UNIVERSITY OF CALIFORNIA

Santa Barbara

Understanding of Amyloid Formation Using Ion Mobility Mass Spectrometry and Latent Models with Variational Autoencoders

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Chemistry

by

Michael Joaquin Tro

Committee in charge:

Professor Michael T Bowers, Chair

Professor Joan-Emma Shea

Professor Mattanjah de Vries

Professor Stuart Feinstein

March 2020

The dissertation of Michael Joaquin Tro is approved.

_____

Joan-Emma Shea

_____

Mattanjah de Vries

_____

Stuart Feinstein

_____

Professor Michael T Bowers, Committee Chair

March 2020

Understanding of Amyloid Formation Using Ion Mobility Mass Spectrometry and Latent Models with Variational Autoencoders

ACKNOWLEDGEMENTS

Thank you to my amazing wife for always supporting me, to my family for being here and feeding me many dinners, my friends who have been there with me throughout graduate school, and, of course, Mike, Joan, and Thomas for teaching me everything I know about how to do research. I couldn't have done it without all their help.

VITA OF MICHAEL JOAQUIN TRO

March 2020

EDUCATION

-Bachelor of Science in Chemistry with a Minor in Computer Science, Santa Clara University, Santa Clara, June 2014

-Doctor of Philosophy in Chemistry and Biochemistry, University of California, Santa Barbara, March 2020 (expected)

PROFESSIONAL EMPLOYMENT

-Undergraduate Researcher, Santa Clara University, January 2011 – June 2014

Advisor: Brian McNelis

Topics: Organic synthesis, Organic solar cells

-Graduate Researcher, University of California, Santa Barbara, August 2014-present

Advisor: Michael T. Bowers

Dissertation: Understanding of Amyloid Formation Using Ion Mobility Mass Spectrometry and Latent Models with Variational Autoencoders

UCSB Teaching Assistant:

-Chemical Thermodynamics (Fall 2019)

-Quantum theory and Spectroscopy (Winter 2016, 2017, 2019, and 2020)

-Statistical Mechanics and Kinetic Theory of Gases (Spring 2018, and 2019)

-General Chemistry Lab (Fall, Winter, and Spring 2014-2015)

-Physical Chemistry Lab (Spring 2017)

PUBLICATIONS

-Michael J Tro, et al. 2015. "Structure-Function Relationships of Fullerene Esters in Polymer Solar Cells: Unexpected Structural Effects on Lifetime and Efficiency: Structure-Function Relationships of Fullerenes in Polymer Solar Cells." *International Journal of Energy Research*

https://doi.org/10.1002/er.3463.

-Almeida, Natália E. C. de, Thanh D. Do, Michael J. Tro, Nichole E. LaPointe, Stuart C. Feinstein, Joan-Emma Shea, and Michael T. Bowers. 2016. "Opposing Effects of Cucurbit[7]Uril and 1,2,3,4,6-Penta- *O* -Galloyl-β- D -Glucopyranose on Amyloid β_{25-35} Assembly." *ACS Chemical Neuroscience* 7 (2): 218–26. https://doi.org/10.1021/acschemneuro.5b00280.

-Almeida, Natália E.C. de, Thanh D. Do, Nichole E. LaPointe, Michael J. Tro, Stuart C. Feinstein, Joan-Emma Shea, and Michael T. Bowers. 2017. "1,2,3,4,6-Penta-O-Galloyl-β-d-Glucopyranose Binds to the N-Terminal Metal Binding Region to Inhibit Amyloid β-Protein Oligomer and Fibril Formation." *International Journal of Mass Spectrometry* 420 (September): 24–34. https://doi.org/10.1016/j.ijms.2016.09.018.

-Do, Thanh D., James W. Checco, Michael J. Tro, Joan-Emma Shea, Michael T. Bowers, and Jonathan V. Sweedler. 2018. "Conformational Investigation of the Structure–Activity Relationship of GdFFD and Its Analogues on an Achatin-like Neuropeptide Receptor of *Aplysia Californica* Involved in the Feeding Circuit." *Physical Chemistry Chemical Physics* 20 (34): 22047–57. https://doi.org/10.1039/C8CP03661F.

