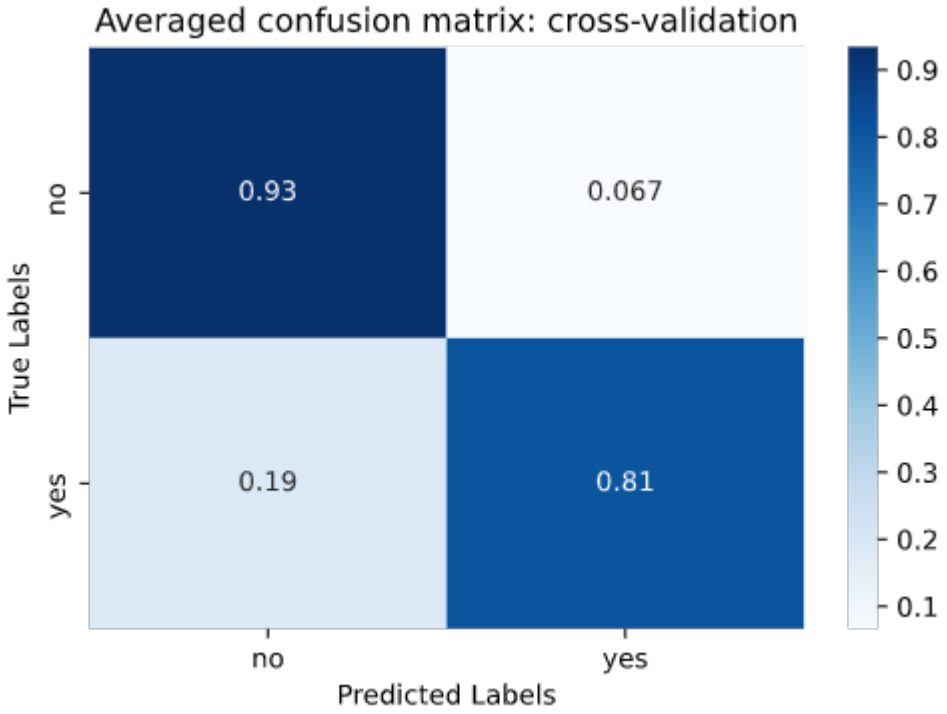
Output data and an IPython notebook are maintained at this GitHub repository:
https://github.com/mocklee/AgGRU-Check

