Classification: Biophysics and Computational Biology

Title: *The Classifying Autoencoder: Gaining Insight to Amyloid Assembly of Peptides and Proteins*

Authors:

Michael J Tro[1†] *mtro@chem.ucsb.edu*

Nathaniel Charest[1†] *ncharest@chem.ucsb.edu*

Zachary Taitz[2] *zachary.taitz@yale.edu*

Joan-Emma Shea[*1] *shea@ucsb.edu*

Michael T Bowers[*1] *bowers@chem.ucsb.edu*

Author Affiliation:

[†] These authors contributed equally to the work presented here.

1**.** Department of Chemistry and Biochemistry, University of California Santa Barbara, Santa Barbara, CA, 93106-9510

2. Department of Chemistry, Yale University, PO Box 208107, New Haven CT 06520 8107

***Corresponding Author*:**

Name: Michael T Bowers

Address: Department of Chemistry and Biochemistry, University of California Santa Barbara, Santa Barbara, CA, 93106-9510

Phone Number: 805-893-2673

Email: bowers@chem.ucsb.edu

Name: Joan-Emma Shea

Address: Department of Chemistry and Biochemistry, University of California Santa Barbara, Santa Barbara, CA, 93106-9510

Phone Number: (805) 893-5604

Email: shea@chem.ucsb.edu

Abstract

Despite the importance of amyloid formation in disease pathology, the understanding of the primary structure – activity relationship for amyloid-forming peptides remains elusive. Here we use a new neural-network based method of analysis: the classifying autoencoder (CAE). This machine learning technique uses specialized architecture of artificial neural networks to provide insight into typically opaque classification processes. The method proves to be robust to noisy and limited datasets, as well as being capable of disentangling relatively complicated rules over datasets. We demonstrate its capabilities by applying the technique to an experimental database (the Waltz database) and demonstrate the CAE's capability to provide insight into a novel descriptor, dimeric isotropic deviation — an experimental measure of the aggregation properties of the amino acids. We measure this value for all 20 of the common amino acids and find correlation between dimeric isotropic deviation and the failure to form amyloids when hydrophobic effects are not a primary driving force in amyloid formation. These applications show the value of the new method and provide a flexible and general framework to approach problems in biochemistry using artificial neural networks.

Introduction

Amyloid aggregates are pathologically associated with numerous diseases and biological functions, with their existence having long drawn the attention of the biochemical and biological scientific communities [1–5]. An amyloid is defined as a proteinaceous fibrillar aggregate where the proteins are arranged with a "cross-β" spine — that is, both the protein backbone and the normal vector of the beta sheet plane are perpendicular to the fibril axis [6,7]. The fibril typically ranges from 60-200 Å in diameter when fully mature, with fibril-like subunits which have been observed in isolation with diameters as small as 10 Å [8–12]. While amyloids are particularly known for association with degenerative diseases such as Alzheimer's [4,13,14], Parkinson's [15], Huntington's [16], type 2 diabetes [17], and amyotrophic lateral sclerosis [18], they may also play beneficial roles. The motif appears in spider silks, egg shells, biofilms and biomechanical scaffolds for human synthetic pathways [2].

Despite the apparent importance of these aggregate structures, specifics regarding the physics driving the formation of these structures remains weakly characterized. There is interest in developing a process which relates a primary structure to that peptide's ability to form amyloids [19–21]. Enough algorithms have been developed on this topic that there exists prediction algorithms which take into account as many other algorithms as possible [22,23]. These meta-predictors improve predictions, but unfortunately step into a major criticism of machine learning based classifiers: hiding insight into why they make the classifications they do [24–26]. This is also common in attempts to use machine learning for understanding amyloid aggregation [24–26]. With these predictors it is possible to get either a positive or negative prediction, but it is hard to examine the process and learn what aspects of the peptide are contributing to for this prediction. For example, there could be multiple mechanisms for amyloid formation, such as one driven by hydrophobic interactions and one driven by electrostatic interaction. Many algorithms would be able to make the correct prediction, but the opaque construction of those algorithms makes it hard to distinguish the difference between the first and second mechanisms. Our aim was to develop a method that addressed these problems; a method which would be able to give us predictions and allow us to easily visualize what factors contributed to those predictions.

One machine learning framework, artificial neural networks (ANNs), offers a powerful approach to classification problems. By using numerical descriptions of a system, many fitting parameters, and a set of data points to learn from, ANNs generate a complicated mapping from the descriptions to an output. This output can be any target set of numbers but is often a numerical representation of a class — for our purposes, whether a peptide sequence is amyloid-forming or not. These classification networks have been employed in a number of fields, from ecological studies to economics to chemistry [27–30]. Attempts have been made to elucidate the inner workings of ANNs [31,32], however these methods can still leave intuition difficult to obtain.

Fundamentally, classification can be viewed as a dimensional reduction problem in which numerous pieces of descriptive input data (an attribute of an amino acid, in our case) must be reduced to a single descriptive dimension (the propensity to aggregate). Autoencoders are an architecture of ANNs that have been applied to the problem of dimensional reduction, capable of reducing relatively complex descriptions of objects to a lower dimension (termed the latent space), and then reconstructing the original description of the object with as much fidelity as can be allowed [33]. Differing versions of the basic autoencoder, perhaps most notably the Variational Autoencoder (VAE) [34] have emerged, with variants typically involving goals beyond dimensional reduction and reconstruction of the data [35]. In this paper, our goal was to develop a method of classification, which we call the classifying autoencoder (CAE), based on prior algorithms[29,30,34], that could offer easily-interpreted insight into our classification task.

For this work, we develop a relation between the attributes of the amino acids in a six amino acid peptide (hexapeptide) and the amyloid propensity of the sequences. The use of hexapeptides means the primary structure will dominate the behavior of a given peptide. While other peptide lengths can also form amyloids, hexapeptides are the shortest length for which a large number of amyloid forming peptides are known[36]. There are relatively few known examples of smaller peptides which form amyloids[37,38]. Longer peptides are more likely to have more complex mechanisms of amyloid formation involving thorough considerations of internal secondary and tertiary structures. We adopt a reductionist paradigm and posit understanding simple systems will help understanding of more complex systems in future work. A database exists in which about one thousand hexapeptides have been experimentally characterized as amyloid or non-amyloid, which we use here[36]. We this database to help prove the concept of our method and explore some of its potential usages, including elucidating the role specific descriptors play in establishing the classification and whether any motifs within these descriptor sequences can be identified as especially related to amyloid formation.

In the next sections we first assess the capability of the CAE to identify motifs by generating a dataset and then using the method to recover the motifs used in generating the dataset. With our method's concept successfully tested, we demonstrate its ability to analyze the relationship between a novel experimentally measured descriptor of a system, and that system's properties. We have called the new descriptor dimeric isotropic deviation (DID). Deviation from isotropic aggregation of amino acids has previously been suggested a parameter predictive of amyloid formation[39] for a small data set (3 peptides) and only 5 amino acids. DID differs from the isotropic deviation previously utilized (explained in Results and Discussion), but these simplifying differences enabled the measurement of all 20 common amino acids, allowing for a more robust exploration of DID and amyloid aggregation over a set of about one thousand peptides [36].

Methods

Generated Database

It was important to first test our architecture on a generated database, so that we could examine the method under a controlled setting. We devised a set of amino acid sequences that were assigned as belonging to an archetype (we use this term to describe a pattern within the sequence, such as alternating amino acid hydrophobicity), and then the sequence classified as positive or negative. Fig. 1 depicts a flow chart of the process used to generate this database.