-Michael J Tro*, Nate Charest*, Zachary Taitz, Joan Shea, and Michael T Bowers. 2019. "The Classifying Autoencoder: Gaining Insight to Amyloid Assembly." *Journal of Physical Chemistry B* 23 (25), 5256-5264.

https://doi.org/10.1021/acs.jpcb.9b03415

-Nate Charest*, Michael J Tro*, Michael T Bowers, Joan Shea. 2019. "Variational Autoencoders: Applications in the Analysis of Molecular Dynamics Data" *(In Perperation)*

AWARDS

-Outstanding Service to the Department during the 2018-2019 academic year

-Outstanding Service to the Department during the 2017-2018 academic year

PRESENTATIONS

-Michael J. Tro, Thanh Do, Nikki LaPointe, Natalia E. C. De Almeida, and Michael T. Bowers. "Amyloid Aggregation Prediction in Peptides." Presented by Tro at Conference of Ion Chemistry and Mass Spectrometry At Lake Arrowhead January 2017

-Michael J. Tro, Nate Charest, Thanh Do, Joan Shea, and Michael T. Bowers. "A Neural Network Approach for Amyloid Recognition" Presented by Tro at Conference of Ion Chemistry and Mass Spectrometry At Lake Arrowhead January 2018

* These authors contributed equally to the work presented here

-Michael J. Tro, Nate Charest, Joan E. Shea, and Michael T. Bowers. “Insights Into Amyloid Formation: Automatic Order Parameter Detection.” Presented by Tro at Conference of Ion Chemistry and Mass Spectrometry At Lake Arrowhead January 2019

-Michael J. Tro, Nate Charest, Joan E. Shea, and Michael T. Bowers. “Amino Acid Packing and its Effect on Amyloid Formation” Presented by Tro at Gordon Research Conference February 2019

ABSTRACT

Understanding of Amyloid Formation Using Ion Mobility Mass Spectrometry and Latent Models with Variational Autoencoders

by

Michael Joaquin Tro

Amyloid fibrils are a solid composed of proteins or peptides arranged in what is known as a cross beta pattern. That is, the fibril is made of beta sheets where the peptide backbone is perpendicular to the fibril axis. This structure has been associated with several difficult to treat diseases including Alzheimer's Disease, type II diabetes, and amyotrophic lateral sclerosis. The characterization of the formation of these fibrils remains poorly understood at a fundamental level.

First, I consider the understanding of the primary structure – activity relationship for amyloid-forming peptides. Here I use a neural-network based method of analysis: the classifying autoencoder (CAE). I demonstrate its capabilities by applying the technique to an experimental database (the Waltz database) to provide insight into a novel descriptor, dimeric isotropic deviation — an experimental measure of the aggregation properties of the amino acids. I find correlation between dimeric isotropic deviation and the failure to form amyloids when hydrophobic effects are not a primary driving force in amyloid formation.

Next, I consider the formation of amyloids from the perspective of a molecular dynamics simulation. I use a similar technique as above to analyze molecular dynamics simulations of amyloid formation. Here we the technique is applied to the internal coordinates of a coarse-grained molecular dynamics simulation of amyloid formation. The method is shown to be able to reduce the ensemble of data to a single variable that tracks evolution in the system and successfully characterizes large-scale system

evolutions with precision exceeding or comparable to more conventional order parameters. In addition, we show it can be used to identify the features of the system which best track the evolution of the system and be used to automatically detect the nucleus of the aggregating system.

Table of Contents

I. Introduction

"Biology today is moving in the direction of chemistry. Much of what is understood in the field is based on the structure of molecules and the properties of molecules in relation to their structure. If you have that basis, then biology isn't just a collection of disconnected facts."

Linus Pauling (Campbell 1986)

A. Proteins

Proteins are the molecular work horses of biology. They are the result of over 4.54 billion years of chemical (and, of course, biological) evolution and are responsible for nearly all of the most remarkable accomplishments of molecular biology. From some of the simplest parts of our bodies, our hair and fingernails, to complex the complex chemical machines involved in DNA replication, proteins are the molecules that allow biology to perform efficient, complex chemistry. The key to the success of proteins is the structure of the protein molecule; as the above Linus Pauling quote states: structure determines properties, and this is the key to understanding molecular biology.