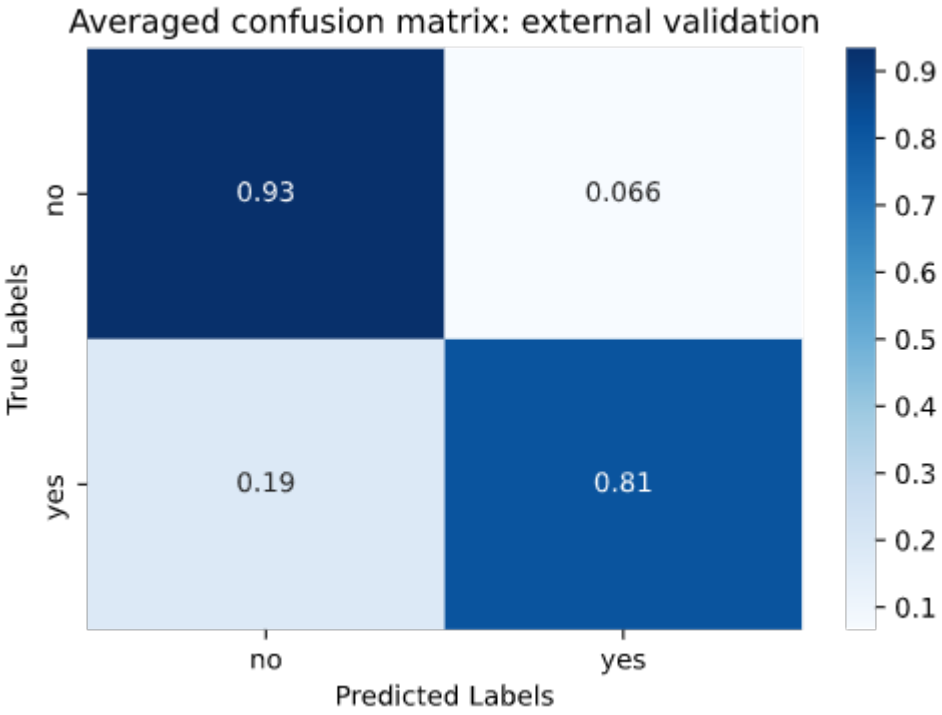
Supplemental Data 3






Averaged receiver operating characteristic: external validation
True Positive Rate
False Positive Rate
Receiver operating characteristic (area = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0


Averaged confusion matrix: cross-validation
True Labels
Predicted Labels
no
yes
0.93
0.067
0.19
0.81
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Averaged confusion matrix: external validation
0.93
0.066
0.19
0.81
no
yes
True Labels
no
yes
Predicted Labels
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Bibliography

[1] M. M. Barnhart and M. R. Chapman. Curli biogenesis and function. *Annual Review of Microbiology*, 60:131–147, 2006. URL https://doi.org/10.1146/annurev.micro.60.080805.142106. i, 1

[2] Y. Zhou, D. Smith, B. J. Leong, K. Brännström, F. Almqvist, and M. R. Chapman. Promiscuous cross-seeding between bacterial amyloids promotes interspecies biofilms. *The Journal of Biological Chemistry*, 287(42):35092–35103, 2012. URL https://doi.org/10.1074/jbc.M112.383737. i, 1

[3] K. Lundmark, G. T. Westermark, A. Olsén, and P. Westermark. Protein fibrils in nature can enhance amyloid protein a amyloidosis in mice: Cross-seeding as a disease mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17):6098–6102, 2005. URL https://doi.org/10.1073/pnas.0501814102. i

[4] R. Morales, I. Moreno-Gonzalez, and C. Soto. Cross-seeding of misfolded proteins: Implications for etiology and pathogenesis of protein misfolding diseases. *PLoS Pathogens*, 9:9, 2013. URL https://doi.org/10.1371/journal.ppat.1003537. i

[5] B. Ren, Y. Zhang, M. Zhang, Y. Liu, D. Zhang, X. Gong, Z. Feng, J. Tang, Y. Chang, and J. Zheng. Fundamentals of cross-seeding of amyloid proteins: An introduction. *Journal of Materials Chemistry B*, 7(46):7267–7282, 2019. URL https://doi.org/10.1039/C9TB01871A. i

[6] L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, and T. D. Lawley. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109, 2021. URL https://doi.org/10.1016/j.cell.2021.01.029. i, 7, 8

[7] J. D. Sipe and A. S. Cohen. Review: History of the amyloid fibril. *Journal of Structural Biology*, 130(2):88–98, 2000. URL https://doi.org/10.1006/jsbi.2000.4221. 1

[8] M. Tanskanen. *"Amyloid"— Historical Aspects.* Amyloidosis, 2013. URL https://doi.org/10.5772/53423. 1

[9] R. Virchow. "Uber eine in gehirn und r"uckenmark des menschen aufgefundene substanz mit der chemischen reaktion der cellulose. *Virchow's Archiv fur pathologische Anatomie und f"ur klinische Medicin*, 6:135–138, 1854. 1

[10] A. S. Cohen and E. Calkins. Electron microscopic observations on a fibrous component in amyloid of diverse origins. *Nature*, 183(4669):1202–1203, 1959. URL https://doi.org/10.1038/1831202a0. 1

[11] L. Bonar, A. S. Cohen, and M. M. Skinner. Characterization of the amyloid fibril as a cross-beta protein. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, 131(4):1373–1375, September 1969. ISSN 0037-9727. doi: 10.3181/00379727-131-34110. 1