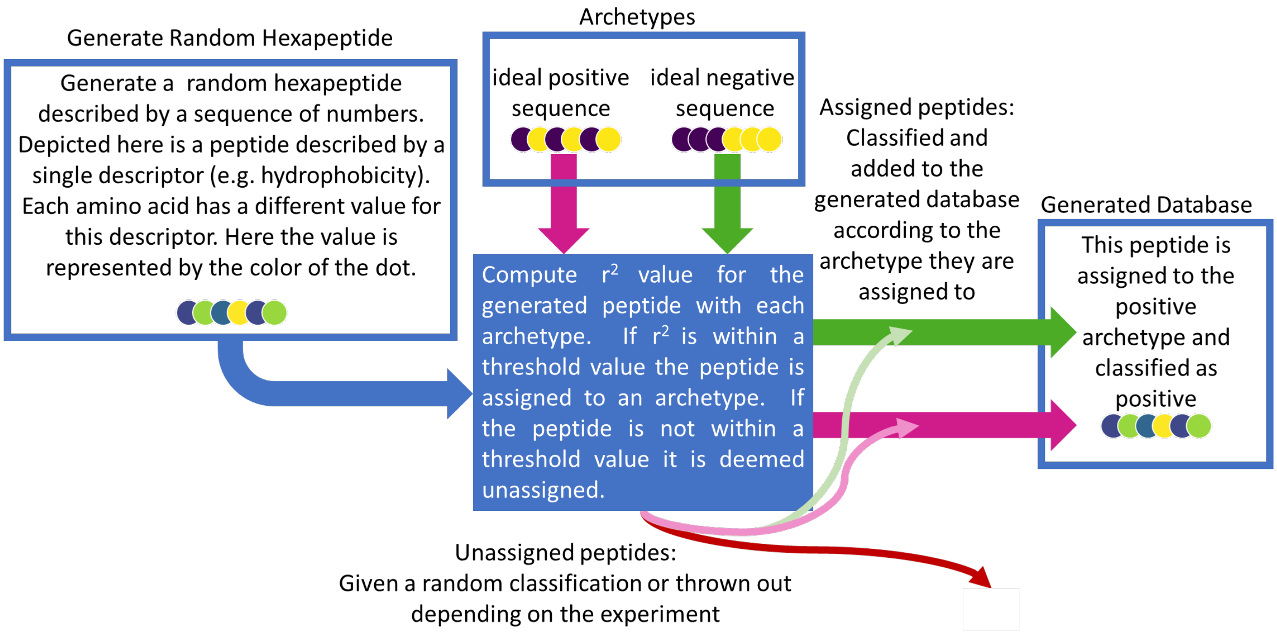
Generate Random Hexapeptide
Generate a random hexapeptide described by a sequence of numbers. Depicted here is a peptide described by a single descriptor (e.g. hydrophobicity). Each amino acid has a different value for this descriptor. Here the value is represented by the color of the dot.
Archetypes
ideal positive sequence
ideal negative sequence
Compute r² value for the generated peptide with each archetype. If r² is within a threshold value the peptide is assigned to an archetype. If the peptide is not within a threshold value it is deemed unassigned.
Assigned peptides: Classified and added to the generated database according to the archetype they are assigned to
Generated Database
This peptide is assigned to the positive archetype and classified as positive
Unassigned peptides:
Given a random classification or thrown out depending on the experiment

Fig. 1 This flow chart depicts the generation of the database. In the example depiction shown in the flow chart a random peptide is added to the database by being assigned to the positive archetype and classified as positive.

We generated two artificial descriptors for our validation database. A descriptor is a property of the system (e.g. the hydrophobicity of an amino acid). The two descriptors were uncorrelated and generated to be linearly distributed between 0 and 1. Peptide archetypes were also defined.   These archetypes are treated as the ideal positive or negative peptide (in the context of amyloid aggregation this assumes that certain patterns would yield an optimal activity, and the activity could be directly correlated to the degree of difference between an archetype's set of descriptor values and a peptide's set of descriptor values). The database was generated by randomly picking hexapeptides and classifying them as positive or negative (e.g. amyloid or not amyloid). These classifications were based on equation 1, which compares a generated peptide to an archetype:

$$r^2 = \Sigma_{i=amino\ acids}\Sigma_{j=descriptors}\left(f_{archetype;i,j} - f_{i,j}\right)^2 \qquad (1)$$

Where $f_{archetype;i,j}$ is the value of the j^{th} descriptor of the i^{th} amino acid in an archetypal peptide, and $f_{i,j}$ is the corresponding value of the peptide that is being classified. This r-squared value between the peptide and the archetypes served as our metric of distance. A peptide is assigned to an archetype with which it has the smallest r-squared value. The peptide is then classified based on which archetype it has been assigned to. In addition, if the peptide is not within a threshold r-squared of any of the archetypes, that peptide was deemed unassigned and either given a random classification or thrown out, depending on which validation test we were performing.

Experimental Database

Given that our goal is to better understand how the physical descriptors of a peptide relate to amyloid activity, we used a database of experimentally-verified peptides. We use the Waltz-DB [36,40] of 1089 hexapeptides that have been experimentally tested for amyloid formation by transmission electron microscopy, dye binding, and Fourier transform infrared spectroscopy. Of the 1089 peptides, 244 form amyloids, and the rest do not. This database is known to have over-representation of peptides similar to the peptide sequence STVIIE. The database was pruned to exclude any peptide which is within three point mutations of the peptide sequence STVIIE. This reduced the database to 946 total peptides. Of the pruned data set, 174 form amyloids; 772 do not.

The model was trained on half of the database, while the other half of the database was used for validation. The ratio of amyloid peptides to non-amyloid peptides was held constant over the training set and the validation set. Other fitting algorithms often use upwards of 66% of the database for training and

34% for validation [24,26]. We opted for a larger validation set at the cost of a smaller training set since the database is relatively small and we wanted to make sure there was enough data in the validation set to get a good idea of how generalizable the model is.

Polarity Descriptor

We found the hydrophobic parameter using the AAindex database [41–43]. This parameter was first measured by Jean-Luc Fauchere, in which the amino acids were dissolved in octanol and water and the relative solubility was measured [44]. We choose this metric for hydrophobicity because it correlates with many of the other hydrophobicity metrics in the database, it performs well for classification, and has a clear experimental basis and intuitive interpretation.

Cross Section Measurements

To measure the DID, amino acid samples were dissolved in water to concentrations between 1 and 12 millimolar. The cross section of the singly charged amino acid, and the cross section of the singly charged dimer cluster of the amino acid were measured using a lab-built ion mobility mass spectrometer which is described in detail elsewhere [45]. Briefly, this instrument uses nano-electrospray ionization to generate ions. The ions enter the instrument from atmosphere into a 10 torr source region. The ions are stored in an ion funnel and pulse injected into a 2-meter-long drift cell which is held at 0.25 torr above the pressure in the ion funnel to maintain a pure helium buffer gas in the drift cell. The ions exit the drift cell through another ion funnel and are mass selected with a quadrupole before being detected. This instrument is notable for minimization of energizing the sample ions at all stages. This allows us to easily measure non-covalently bound assemblies such as the amino acid clusters reported here.