A protein is, itself, a heteropolymer made up of many amino acid monomers. That is, the protein is a chain of linked molecules where each link in the chain can be any amino acid. While there are twenty canonical amino acids, there also exist some unusual amino acids. In this thesis we will only consider the twenty.

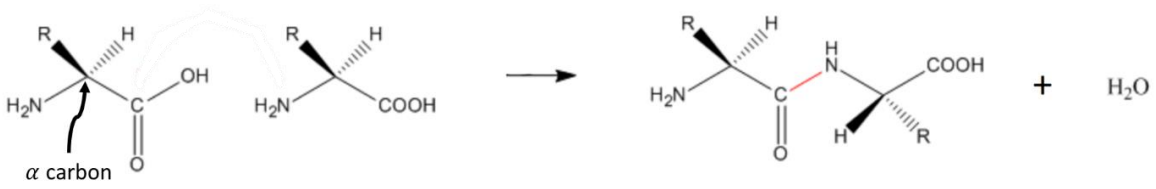


Figure I-1 Basic representation of the peptide bond which links two amino acids in a protein. Here the R group is defined by which amino acid which is represented. The α carbon of only the first reactant has been highlighted. On the reactants side of the above equation the peptide bond is highlighted in red. ("Amino Acids and Proteins - Biological Molecules - MCAT Review" n.d.)

Error! Reference source not found. depicts a basic representation of the bond involved between two amino acids, the peptide bond. In this depiction R can be any of the 20 amino acids[†]. The carbon to which the side chain is bonded is referred to as the α carbon. Almost all of the 20 canonical amino acids share this backbone structure, with the exception of proline where an additional bond from the the R group comes back to bond to the amine, resulting in a five-member-ring including the nitrogen, the alpha carbon, and three other carbons. This has important implications since it constrains the torsion angles that may be accessed by the nitrogen-α carbon bond.

Additionally, the α carbon is a chiral center. For almost all the 20 canonical amino acids the vast majority of them are found in the L-isoform. One exception to this is glycine, which has two hydrogens bonded to the α carbon and is not a chiral center.

Often the length and complexity of these chains of amino acids is implied by the word we use to describe the chain. A shorter chains of amino acids Amino acids can be chained together in small numbers (< ~50) to form peptides, while longer chains may be called polypeptides. Proteins are made up of one more polypeptides, and serve some biological function.

[†]A table of all 20 amino acids can be found in the appendix

The structure of a protein is separated in to four levels, primary, secondary, tertiary, and quaternary structure. The primary structure is the sequence of amino acids that make up the protein. Often, and for the rest of this thesis, a peptide's primary structure is represented by each amino acid's three letter or single letter code starting from the unbound amino group (the n-terminus) to the unbound carboxylic acid group (the c-terminus).

The secondary structure refers to the local structure along the chain. There are two major classes of secondary structure, α-helix, and β -sheet. The α-helix, as the name implies, is a helix, often with three to four amino acids per full rotation of the helix. In this secondary structure, the side chains are on the outside of the helix, and the carbon back is on the inside of the helix. A major driving force in the stability of an α-helix is hydrogen bonding formed between the carbonyl group of one amino acid and the amino group of an amino acid right next to it in the helix, four residues away in primary structure. The β-sheet on the other hand is a relatively straight region requiring at least two strands to run parallel to each other. This is driven by hydrogen bonding, again between carbonyl and amino groups between the two chains, but here between the two strands. The result is that the side chains extend roughly parallel to the normal plane of the beta sheet, and each consecutive side chain in the primary sequence extends on opposite sides of the sheet. There are different varieties of these major classes of secondary structure, but they are out of the scope of this thesis. Notably, some peptides may lack secondary structure, and can will be referred to as random coil. These are particularly important to the work in this thesis, since their lack of structure leaves them free to adopt problematic structures.

While tertiary and quaternary structure are not as pertinent to this thesis, they are briefly described here for completeness sake. The tertiary structure refers to how the chains fold in on them selves in order to form a three-dimensional molecule. Finally, the quaternary structure is the interaction between two three dimensional structures to form a fully functioning protein.