[12] V. N. Uversky and A. L. Fink. Conformational constraints for amyloid fibrillation: The importance of being unfolded. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1698(2):131–153, 2004. URL https://doi.org/10.1016/j.bbapap.2003.12.008. 1

[13] J.-C. Rochet and P. T. Lansbury. Amyloid fibrillogenesis: Themes and variations. *Current Opinion in Structural Biology*, 10(1):60–68, 2000. URL https://doi.org/10.1016/S0959-440X(99)00049-4. 1

[14] F. Chiti and C. M. Dobson. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology*, 5(1):15–22, 2009. URL https://doi.org/10.1038/nchembio.131. 1

[15] Y. Kim, J. S. Wall, J. Meyer, C. Murphy, T. W. Randolph, M. C. Manning, A. Solomon, and J. F. Carpenter. Thermodynamic modulation of light chain amyloid fibril formation. *Journal of Biological Chemistry*, 275(3):1570–1574, 2000. URL https://doi.org/10.1074/jbc.275.3.1570. 1

[16] I. Javed, Z. Zhang, J. Adamcik, N. Andrikopoulos, Y. Li, D. E. Otzen, S. Lin, R. Mezzenga, T. P. Davis, F. Ding, and P. C. Ke. Accelerated amyloid beta pathogenesis by bacterial amyloid fapc. *Advanced Science*, 7(18):2001299, 2020. URL https://doi.org/10.1002/advs.202001299. 1

[17] M. I. Ivanova, Y. Lin, Y.-H. Lee, J. Zheng, and A. Ramamoorthy. Biophysical processes underlying cross-seeding in amyloid aggregation and implications in amyloid pathology. *Biophysical Chemistry*, 269(10650):7, 2021. URL https://doi.org/10.1016/j.bpc.2020.106507. 1

[18] D. M. Fowler, A. V. Koulov, W. E. Balch, and J. W. Kelly. Functional amyloid – from bacteria to humans. *Trends in Biochemical Sciences*, 32(5):217–224, 2007. URL https://doi.org/10.1016/j.tibs.2007.03.003. 1

[19] K. Schwartz and B. R. Boles. Microbial amyloids – functions and interactions within the host. *Current Opinion in Microbiology*, 16(1):93–99, 2013. URL https://doi.org/10.1016/j.mib.2012.12.001. 1

[20] Timothy R Sampson, Collin Challis, Neha Jain, Anastasiya Moiseyenko, Mark S Ladinsky, Gauri G Shastri, Taren Thron, Brittany D Needham, Istvan Horvath, Justine W Debelius, Stefan Janssen, Rob Knight, Pernilla Wittung-Stafshede, Viviana Gradinaru, Matthew Chapman, and Sarkis K Mazmanian. A gut bacterial amyloid promotes -synuclein aggregation and motor impairment in mice. *eLife*, 9:e53111, February 2020. ISSN 2050-084X. doi: 10.7554/eLife.53111. URL https://doi.org/10.7554/eLife.53111. Publisher: eLife Sciences Publications, Ltd. 1

[21] R. P. Friedland and M. R. Chapman. The role of microbial amyloid in neurodegeneration. *PLoS Pathogens*, 13:12, 2017. URL https://doi.org/10.1371/journal.ppat.1006654. 1

[22] R. M. Chakaroun, L. Massier, and P. Kovacs. Gut microbiome, intestinal permeability, and tissue bacteria in metabolic disease: Perpetrators or bystanders? *Nutrients*, 12:4, 2020. URL https://doi.org/10.3390/nu12041082. 1

[23] J. R. Kelly, P. J. Kennedy, J. F. Cryan, T. G. Dinan, G. Clarke, and N. P. Hyland. Breaking down the barriers: The gut microbiome, intestinal permeability and stress-related psychiatric disorders. *Frontiers in Cellular Neuroscience*, 9, 2015. URL https://doi.org/10.3389/fncel.2015.00392. 1

[24] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi. The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems. *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, 28(2):203–209, 2015. 1