To measure the cross section, the ions traverse the drift cell at various drift voltages. The time it takes to reach the detector is $t_A = \frac{l^2}{K_0}\left(\frac{T}{760}\right)\left(\frac{P}{V}\right) + t_0$, where l is the cell length, T the temperature, V the voltage across the cell, P the pressure in the cell, and t_0 the time from exiting the drift cell to the detector recorded for mobility calculations[46]. The reduced mobility, K_0, is related to the cross section by the equation $\sigma \approx \frac{3e}{16N_0}\left(\frac{2\pi}{\mu k_B T}\right)^{\frac{1}{2}}\left(\frac{1}{K_0}\right)$. Here e is the charge of the ion, N_0 is the number density of the buffer gas, μ is the reduced mass of the buffer gas and the ion, k_B is the Boltzmann constant, and σ is the cross section of the ion [47].

Software

All neural nets were constructed and trained using the Keras software package[48] with the Tensorflow backend[49].

Results and Discussion

Developing the Classifying Autoencoder

Classification is a specific type of dimensional reduction. We hypothesize we can learn more about why the classifying model is making its predictions by combining it with a variational autoencoder (VAE). A primer of VAEs can be found in the supporting information, but briefly, a VAE is an unsupervised neural-network-based dimensional reduction algorithm which seeks a robust reduced representation of a data set. As with any fitting algorithm, it quantifies the quality of the fit by defining and minimizing a loss function. For standard linear regression this is typically the sum of squares of the residuals, r^2. The VAE has a two-term loss function. The first term relates the fidelity between the reduced representation and the original representation. This is called the reconstruction term since it is a measure of how well the model can reconstruct the original representation if only given the reduced representation. The second term adds noise to the data during training. These competing loss terms lead to robust reduced representations.

We used the underlying architecture and concept of the VAE, but added another term to the loss function to make the reduced representation also function as a classification metric. We have called this the classifying autoencoder (CAE), and depicted it in Fig. 2. Inputs (a description of the peptide) are fed into the model via the input nodes. The depiction in Fig. 2 shows only four input nodes, but in the final model there will be an input node for each value that represents the peptide, i.e. the number of descriptors times the number of amino acids in the peptide. The hidden layers add more fitting parameters. The nodes labeled μ represent what is termed the latent space. Typically, the term latent space is used to refer to the space of the reduced representation. Here, these values are also used as the prediction. The latent space is two dimensional, one for the amyloid propensity and one for the non-amyloid propensity. A peptide is classified depending which node outputs a higher value. The nodes labeled $N(\mu, \sigma^2)$ inject noise into the data during training. This noise is in the form of a normal distribution centered at the reduced representation, μ, and has a standard deviation, σ^2. The nodes to the right of the nodes labeled $N(\mu, \sigma^2)$ (the decoder) attempt to reconstruct the original input. For a more detailed explanation of this please see the primer of VAEs in the supporting information.

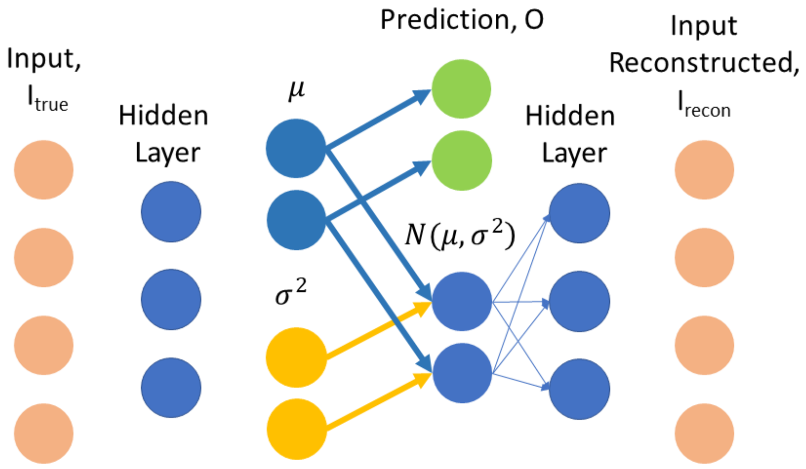

Fig. 2 Depicted is the architecture of a classifying autoencoder (CAE) with four inputs, and one hidden layer with three nodes. The latent space in this model is two-dimensional and is labeled μ. These nodes are also used as the prediction layer. Noise is added to the latent space at the nodes labeled $N(\mu, \sigma^2)$; this noise is in the form of a normal distribution centered on the reduced

representation, μ, with a standard deviation, σ^2. The decoder (all nodes to the right of the nodes which introduce the noise) tries to reconstruct the input. Dense connectivity (see primer of VAEs in Supporting Information) can be assumed for all layers not drawn explicitly. In addition, connectivity has been drawn explicitly at the latent space to highlight that the output nodes are not fed into the decoder, but the nodes labeled $N(\mu,\sigma)$ are fed to the decoder.

Validation on a Constructed Data Set

To verify that the CAE successfully elucidates and reconstructs characteristic archetypes, we generated an artificial peptide database as discussed in Methods. Using the constructed database allowed us to verify the model was performing its intended functions, while also testing how sensitive the model is to potential issues within the database, such as small database sizes or flawed results. We did this in two ways. First, we allowed for some unassigned peptides (peptides which were not in the neighborhood of any archetype) to be given a random class to see if the model would be able to see through the resulting noise. This tests situations where the descriptor we are using contributes to the amyloid activity of some of the peptides, while other peptides are dominated by a mechanism unrelated to the descriptors that have been chosen. The second test only used peptides that have been assigned to an archetype but introduced a stochastic element to classifications. When a peptide was assigned to an archetype it was classified as amyloidogenic or non-amyloidogenic according to a probability. The second scenario captures errors in the experimental data in the database, or an amyloid mechanism that only partially relates to the chosen descriptors.

All models are trained on 500 peptides and validated with 500 different peptides. Peptides are described with two descriptors per amino acid. The axes for each plot in Fig. 3 are the values in the latent space; that is the values output by the two nodes labeled μ in Fig. 2. The y-axis is the positive prediction axis, and the x-axis is the negative prediction axis. A peptide is classified as positive if its positive prediction value is greater than its negative prediction value. Thus, if a peptide falls above the red line on the plots that peptide is predicted positive, while falling below the red line is a negative prediction. Figs. 3 A, D, and E plot each peptide in the database according to where they fall in latent space. Figs. 3 B and C show the reconstructed description of the peptide for equally spaced points in latent space. These types of plots will be referred as reconstruction plots.