The final structure of a proteins varies wildly, from the relatively simple structure of ubiquitin, basically a tangled ball (Smith et al. 2019) which, as can be inferred from its name, is found ubiquitously, and serves many different functions throughout the cell (Komander and Rape 2012). To the exotic and beautifully complex pilus machine, used in bacteria mobility and surface adhesion (Chang et al. 2016). These two proteins are visualized in Figure I-2.

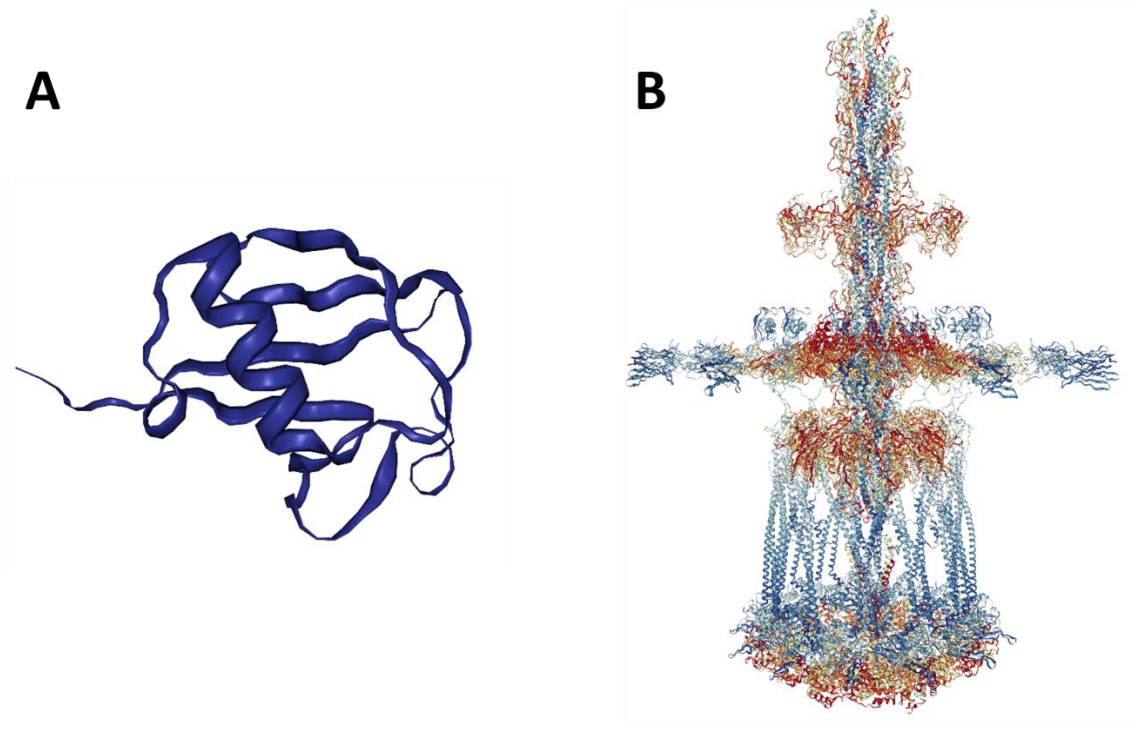


Figure I-2 A. Cartoon depiction of ubiquitin take from an NMR study. B. Assembly of pilus machine determined with electron microscopy. Here the color scheme is only to help the eye make sense of the complicated structure.

The mechanism by which the protein reaches these final structures can be as complex as the structure is itself, and even perplexed scientists for some time. What came to be known as the Leventhal Paradox stated that even a peptide with only 150 amino acids could have 3^{100} possible configurations (based on the backbone bonds, and their low energy conformations), and unless there was some directed folding mechanism, that peptide could never randomly find its native state (Levinthal 1969). While protein folding remains an active area of research (Elías-Villalobos et al.

2019; Puchades, Sandate, and Lander 2020), epically when considering refolding and chaperone assisted folding schemes, it is generally accepted that the solution to Leventhal's paradox is a funnel shaped energy landscape, as depicted in Figure I-3. When a protein starts in an unfolded state, it will descend through the contours of that energy landscape until it reaches a global minimum, which should be its native state.