[25] M.-S. Kim, Y. Kim, H. Choi, W. Kim, S. Park, D. Lee, D. K. Kim, H. J. Kim, H. Choi, D.-W. Hyun, J.-Y. Lee, E. Y. Choi, D.-S. Lee, J.-W. Bae, and I. Mook-Jung. Transfer of a healthy microbiota reduces amyloid and tau pathology in an alzheimer's disease animal model. *Gut*, 69(2):283–294, 2020. URL https://doi.org/10.1136/gutjnl-2018-317431. 1

[26] K. Kowalski and A. Mulak. Brain-gut-microbiota axis in alzheimer's disease. *Journal of Neurogastroenterology and Motility*, 25(1):48–60, 2019. URL https://doi.org/10.5056/jnm18087. 1

[27] P. Arosio, T. P. J. Knowles, and S. Linse. On the lag phase in amyloid fibril formation. *Physical Chemistry Chemical Physics*, 17(12):7606–7618, 2015. URL https://doi.org/10.1039/c4cp05563b. 1

[28] Raimon Sabate, Frederic Rousseau, Joost Schymkowitz, Cristina Batlle, and Salvador Ventura. Amyloids or prions? That is the question. *Prion*, 9(3):200–206, June 2015. ISSN 1933-6896. doi: 10.1080/19336896.2015.1053685. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4601216/. 1, 2

[29] A. Esteras-Chopo, L. Serrano, and M. L. de la. Paz. The amyloid stretch hypothesis: Recruiting proteins toward the dark side. *Proceedings of the National Academy of Sciences*, 102(46):16672–16677, 2005. URL https://doi.org/10.1073/pnas.0505905102. 1

[30] L. Zhang and J. D. Schmit. Theory of amyloid fibril nucleation from folded proteins. *Israel Journal of Chemistry*, 57(7):738–749, 2017. URL https://doi.org/10.1002/ijch.201600079. 1

[31] E. M. Sigurdsson, T. Wisniewski, and B. Frangione. Infectivity of amyloid diseases. *Trends in Molecular Medicine*, 8(9):411–413, 2002. URL https://doi.org/10.1016/S1471-4914(02)02403-6. 1

[32] R. Sabate. When amyloids become prions. *Prion*, 8(3):233–239, 2014. URL https://doi.org/10.4161/19336896.2014.968464. 1

[33] F. ul A. A. Minhas, E. D. Ross, and A. Ben-Hur. Amino acid composition predicts prion activity. *PLOS Computational Biology*, 13:4, 2017. URL https://doi.org/10.1371/journal.pcbi.1005465. 2

[34] M. R. H. Krebs, L. A. MorozovaRoche, K. Daniel, C. V. Robinson, and C. M. Dobson. Observation of sequence specificity in the seeding of protein amyloid fibrils. *Protein Science*, 13(7):1933–1938, 2004. URL https://doi.org/10.1110/ps.04707004. 2

[35] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. L. de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. Serpell, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242, 2010. URL https://doi.org/10.1038/nmeth.1432. 2

[36] C. Família, S. R. Dennison, A. Quintas, and D. A. Phoenix. Prediction of peptide and protein propensity for amyloid formation. *PLOS ONE*, 10:8, 2015. URL https://doi.org/10.1371/journal.pone.0134679. 2

[37] E. D. Ross, K. S. MacLea, C. Anderson, and A. Ben-Hur. A bioinformatics method for identifying q/n-rich prion-like domains in proteins. In D. M. Hatters and A. J.