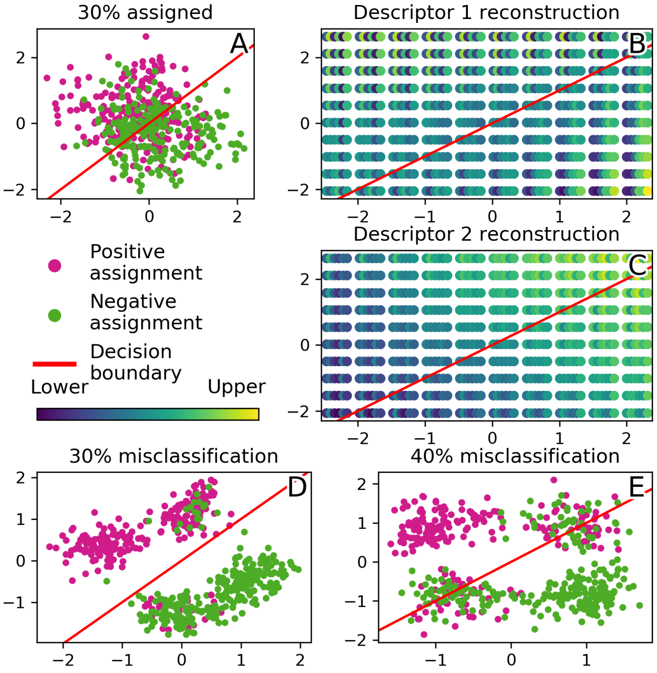

Fig. 3 For all plots, the y-axis is the positive prediction axis, and the x-axis is the negative prediction axis. A peptide is predicted positive if it is falls above the red line and is predicted negative if it falls below the red line. Plot (A) shows where each peptide is encoded by the CAE, while plots (B) and (C) show the reconstructed description at each point in latent space for that same model. Plots (D) and (E) are each a separate model (see text) and show where each peptide in their database encodes to in latent space.

Figs. 3 A-C are generated from the same model. Of the 1000 peptides 27% are assigned to an archetype, and subsequently classified as either positive or negative depending on the archetype. The positive archetype is LULULU (U = upper, L = lower) in descriptor 1 and LLLLLL or UUUUUU in descriptor 2. The negative archetype is LLLUUU for descriptor 1 and either LLLLLL or UUUUUU in descriptor 2. All peptides not assigned to an archetype were not within a threshold r-squared distance of any of these archetypes and were classified randomly.

In Fig. 3 A, each peptide is encoded to two numbers (the values output by μ) and plotted according to those values, showing how the peptides are arranged in latent space. The color of the marker represents the peptide's classification in the database; pink markers are positive, while green are negative. In Fig. 3 B and C, we visualize the reconstructed description of the peptide, descriptors 1 and 2, respectively, in latent space. This representation of the peptide description arranged in latent space, the reconstruction plot, is the key to gaining intuition from the CAE, as it visualizes the different regions of positive and negative predictions that the CAE identified.

Figs. 3 D and E depict different models than Figs. 3 A-C. These use a database generated to mimic a set of experiments that yielded occasionally flawed results. In Figs. 3 D and E all peptides in the database are within a threshold distance to one of four archetypes. For the two positive archetypes, one archetype was always classified positive, while the other archetype was misclassified at the rate indicated in the plot title. The negative archetypes were assigned in the same way.

These results show the models can simultaneously sort the data into the positive and negative classifications and identify the original archetypes used to generate the data. In the top left of Fig. 3 A, the positive prediction region of the latent space, positive peptides have been separated from a mixture of positive and negative peptides, correctly predicting those peptides as positive. The corresponding region in the reconstruction plot, Figs. 3 B and C, correctly reflects the positive archetypes. This happens similarly for the negative prediction region. We, also, learn how to interpret the reconstruction plot by examining Figs. 3 B and C. The middle of Fig. 3 A, the data's latent space distribution, shows mixed positive and negative peptides; in Figs. 3 B and C, the reconstruction plot, this region shows no evidence of the positive or negative archetypes. However, as we move to the top left of the data's latent space distribution, Fig. 3 A, we see a separation of positive classifications from the mixture of classifications; when we follow this trajectory in the reconstruction plot, Figs. 3 B and C, the positive archetype emerges. The separation of a single class from a mixture of classes can tell us about the trend that contributed to that separation.

In Figs 3 D and E, the model correctly shows four clusters in the latent space, according to the four archetypes used to construct the database. The reconstruction plot (Fig. S5) correctly reflects the four archetypes. This gives us insight to how the model deals with the uncertainty in the data. In Fig. 3 D, the archetype which has been 80% classified positive and 20% classified negative is placed in the positive prediction region of the latent space. However, this is nearer the decision boundary (the red line) than the cluster associated with the 100% positive archetype, suggesting the model identified the ambiguous archetype. Further, in Fig. 3 E, the ambiguous archetype was associated 60% to one classification and 40% to the other. In this case, the cluster that represents the ambiguous archetype is placed nearly atop the decision boundary, leaning slightly positive. The method is capable of making identifications regarding how an archetype leans, in addition to characterizing archetypes that are certainly associated with activities.

We note our validations show our method works with large databases that are typically used in machine learning (N = 10,000; Fig. S7), but crucially also with the limited databases we have available for amyloid studies (N = 1000; as shown here). This suggests potential generalizability of the models to problems associated with relatively small databases, such as the Waltz database we use later [36].

Ultimately, these results demonstrate the CAE's ability to relate sequences to an interesting activity. Even adding disturbances to the ideality of an artificially constructed database, the CAE was able to mine the patterns associated with the class of interest, and discern when a pattern had a leaning, rather than a fixed identity. This suggests the validity of this method for the task at hand: identifying characteristics and motifs of sequences that yield amyloidogenic behavior.

CAE on an Experimental Database: Hydrophobicity

Metrics related to hydrophobicity were found to be the most effective descriptors, and such a metric is used in both descriptors examined here. In Fig. 4 B (and later in Fig. 5 B) we can see a region of peptides with yellow or green amino acids in the middle (positions 3 and 4) and dark green or blue amino acids on the ends. This means peptides in this region of the latent space tend to be hydrophobic in the middle, and more hydrophilic on the ends, suggesting this type of amyloid fibril buries the hydrophobic core by stacking while the hydrophilic ends on the outside interact with water. It should be noted in both cases much of this region is an extrapolation by the model (there are few data points in the latent space in these regions). While extrapolation must be taken with caution, this motif in this "most-likely amyloid" region is worth noting due to the intuitive sense that hydrophobic amino acids should be buried away from the solvent. This motivates further investigation on sequences capturing this motif. In other words, if a goal is to investigate the coarse forces driving amyloid formation or design new amyloid forming peptides, the CAE's extrapolation can be a hypothesis to pursue.

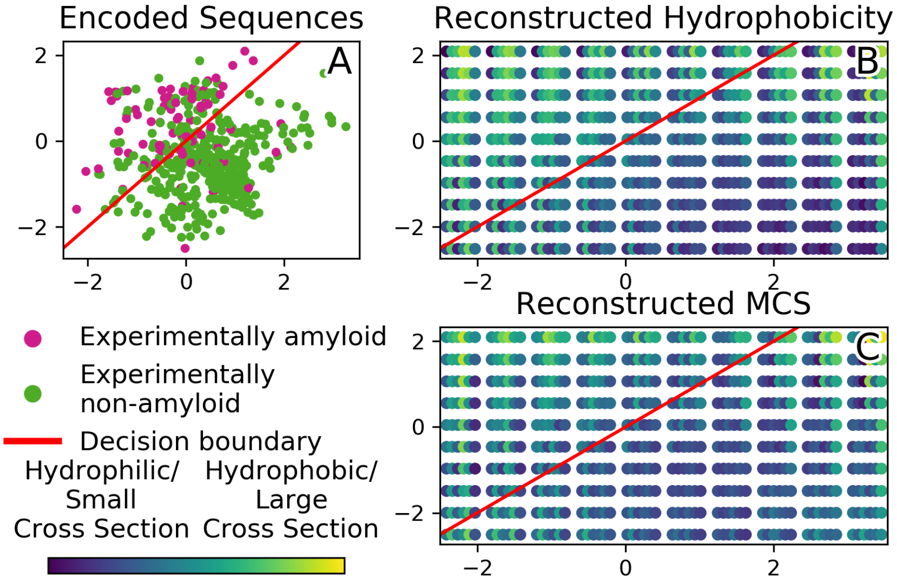

Fig. 4 Representative model trained using hydrophobicity and monomer cross section (MCS). (A) All sequences in the validation set plotted in latent space. Each axis here is the value of one of the latent space nodes. The color of the point represents the experimental classification of that peptide. (B) and (C) show the reconstructed descriptions. The axes here also represent values in the latent space, but the markers represent peptide descriptions. Each group of six dots represents a peptide, and the color of that dot represents the reconstructed descriptor value at that point in latent space. (B) depicts the hydrophobicity of the peptides, where yellow is hydrophobic, and blue is hydrophilic. (C) depicts the monomer cross section (MCS) of each point in latent space. Here yellow is a large cross section, and blue a small cross section.