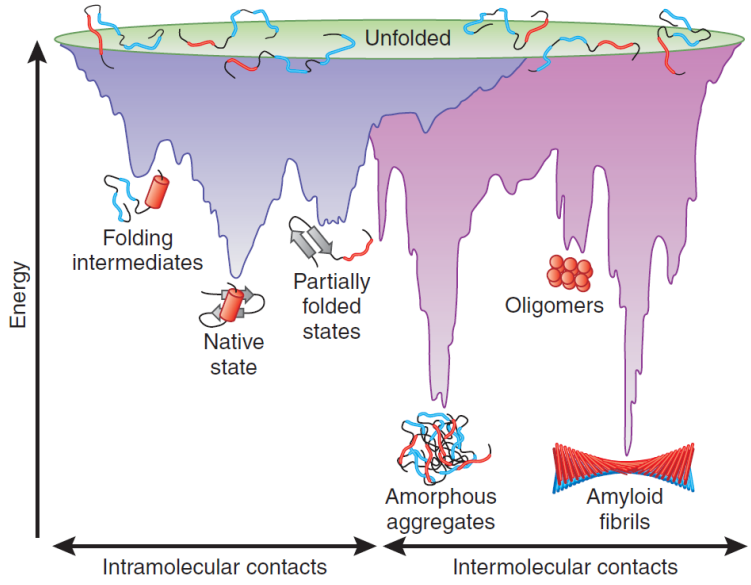


Figure I-3 A funnel shaped energy landscape. If the protein is unfolded and by itself it will descend the energy landscape of the funnel until it reaches its final state. However, also depicted here are competing states that are a result of inter molecular interactions with other proteins.

However, Figure I-3 also depicts possible metastable states, and competing states that are the result of intermolecular interactions. These other states complicate our understanding of the folding mechanism of the proteins, and it has even been hypothesized that most proteins are metastable, and in fact the global minimum of proteins are aggregate structures (Thirumalai and Reddy 2011). One of the most common, and important, protein aggregate structures is known as the amyloid fibril. Which will be discussed in detail in the next section.

B. Amyloids

Due to the complicated nature of amyloids, the term itself can sometimes be used in a variety of circumstances. I will keep to the nomenclature set out by the International Society of Amyloidosis (ISA) (Benson et al. 2019). Here an amyloid is biological deposit consisting of mostly protein, where a large amount of that protein is a fibril which has adopted what is known as a "cross-β" structure. The structure of the fibril is shown in Figure I-4. The fibril itself can range from about 60-200 Å in diameter (Makin and Serpell 2002; Fitzpatrick et al. 2013; Sipe and Cohen 2000). The peptide chains in the beta sheets run perpendicular to the fibril, as does the normal plane of the beta sheet. This results in the x-ray diffraction pattern shown in Figure I-4 A: an intersheet spacing of ~10 Å which is perpendicular to a interstrand spacing of ~4.8 Å.