Hannan, editors, *Tandem Repeats in Genes, Proteins, and Disease: Methods and Protocols*, pages 219–228. Humana Press, 2013. URL https://doi.org/10.1007/978-1-62703-438-8_16. 2

[38] A. K. Lancaster, A. Nutter-Upham, S. Lindquist, and O. D. King. Plaac: A web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics (Oxford, England)*, 30(17):2501–2502, 2014. URL https://doi.org/10.1093/bioinformatics/btu310. 2, 4, 7

[39] R. Zambrano, O. Conchillo-Sole, V. Iglesias, R. Illa, F. Rousseau, J. Schymkowitz, R. Sabate, X. Daura, and S. Ventura. Prionw: A server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores. *Nucleic Acids Research*, 43:W331–W337, 2015. URL https://doi.org/10.1093/nar/gkv490. 2

[40] Konstantinos Tenidis, Michaela Waldner, Jürgen Bernhagen, Wolfgang Fischle, Michael Bergmann, Marco Weber, Marie-Luise Merkle, Wolfgang Voelter, Herwig Brunner, and Aphrodite Kapurniotu. Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *Journal of Molecular Biology*, 295(4):1055–1071, January 2000. ISSN 0022-2836. doi: 10.1006/jmbi.1999.3422. URL https://www.sciencedirect.com/science/article/pii/S0022283699934228. 2

[41] V. A. Iconomidou, G. D. Chryssikos, V. Gionis, A. S. Galanis, P. Cordopatis, A. Hoenger, and S. J. Hamodrakas. Amyloid fibril formation propensity is inherent into the hexapeptide tandemly repeating sequence of the central domain of silkmoth chorion proteins of the a-family. *Journal of Structural Biology*, 156(3):480–488, 2006. URL https://doi.org/10.1016/j.jsb.2006.08.011. 2

[42] J. Zheng, B. Ma, C.-J. Tsai, and R. Nussinov. Structural stability and dynamics of an amyloid-forming peptide gnnqqny from the yeast prion sup-35. *Biophysical Journal*, 91(3):824–833, 2006. URL https://doi.org/10.1529/biophysj.106.083246. 2

[43] J. Stanislawski, M. Kotulska, and O. Unold. Machine learning methods can replace 3d profile method in classification of amyloidogenic hexapeptides. *BMC Bioinformatics*, 14(1):21, 2013. URL https://doi.org/10.1186/1471-2105-14-21. 2

[44] M. Emily, A. Talvas, and C. Delamarche. Metamyl: A meta-predictor for amyloid proteins. *PLOS ONE*, 8:11, 2013. URL https://doi.org/10.1371/journal.pone.0079722. 2

[45] R. Sant'Anna, C. Braga, N. Varejão, and K. M. Pimenta. The importance of a gatekeeper residue on the aggregation of transthyretin. *The Journal of Biological Chemistry*, 289(41):28324–28337, 2014. URL https://doi.org/10.1074/jbc.M114.563981. 2

[46] X. Wang, Y. Zhou, J.-J. Ren, N. D. Hammer, and M. R. Chapman. Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. *Proceedings of the National Academy of Sciences*, 107(1):163–168, 2010. URL https://doi.org/10.1073/pnas.0908714107. 2

[47] S. G. Estácio, S. S. Leal, J. S. Cristóvão, P. F. N. Faísca, and C. M. Gomes. Calcium binding to gatekeeper residues flanking aggregation-prone segments underlies non-fibrillar amyloid traits in superoxide dismutase 1 (sod1). *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1854(2):118–126, 2015. URL https://doi.org/10.1016/j.bbapap.2014.11.005. 2

[48] M. Niu, Y. Li, C. Wang, and K. Han. Rfamyloid: A web server for predicting amyloid proteins. *International Journal of Molecular Sciences*, 19:7, 2018. URL https://doi.org/10.3390/ijms19072071. 2, 3, 4

[49] Y. Li, Z. Zhang, Z. Teng, and X. Liu. Predamyl-mlp: Prediction of amyloid proteins using multilayer perceptron. *Computational and Mathematical Methods in Medicine*, 2020, 2020. URL https://doi.org/10.1155/2020/8845133. 2, 3, 4

[50] P. Charoenkwan, S. Kanthawong, C. Nantasenamat, Md. M. Hasan, and W. Shoombuatong. iamy-scm: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics*, 113(1):689–698, 2021. URL https://doi.org/10.1016/j.ygeno.2020.09.065. 2, 3, 4

[51] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012. URL https://doi.org/10.1093/bioinformatics/bts565. 3