There also exists some signs of the West et al result [50] of NPNPNP (P = hydrophobic (polar), N = hydrophilic (non-polar) within the core residues (2, 3, 4, 5) in both models. This patterning is also characterized by less extreme hydrophobicity, suggesting this motif is preferred by those residues with moderate hydrophobicity values. It is worth noting this region has been interpolated as there are many amyloid points in this region of the latent space; we can then be more confident that these motifs are well-represented within the database. Observations based on this interpolated region could also provide grounds to investigate forces driving amyloid formation or to inspire novel amyloid forming peptides.

These two motifs are consistently represented in the amyloid region independently of the second descriptor, giving further confidence these motifs are mirrored in the data.

CAE on an Experimental Database: Monomeric Cross Section

Fig. 4 represents a model trained using Monomeric Cross Section (reported in Table 1) and hydrophobicity as the descriptors. The populated region on the amyloid side tends to include mid-to-large residues, while the populated region in the non-amyloid side tends to include small residues. This could suggest a preference for bulky side chains, perhaps to help drive amyloid stability through surface-area dependent forces such as van der Waals. Additionally, on the amyloid side, there is some alternation of large and small residues. We also note that hydrophobicity similarly alternates in the same region of latent space. Perhaps this alludes to a connection between the size of a side chain and its potential for stronger hydrophobic-related forces resulting in a preference for sequences that alternate large, hydrophobic residues and small, hydrophilic residues[51–53].

Monomer Cross Section and Dimeric Isotropic Deviation

Amino acid	**Monomer Cross Section (Å$\pm$Standard Deviation)**	**Dimeric isotropic deviation, Δi_2 $\times 100$ ($\pm$ Standard Deviation)**
Glutamic acid	61.9 $\pm$ 0.3	-6.1 $\pm$ 0.3

Leucine	65.3 ± 0.2	-5.8 ± 0.4
Isoleucine	64.2 ± 0.4	-4.8 ± 0.5
Glutamine	63.3 ± 0.1	-4.7 ± 0.4
Valine	58.8 ± 0.2	-4.7 ± 0.7
Methionine	65.6 ± 0.3	-4.5 ± 0.2
Proline	56.5 ± 0.2	-3.2 ± 0.2
Histidine	66.3 ± 0.4	-2.9 ± 0.5
Threonine	56.3 ± 0.3	-2.5 ± 0.6
Aspartic acid	57.8 ± 0.4	-1.7 ± 0.5
Arginine	71.8 ± 0.2	-1.0 ± 0.4
Asparagine	59.0 ± 0.4	-0.1 ± 0.5
Lysine	65.4 ± 0.2	0.5 ± 0.4
Alanine	50.6 ± 0.3	0.8 ± 0.3
Serine	52.1 ± 0.5	1.2 ± 0.9
Phenylalanine	72.0 ± 0.4	3.5 ± 0.6
Tyrosine	75.2 ± 0.3	4.0 ± 0.0
Tryptophan	81.3 ± 0.7	6.0 ± 1.0
Cysteine	55.7 ± 0.3	9.2 ± 0.3
Glycine	49.1 ± 0.4	11.6 ± 0.5

Table 1 Experimentally measured monomer cross section and dimeric isotropic deviation ($\mathit{\Delta} i_2$) for each amino acid. The $\mathit{\Delta} i_2$ have been multiplied by 100 for ease of reading. Convention dictates a negative value is associated with growth larger than isotropic prediction, zero is isotropic growth, and a positive deviation growth more compact than the isotropic prediction.

Introducing Dimeric Isotropic Deviation (DID)

To offer insight into isotropic deviation, consider growth around a sphere as material is added. If that volume is distributed equally around the object, isotropically, it is straight-forward to write an equation which predicts the cross section when material is added: $\sigma^{iso} = \sigma_0 \left(\frac{V}{V_0}\right)^{2/3}$, where V_0 is the original volume of the sphere , V the final volume of the sphere, σ_0 the cross section of the original sphere, and σ^{iso} the cross section given isotropic addition of volume. If that volume is not added isotropically, or the overall density changes, the system will deviate from that prediction. In the same way, if we calculate the volume of an amino acid based off our experimentally measured cross section and assume isotropic growth, we can predict the cross section of an oligomer (in this case, a cluster of amino acids) based on the volume of the monomer using the equation $\sigma_n^{iso} = \sigma_1^{exp}\, n^{2/3}$, where n is the number of amino acid molecules in the oligomer [39]. Most amino acids do not grow isotropically, and we call the degree of deviation from this growth isotropic deviation.

It is intuitive that this property of amino acid aggregation could be used to make predictions about the aggregation properties of peptides since it reflects some degree of order in the amino acid aggregates. In the Do paper, isotropic deviation is measured for different large order oligomers (n = 20 to 30), but was only measured for five amino acids, and verified on three peptides [39]. As we collected more data on aggregation of amino acids, we found that this value was oligomer size dependent (Fig. S3). We also found the monomer and dimer to be the only oligomer sizes that could we could consistently observe across all amino acids. The desire for a systematic metric for all amino acids drove the development of what we call DID (reported in Table 1). For the data available, comparison of Do's measure and DID does not show strong correlation, however DID's basis in peptide packing behavior suggests a potential relation to amyloid formation.

DID is calculated as follows. We have measured the cross section of the singly charged monomer and the singly charged dimer of each of the 20 canonical amino acids (arrival time distributions and cross sections in Fig. S4). If the dimer cross section is larger than the isotropic prediction, convention dictates a negative isotropic deviation is obtained, which we will refer to as extended growth. An experimental dimer cross section which is smaller than the isotropic prediction, compact growth, results in a positive isotropic deviation according to the equation, $\Delta i_2 = \left(1 - \frac{\sigma_2^{exp}}{\sigma_2^{iso}}\right)$. Here σ_2^{exp} is the experimentally measured cross section of the dimer with one charge, and σ_2^{iso} the isotropic prediction of the dimer based on the singly charged monomer's cross section.

Use of DID (a descriptor to be assessed) along with hydrophobicity (a known strong descriptor) shows an important power of the CAE: the ability to assess the relationship between a potential descriptor and classification. The strong descriptor essentially scaffolds the latent space's shape, ensuring good classifications, while the other descriptor can then be used to refine details within the latent space, either indicating that descriptor's relationship to the activity through meaningful contributions, or no such relationship through a lack of systematic contributions. This process is illustrated below.