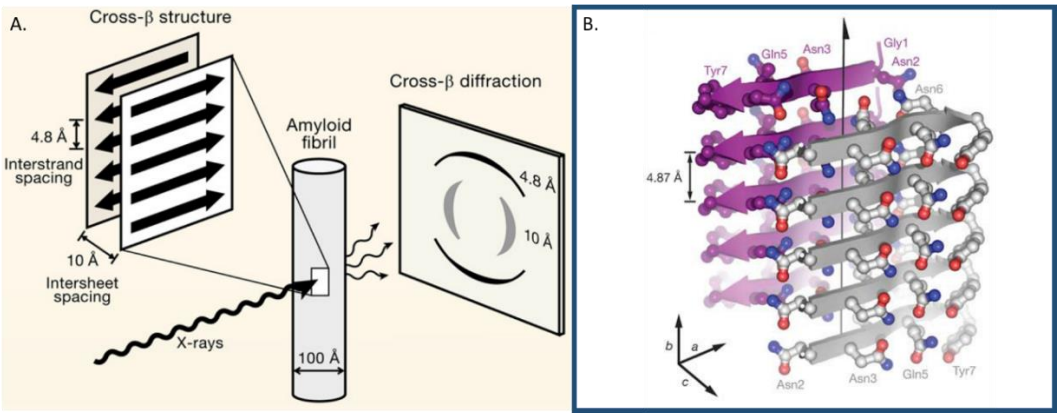


Figure I-4 Left: a cartoon drawing of an amyloid fibril during an x-ray diffraction experiment (Eisenberg and Jucker 2012). Right: single crystal x-ray diffraction results of an amyloid fibril. The primary structure of this peptide is GNNQQNY, a peptide derived from the yeast protein Sup35 (Nelson et al. 2005).

As suggested in Figure I-3 this structure can be a lower energy state then the protein in its native state, and, again, some have even suggested this to be universal of all proteins (Thirumalai and Reddy 2011). This phenomenon is interesting for a number of reasons as will be disused in further detail in the next two sections.

i. Amyloid Diseases

The amyloid state has been under intense study due to its complex relationship to disease. Many diseases have been found to be associated with disease, yet there are varying degrees of understanding when it comes to the disease mechanism.

In amyloidosis the fibrils are the cause the disease. There are two general forms of amyloidosis, systematic and localized. As the name suggest, in systematic amyloidosis amyloids can be found in the whole body or rather, amyloids can be found in many parts of the body. In localized amyloidosis the amyloids are found in a localized region. In this class of diseases, it is known that the fibril itself is the pathogenic agent. One example is AA amyloidosis in which the protein amyloid A, derived from the protein serum amyloid A, forms amyloids in across the body (Benson et al. 2019).

Alternatively, diseases in which the fibrils are not necessarily the pathogenic species exist as well, this includes Alzheimer's, where clearly some relationship between the disease and the fibril exists, but the fibril itself can not be implicated as the main pathogenic species (Sakono and Zako 2010; Lesné et al. 2013; Dodart et al. 2002). In the case of Alzheimer's it is suspected that the toxic species is an intermediate in the fibril formation process (Bernstein et al. 2009; Sakono and Zako 2010).

While much of the conversation on amyloids is centered on disease (for good reason), it is also interesting to note that amyloids are not always toxic.

ii. Functional Amyloids

Due to the unique properties of amyloids, nature has found a use for them. Interestingly the aspects of amyloids that make the resulting diseases difficult to treat are also the same aspects that have made them a useful tool for nature. That is the strong hydrogen bonds between the strands, their unique morphology, and their ability to aggregate are all aspects which I will show are used by organisms to their benefit (Fowler et al. 2007).

With respect to the strength of the hydrogen bonded back bone, an interesting example is spider silk. Proteins known as spidroins form beta sheet rich crystalline-like structures within the spider silk. Further study of these beta sheet rich structures test positive for many amyloid tests including ThT fluorescence, Congo Red binding, and x-ray diffraction. While the strength of hydrogen bonds is one of the reasons it is so break up amyloids in disease, here that same strength lets it to contribute to the strength of spider silk (Kenney et al. 2002).

Another, interesting example of a functional amyloid is in the synthesis of melanin (Fowler et al. 2005). The protein Pmel17, forms amyloid fibrils which both sequester toxic melanin intermediates during the synthesis, and eventually template the formation of melanin polymer along the amyloid fibril. Here both the chemical and morphological properties are taken advantage of in order to carry out the synthesis. Interestingly this whole process is carried out inside a vesical in the cell, where the process is carefully moderated. While the amyloid in this case is being used by the cell, the cell also takes care to sequester the fibrils to protect itself from the negative effects of amyloids (Fowler et al. 2005).

Lastly, and one of my favorite examples functional amyloids, is the way Escherichia coli creates an extra cellular matrix. It does this by secreting a peptide called CsgA. These proteins will eventually be the monomers in the amyloid assembly process. In order to control the formation of

these amyloid fibrils, however, the cell also produces a protein, CsgB, a membrane bound protein on the outside of the cell. This protein nucleates the fibril formation; CsgB will template the first unit of CsgA which will promote assembly of more CsgA on top of it (Fowler et al. 2007).