[52] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou. Pse-in-one: A web server for generating various modes of pseudo components of dna, rna, and protein sequences. *Nucleic Acids Research*, 43:W65–W71, 2015. URL https://doi.org/10.1093/nar/gkv458. 3

[53] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158. URL https://doi.org/10.1093/bioinformatics/btl158. 3

[54] Quan Zou, Jiancang Zeng, Liujuan Cao, and Rongrong Ji. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, 173:346–354, January 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2014.12.123. URL https://www.sciencedirect.com/science/article/pii/S0925231215012801. 3

[55] Yonge Feng and Liaofu Luo. Use of tetrapeptide signals for protein secondary-structure prediction. *Amino Acids*, 35(3):607–614, October 2008. ISSN 1438-2199. doi: 10.1007/s00726-008-0089-7. URL https://doi.org/10.1007/s00726-008-0089-7. 3

[56] Rhys Heffernan, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18):2842–2849, September 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx218. URL http://academic.oup.com/bioinformatics/article/33/18/2842/3738544. Publisher: Oxford Academic. 4

[57] Mihaly Varadi, Greet DeăBaets, Wim F Vranken, Peter Tompa, and Rita Pancsa. AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Research*, 46(Database issue):D387–D392, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx950. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753394/. 4

[58] UniProt Consortium. Uniprot: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D480–D489, 2021. URL https://doi.org/10.1093/nar/gkaa1100. 5, 7

[59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *In Advances in Neural Information Processing Systems*, 32:8024–8035, 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. 5

[60] F. Pedregosa, Ga"el Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, Oct 2011. URL https://dl.acm.org/doi/10.5555/1953048.2078195. 5

[61] Marion Dalmasso, Colin Hill, and R. Paul Ross. Exploiting gut bacteriophages for human health. *Trends in Microbiology*, 22(7):399–405, July 2014. ISSN 0966-842X. doi: 10.1016/j.tim.2014.02.010. URL https://www.sciencedirect.com/science/article/pii/S0966842X14000456. 7

[62] Andrey N. Shkoporov and Colin Hill. Bacteriophages of the Human Gut: The Known Unknown of the Microbiome. *Cell Host & Microbe*, 25(2):195–209, February 2019. ISSN 1931-3128. doi: 10.1016/j.chom.2019.01.017. URL http://www.sciencedirect.com/science/article/pii/S1931312819300575. 7

[63] George V. Tetz, Kelly V. Ruggles, Hua Zhou, Adriana Heguy, Aristotelis Tsirigos, and Victor Tetz. Bacteriophages as potential new mammalian pathogens. *Scientific Reports*, 7(1):7043, August 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-07278-6. URL http://www.nature.com/articles/s41598-017-07278-6. Number: 1 Publisher: Nature Publishing Group. 7

[64] Thomas D. S. Sutton and Colin Hill. Gut Bacteriophage: Current Understanding and Challenges. *Frontiers in Endocrinology*, 10, 2019. ISSN 1664-2392. doi: 10.3389/fendo.2019.00784. URL https://www.frontiersin.org/articles/10.3389/fendo.2019.00784/full. Publisher: Frontiers. 7

[65] R. Sausset, M. A. Petit, V. Gaboriau-Routhiau, and M. De Paepe. New insights into intestinal phages. *Mucosal Immunology*, 13(2):205–215, March 2020. ISSN 1935-3456. doi: 10.1038/s41385-019-0250-5. URL http://www.nature.com/articles/s41385-019-0250-5. Number: 2 Publisher: Nature Publishing Group. 7

[66] Pedro Blanco-Picazo, Dietmar Fernández-Orth, Maryury Brown-Jaque, Elisenda Miró, Paula Espinal, Lorena Rodríguez-Rubio, Maite Muniesa, and Ferran Navarro. Unravelling the consequences of the bacteriophages in human samples. *Scientific Reports*, 10(1):6737, April 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-63432-7. URL http://www.nature.com/articles/s41598-020-63432-7. Number: 1 Publisher: Nature Publishing Group. 7