CAE on an Experimental Database: Dimeric Isotropic Deviation (DID)

Here we probe the relationship between DID and amyloid propensity. For the most part, Fig. 5 C shows few features in the amyloid region and the peptides are generally on the extended side of DID. The top left shows some signs of compact DID. This is also the same region where the hydrophobic core motif is represented. Like the monomer cross section result, here the hydrophobicity is likely the larger factor governing amyloid formation, as evidenced by the larger diversity of hydrophobicity motifs in the amyloid region. In the non-amyloid region, there exists a region of mixed amyloid and non-amyloid points (middle of the plots), as well as a region of pure non-amyloid points (the right of the plots), reminiscent to the pattern we saw in the distribution of points during the first validation experiment (Fig. 3 A). Within these regions the hydrophobicity motifs have relatively low diversity, being generally hydrophilic, while there is greater diversity in the DID motifs. Critically, as one moves deeper into the non-amyloid region, one observes a rise in the compactness of the residues. Thus, in the same way the model from Fig. 3 A determined the archetype in the pure green region, the CAE has determined a strong relationship between compactness and a failure to grow fibrils – the extrapolated "least amyloidogenic" peptides (those that would appear in the bottom right of Fig. 5 C) are most strongly characterized by a higher degree of compactness, with less distinguishing features in hydrophobicity representation.

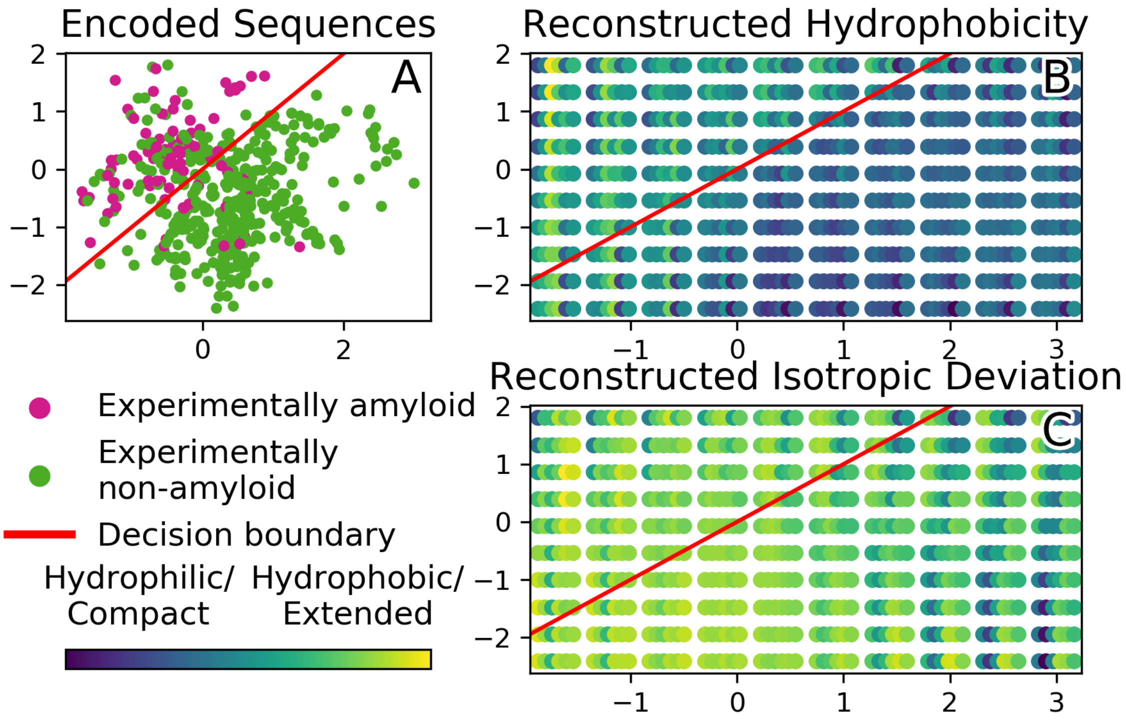

Fig. 5 Representative model trained using hydrophobicity and DID. This figure is the same representation of a model as Fig. 4except (C) depicts the DID of each point in latent space. Here yellow is extended growth, and blue is compact growth.

These results provide potential insight about how DID relates to amyloid formation. Namely, compact growth of the amino acids could block the amyloid process of the peptide when hydrophobic interactions are not a significant driving force of amyloid formation. While it was not found that DID could be used by itself to attain reliable correlations with amyloid-forming behavior, likely due to the specificity of the interaction observed at the dimer level, the CAE determined that DID could be strongly related to a failure to form fibrils. Further, from this observation we may gain some insight about the differences between amyloid forming hexapeptides, and larger proteins. The residue with the most compact isotropic deviation is glycine, and indeed the peptides in the non-amyloid forming/compact isotropic deviation region of the latent space are rich in glycine. This is a curious result since amyloids are often associated with glycine rich proteins as they tend to be intrinsically disordered [54,55]. Further, it has also recently been shown that glycine is an essential residue in cylindrin formation, structure that may be responsible for breaching the plasma membrane potentially leading to neuron death. However, for cylindrin formation peptide lengths on 11 or more amino acids are required. Here, however, we see the opposite trend. Perhaps amyloid structures for small hexapeptides are destabilized by the lack of side chains from glycine. Larger proteins have more backbone interactions and other non-glycine side chains to stabilize the amyloid structure. This observation may help in understanding how to use data taken on hexapeptides to make predictions about proteins. Precise mechanistic insight is beyond the capability of this method. However, its ability to obtain correlations may motivate more detailed experiments or simulations which can investigate the hypotheses yielded by the trends within the CAE's latent space representations.

Conclusions

Here we develop a method combining the techniques of an artificial neural network classifier and the variational autoencoder (VAE) to analyze a set of experimental data and produce relationships between properties of the peptides and their amyloidogenic activity. This method was validated on a set of artificially generated data, demonstrating its ability to perform the functions intended as well as demonstrate a robustness to both noisy and limited datasets – common features of currently available data for biochemical assembly systems.

The CAE was then applied to the experimentally verified Waltz database to mine important motifs correlated to amyloidogenic behavior. The CAE was able to rediscover previously observed relationships regarding hydrophobicity and steric size and additionally establish a link between DID and amyloidogenic activity. This observation demonstrates its ability to provide relationships between relatively complex input spaces and a reduced-dimension output associated with whether a peptide produces amyloid fibrils. This capability enabled us to observe an extrapolated but intuitive suggestion that hexapeptides with highly hydrophobic, bulky cores and hydrophilic, smaller termini will be among the most likely to form fibrils. We were also able to detect that the database has a strong representation of sequences in which alternating patterns of hydrophobic and intermediate residues correlate to amyloid formation.

In addition, we used this method to elucidate the relationship between novel descriptors (such as the newly reported DID) and activities of interest. The CAE was able to extract trends within the DID of peptides, and demonstrate a relationship to amyloidogenicity, even though this relationship only weakly contributed to the overall score of the model. The hydrophobicity of the peptide dominates in this database, but we are still able to observe cases where hydrophobic forces did not strongly contribute, and compact amino acid growth could be clearly associated with failure to form amyloid.

This method can easily be generalized to analyze many problems that involve understanding complicated data. There are no restrictions on the number of classes or inputs that can be considered, and while we use classification in the latent space, other loss functions could be used to alter the meaning of the axes. While we demonstrated this works on relatively small datasets, we took great care to avoid overfitting. The more inputs (and thus hidden layer fitting parameters) and the smaller the dataset, the more likely the model will overfit.