[67] Chandrabali Ghose, Melissa Ly, Leila K. Schwanemann, Ji Hyun Shin, Katayoon Atab, Jeremy J. Barr, Mark Little, Robert T. Schooley, Jessica Chopyk, and David T. Pride. The Virome of Cerebrospinal Fluid: Viruses Where We Once Thought There Were None. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.02061. URL https://www.frontiersin.org/articles/10.3389/fmicb.2019.02061/full. Publisher: Frontiers. 7

[68] George Tetz and Victor Tetz. Prion-Like Domains in Phagobiota. *Frontiers in Microbiology*, 8, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02239. URL https://www.frontiersin.org/articles/10.3389/fmicb.2017.02239/full. Publisher: Frontiers. 7, 8

[69] Bryan Maloney and Debomoy K. Lahiri. The Alzheimer's amyloid -peptide (A) binds a specific DNA A-interacting domain (AID) in the APP, BACE1, and APOE promoters

in a sequence-specific manner: characterizing a new regulatory motif. *Gene*, 488(1-2):1–12, November 2011. ISSN 1879-0038. doi: 10.1016/j.gene.2011.06.004. URL https://doi.org/10.1016/j.gene.2011.06.004. 7

[70] Sashank Agrawal, Pan-Hsien Kuo, Lee-Ya Chu, Bagher Golzarroshan, Monika Jain, and Hanna S. Yuan. RNA recognition motifs of disease-linked RNA-binding proteins contribute to amyloid formation. *Scientific Reports*, 9(1):6171, April 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-42367-8. URL http://www.nature.com/articles/s41598-019-42367-8. Number: 1 Publisher: Nature Publishing Group. 7

[71] Massimiliano Meli, Maria Gasset, and Giorgio Colombo. Are Amyloid Fibrils RNA-Traps? A Molecular Dynamics Perspective. *Frontiers in Molecular Biosciences*, 5, 2018. ISSN 2296-889X. doi: 10.3389/fmolb.2018.00053. URL https://www.frontiersin.org/articles/10.3389/fmolb.2018.00053/full. Publisher: Frontiers. 7

[72] Rajaraman Krishnan, Haim Tsubery, Ming Y. Proschitsky, Eva Asp, Michal Lulu, Sharon Gilead, Myra Gartner, Jonathan P. Waltho, Peter J. Davis, Andrea M. Hounslow, Daniel A. Kirschner, Hideyo Inouye, David G. Myszka, Jason Wright, Beka Solomon, and Richard A. Fisher. A Bacteriophage Capsid Protein Provides a General Amyloid Interaction Motif (GAIM) That Binds and Remodels Misfolded Protein Assemblies. *Journal of Molecular Biology*, 426(13):2500–2519, June 2014. ISSN 0022-2836. doi: 10.1016/j.jmb.2014.04.015. URL https://www.sciencedirect.com/science/article/pii/S0022283614001995. 7

[73] Jonathan M. Levenson, Sally Schroeter, Jenna C. Carroll, Valerie Cullen, Eva Asp, Ming Proschitsky, Charlotte H. Y. Chung, Sharon Gilead, Muhammad Nadeem, Hemraj B. Dodiya, Shadiyat Shoaga, Elliott J. Mufson, Haim Tsubery, Rajaraman Krishnan, Jason Wright, Beka Solomon, Richard Fisher, and Kimberley S. Gannon. NPT088 reduces both amyloid- and tau pathologies in transgenic mice. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 2(3):141–155, September 2016. ISSN 2352-8737. doi: 10.1016/j.trci.2016.06.004. URL https://www.sciencedirect.com/science/article/pii/S235287371630018X. 7

[74] J. Rodney Brister, Danso Ako-adjei, Yiming Bao, and Olga Blinkova. NCBI Viral Genomes Resource. *Nucleic Acids Research*, 43(D1):D571–D577, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1207. URL https://doi.org/10.1093/nar/gku1207. 8