We believe we have successfully illustrated a quick and understandable analysis of high dimensional, nonlinearly dependent data. We set out to probe the relationship between DID and amyloid formation, and our method offered a relatively rapid way to obtain correlations of significance. The general approach established here could be used to mine databases for directions to take when considering future experiments. As science continues to move to higher throughput methods, higher dimensionality, and more complicated systems, machine learning methods have flourished at the cost of physical/chemical insight. Here we have used a prescription to open the black box and have offered a way to gain intuitive insight to the system which has been modeled, while retaining the full power of machine learning’s modeling abilities.

Supporting Information

Supplementary text

Figs. S1 to S7

Primer of a variational autoencoder (VAE)

Acknowledgement: We greatly appreciate the support of the National Science Foundation under grant CHE-1565941 (MTB) and grant MCB-1716956 (JS). We also acknowledge support from the Center for Scientific Computing from the CNSI, MRL: an NSF MRSEC (DMR-1720256).

References

(1) Chiti, F.; Dobson, C. M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.* **2006**, *75* (1), 333–366. https://doi.org/10.1146/annurev.biochem.75.101304.123901.

(2) Fowler, D. M.; Koulov, A. V.; Balch, W. E.; Kelly, J. W. Functional Amyloid – from Bacteria to Humans. *Trends Biochem. Sci.* **2007**, *32* (5), 217–224. https://doi.org/10.1016/j.tibs.2007.03.003.

(3) Zhao, W.-Q.; Townsend, M. Insulin Resistance and Amyloidogenesis as Common Molecular Foundation for Type 2 Diabetes and Alzheimer's Disease. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **2009**, *1792* (5), 482–496. https://doi.org/10.1016/j.bbadis.2008.10.014.

(4) Bernstein, S. L.; Dupuis, N. F.; Lazo, N. D.; Wyttenbach, T.; Condron, M. M.; Bitan, G.; Teplow, D. B.; Shea, J.-E.; Ruotolo, B. T.; Robinson, C. V.; et al. Amyloid-β Protein Oligomerization and the Importance of Tetramers and Dodecamers in the Aetiology of Alzheimer's Disease. *Nat. Chem.* **2009**, *1* (4), 326–331. https://doi.org/10.1038/nchem.247.

(5) Bleiholder, C.; Dupuis, N. F.; Wyttenbach, T.; Bowers, M. T. Ion Mobility–Mass Spectrometry Reveals a Conformational Conversion from Random Assembly to β-Sheet in Amyloid Fibril Formation. *Nat. Chem.* **2011**, *3* (2), 172–177. https://doi.org/10.1038/nchem.945.

(6) Astbury, W. T.; Dickinson, S.; Bailey, K. The X-Ray Interpretation of Denaturation and the Structure of the Seed Globulins. *Biochem. J.* **1935**, *29* (10), 2351-2360.1.

(7) Morriss-Andrews, A.; Shea, J.-E. Computational Studies of Protein Aggregation: Methods and Applications. *Annu. Rev. Phys. Chem.* **2015**, *66* (1), 643–666. https://doi.org/10.1146/annurev-physchem-040513-103738.

(8) Economou, N. J.; Giammona, M. J.; Do, T. D.; Zheng, X.; Teplow, D. B.; Buratto, S. K.; Bowers, M. T. Amyloid β-Protein Assembly and Alzheimer's Disease: Dodecamers of Aβ42, but Not of Aβ40, Seed Fibril Formation. *J. Am. Chem. Soc.* **2016**, *138* (6), 1772–1775. https://doi.org/10.1021/jacs.5b11913.

(9) Makin, O. S.; Serpell, L. C. Examining the Structure of the Mature Amyloid Fibril. *Biochem. Soc. Trans.* **2002**, *30* (4), 521–525. https://doi.org/10.1042/.

(10) Jiménez, J. L.; Nettleton, E. J.; Bouchard, M.; Robinson, C. V.; Dobson, C. M.; Saibil, H. R. The Protofilament Structure of Insulin Amyloid Fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (14), 9196–9201. https://doi.org/10.1073/pnas.142459399.

(11) Fitzpatrick, A. W. P.; Debelouchina, G. T.; Bayro, M. J.; Clare, D. K.; Caporini, M. A.; Bajaj, V. S.; Jaroniec, C. P.; Wang, L.; Ladizhansky, V.; Müller, S. A.; et al. Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril. *Proc. Natl. Acad. Sci.* **2013**, *110* (14), 5468–5473. https://doi.org/10.1073/pnas.1219476110.

(12) Sipe, J. D.; Cohen, A. S. Review: History of the Amyloid Fibril. *J. Struct. Biol.* **2000**, *130* (2), 88–98. https://doi.org/10.1006/jsbi.2000.4221.

(13) Jarrett, J. T.; Lansbury, P. T. Seeding "One-Dimensional Crystallization" of Amyloid: A Pathogenic Mechanism in Alzheimer's Disease and Scrapie? *Cell* **1993**, *73* (6), 1055–1058. https://doi.org/10.1016/0092-8674(93)90635-4.

(14) Stelzmann, R. A.; Norman Schnitzlein, H.; Reed Murtagh, F. An English Translation of Alzheimer's 1907 Paper, "Über Eine Eigenartige Erkankung Der Hirnrinde." *Clin. Anat.* **1995**, *8* (6), 429–431. https://doi.org/10.1002/ca.980080612.

(15) Maries, E.; Dass, B.; Collier, T. J.; Kordower, J. H.; Steece-Collier, K. The Role of α-Synuclein in Parkinson's Disease: Insights from Animal Models. *Nat. Rev. Neurosci.* **2003**, *4* (9), 727–738. https://doi.org/10.1038/nrn1199.

(16) Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Huntingtin-Encoded Polyglutamine Expansions Form Amyloid-like Protein Aggregates In Vitro and In Vivo. *Cell* **1997**, *90* (3), 549–558. https://doi.org/10.1016/S0092-8674(00)80514-0.

(17) Westermark, P.; Andersson, A.; Westermark, G. T. Islet Amyloid Polypeptide, Islet Amyloid, and Diabetes Mellitus. *Physiol. Rev.* **2011**, *91* (3), 795–826. https://doi.org/10.1152/physrev.00042.2009.

(18) Elam, J. S.; Taylor, A. B.; Strange, R.; Antonyuk, S.; Doucette, P. A.; Rodriguez, J. A.; Hasnain, S. S.; Hayward, L. J.; Valentine, J. S.; Yeates, T. O.; et al. Amyloid-like Filaments and Water-Filled Nanotubes Formed by SOD1 Mutant Proteins Linked to Familial ALS. *Nat. Struct. Mol. Biol.* **2003**, *10* (6), 461–467. https://doi.org/10.1038/nsb935.

(19) Walsh, I.; Seno, F.; Tosatto, S. C. E.; Trovato, A. PASTA 2.0: An Improved Server for Protein Aggregation Prediction. *Nucleic Acids Res.* **2014**, *42* (Web Server issue), W301–W307. https://doi.org/10.1093/nar/gku399.

(20) Tartaglia, G. G.; Vendruscolo, M. The Zyggregator Method for Predicting Protein Aggregation Propensities. *Chem. Soc. Rev.* **2008**, *37* (7), 1395–1401. https://doi.org/10.1039/B706784B.

(21) Thompson, M. J.; Sievers, S. A.; Karanicolas, J.; Ivanova, M. I.; Baker, D.; Eisenberg, D. The 3D Profile Method for Identifying Fibril-Forming Segments of Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (11), 4074–4078. https://doi.org/10.1073/pnas.0511295103.

(22) Emily, M.; Talvas, A.; Delamarche, C. MetAmyl: A METa-Predictor for AMYLoid Proteins. *PLOS ONE* **2013**, *8* (11), e79722. https://doi.org/10.1371/journal.pone.0079722.

(23) Tsolis, A. C.; Papandreou, N. C.; Iconomidou, V. A.; Hamodrakas, S. J. A Consensus Method for the Prediction of 'Aggregation-Prone' Peptides in Globular Proteins. *PLOS ONE* **2013**, *8* (1), e54175. https://doi.org/10.1371/journal.pone.0054175.

(24) Stanislawski, J.; Kotulska, M.; Unold, O. Machine Learning Methods Can Replace 3D Profile Method in Classification of Amyloidogenic Hexapeptides. *BMC Bioinformatics* **2013**, *14*, 21. https://doi.org/10.1186/1471-2105-14-21.

(25) Kotulska, M.; Unold, O. On the Amyloid Datasets Used for Training PAFIG - How (Not) to Extend the Experimental Dataset of Hexapeptides. *BMC Bioinformatics* **2013**, *14*, 351. https://doi.org/10.1186/1471-2105-14-351.

(26) Kim, C.; Choi, J.; Lee, S. J.; Welsh, W. J.; Yoon, S. NetCSSP: Web Application for Predicting Chameleon Sequences and Amyloid Fibril Formation. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W469–W473. https://doi.org/10.1093/nar/gkp351.

(27) Olden, J. D.; Jackson, D. A. Illuminating the “Black Box”: A Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks. *Ecol. Model.* **2002**, *154* (1–2), 135–150. https://doi.org/10.1016/S0304-3800(02)00064-9.

(28) White, H. Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. In *IEEE 1988 International Conference on Neural Networks*; **1988**; pp 451–458 vol.2. https://doi.org/10.1109/ICNN.1988.23959.

(29) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(30) Brunner, G.; Konrad, A.; Wang, Y.; Wattenhofer, R. MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. *arXiv:1809.07600* **2018**.

(31) Hermundstad, A. M.; Brown, K. S.; Bassett, D. S.; Carlson, J. M. Learning, Memory, and the Role of Neural Network Architecture. *PLOS Comput. Biol.* **2011**, *7* (6), e1002063. https://doi.org/10.1371/journal.pcbi.1002063.

(32) Andrews, R.; Diederich, J.; Tickle, A. B. Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. *Knowl.-Based Syst.* **1995**, *8* (6), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4.

(33) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313* (5786), 504–507. https://doi.org/10.1126/science.1127647.

(34) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114* **2013**.

(35) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J. Chem. Phys.* **2018**, *148* (24), 241703. https://doi.org/10.1063/1.5011399.

(36) Beerten, J.; Van Durme, J.; Gallardo, R.; Capriotti, E.; Serpell, L.; Rousseau, F.; Schymkowitz, J. WALTZ-DB: A Benchmark Database of Amyloidogenic Hexapeptides. *Bioinformatics* **2015**, *31* (10), 1698–1700. https://doi.org/10.1093/bioinformatics/btv027.

(37) Reches, M.; Porat, Y.; Gazit, E. Amyloid Fibril Formation by Pentapeptide and Tetrapeptide Fragments of Human Calcitonin. *J. Biol. Chem.* **2002**, *277* (38), 35475–35480. https://doi.org/10.1074/jbc.M206039200.

(38) Reches, M.; Gazit, E. Amyloidogenic Hexapeptide Fragment of Medin: Homology to Functional Islet Amyloid Polypeptide Fragments. *Amyloid* **2004**, *11* (2), 81–89. https://doi.org/10.1080/13506120412331272287.

(39) Do, T. D.; de Almeida, N. E. C.; LaPointe, N. E.; Chamas, A.; Feinstein, S. C.; Bowers, M. T. Amino Acid Metaclusters: Implications of Growth Trends on Peptide Self-Assembly and Structure. *Anal. Chem.* **2016**, *88* (1), 868–876. https://doi.org/10.1021/acs.analchem.5b03454.

(40) Schymkowitz, J.; Rousseau, F. Peptide sequences | WALTZ-DB http://waltzdb.switchlab.org/ (accessed Jan 24, **2019**).

(41) Kawashima, S.; Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **2000**, *28* (1), 374.

(42) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2008**, *36* (suppl_1), D202–D205. https://doi.org/10.1093/nar/gkm998.

(43) Tomii, K.; Kanehisa, M. Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins. *Protein Eng.* **1996**, *9* (1), 27–36.

(44) Fauchere, Jean Luc; Pliska, Vladimir. Hydrophobic Parameters π of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino Acid Amides. *Eur. J. Med. Chem.* **1983**, *18* (3), 369–375.

(45) Kemper, P. R.; Dupuis, N. F.; Bowers, M. T. A New, Higher Resolution, Ion Mobility Mass Spectrometer. *Int. J. Mass Spectrom.* **2009**, *287* (1–3), 46–57. https://doi.org/10.1016/j.ijms.2009.01.012.

(46) Gidden, J.; Ferzoco, A.; Baker, E. S.; Bowers, M. T. Duplex Formation and the Onset of Helicity in Poly d(CG)n Oligonucleotides in a Solvent-Free Environment. *J. Am. Chem. Soc.* **2004**, *126* (46), 15132–15140. https://doi.org/10.1021/ja046433+.

(47) Mason, E. A.; McDaniel, E. W. *Transport Properties of Ions in Gases*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, **1988**. https://doi.org/10.1002/3527602852.

(48) Chollet, F.; Others. *Keras*; **2015**.

(49) Martín Abadi; Ashish Agarwal; Paul Barham; Eugene Brevdo; Zhifeng Chen; Craig Citro; Greg S. Corrado; Andy Davis; Jeffrey Dean; Matthieu Devin; et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.

(50) West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R.; Hecht, M. H. De Novo Amyloid Proteins from Designed Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (20), 11211–11216.

(51) Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. A Database of 660 Peptide Ion Cross Sections: Use of Intrinsic Size Parameters for Bona Fide Predictions of Cross Sections. *J. Am. Soc. Mass Spectrom.* **1999**, *10* (11), 1188–1211. https://doi.org/10.1016/S1044-0305(99)00079-3.

(52) Dilger, J. M.; Glover, M. S.; Clemmer, D. E. A Database of Transition-Metal-Coordinated Peptide Cross-Sections: Selective Interaction with Specific Amino Acid Residues. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (7), 1293–1303. https://doi.org/10.1007/s13361-016-1592-9.

(53) Counterman, A. E.; Clemmer, D. E. Volumes of Individual Amino Acid Residues in Gas-Phase Peptide Ions. *J. Am. Chem. Soc.* **1999**, *121* (16), 4031–4039. https://doi.org/10.1021/ja984344p.

(54) Fink, A. L. Natively Unfolded Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15* (1), 35–41. https://doi.org/10.1016/j.sbi.2005.01.002.

(55) Uversky, V. N. Targeting Intrinsically Disordered Proteins in Neurodegenerative and Protein Dysfunction Diseases: Another Illustration of the D2 Concept. *Expert Rev. Proteomics* **2010**, *7* (4), 543–564. https://doi.org/10.1586/epr.10.36.