In all these cases nature has figured out a way to not only found a way to survive alongside amyloid fibrils, but to use them to their benefit. This raises two interesting questions for us as researchers. The first is how do these organisms protect themselves? Perhaps by studying these systems we may learn new ways to combat amyloid diseases. The second question is, is there a way we can use amyloids to our advantage much like nature has. The materials research has long been looking at self-organization to create new materials, and the idea of using amyloid systems has attracted some interest. For all these applications, however, a better fundamental understanding of amyloid is key.

C. Conclusions

In conclusion, amyloids are important. They are important due to their disease applications, of course, but also to biology in general. Furthermore, they may yet have yet unknown uses to us. Despite this, there remains much unknown to us about this structure. We have yet to invent a good way to stop many of the diseases that are related to amyloids. Despite massive efforts to solve the Alzheimer's problem (Cummings, Reiber, and Kumar 2018) the problem remains unsolved, and recently the pharmaceutical industry has started pulling back on Alzheimer's research (*BBC News* 2018). This only highlights the need for a fundamental understanding of the disease pathogenesis, including understanding of fibril formation.

One surprising aspect missing from this understanding of amyloid fibrils, is the ability to predict amyloid propensity of a protein, or section of protein, to form amyloids using only the primary sequence of the protein. There have been many attempts (Walsh et al. 2014; Tartaglia and

Vendruscolo 2008; Thompson et al. 2006; Do et al. 2016), but there has been failure to find consensus on a method which works very well. There, in fact, has even been a trend to use as many different prediction algorithms as possible in order to come up with some meta method, but even these are not perfect (Emily, Talvas, and Delamarche 2013; Tsolis et al. 2013).

Since amyloids seem to show up in so many different and complex contexts, some method to learn amyloid propensity from primary sequence with high throughput would be highly valuable. For example, one could scan the protein data base to find amyloid forming hot spots to find new amyloid forming proteins that may or may not be involved in disease. One could scan a known fibril forming peptide and find the amyloid hot spot, and design blockers for that specific region of peptide in order to stop or reverse disease progress. In addition, this could be a way to help design amyloid materials. Finally, the method itself may reveal aspects of amyloids that are yet not understood, which could lead to yet further unexpected benefits.

D. References

"Amino Acids and Proteins - Biological Molecules - MCAT Review." n.d. Accessed January 7, 2020. http://mcat-review.org/amino-acids-proteins.php.

BBC News. 2018. "Pharma Giant Ending Alzheimer's Research," January 10, 2018, sec. Health. https://www.bbc.com/news/health-42633871.

Benson, Merrill D., Joel N. Buxbaum, David S. Eisenberg, Giampaolo Merlini, Maria J. M. Saraiva, Yoshiki Sekijima, Jean D. Sipe, and Per Westermark. 2019. "Amyloid Nomenclature 2018: Recommendations by the International Society of Amyloidosis (ISA) Nomenclature Committee." *Amyloid* 0 (0): 1–5. https://doi.org/10.1080/13506129.2018.1549825.

Bernstein, Summer L., Nicholas F. Dupuis, Noel D. Lazo, Thomas Wyttenbach, Margaret M. Condron, Gal Bitan, David B. Teplow, et al. 2009. "Amyloid-β Protein Oligomerization and the Importance of Tetramers and Dodecamers in the Aetiology of Alzheimer's Disease." *Nature Chemistry* 1 (4): 326–31. https://doi.org/10.1038/nchem.247.

Campbell, Neil A. 1986. "Crossing the Boundaries of ScienceApplying Chemistry and Physics to Study Biological Molecules, Linus Pauling Has Shown Us That No Discipline Is an Island." *BioScience* 36 (11): 737–39. https://doi.org/10.2307/1310282.

Chang, Yi-Wei, Lee A. Rettberg, Anke Treuner-Lange, Janet Iwasa, Lotte Søgaard-Andersen, and Grant J. Jensen. 2016. “Architecture of the Type IVa Pilus Machine.” *Science* 351 (6278). https://doi.org/10.1126/science.aad2001.

Cummings, Jeffrey, Carl Reiber, and Parvesh Kumar. 2018. “The Price of Progress: Funding and Financing Alzheimer’s Disease Drug Development.” *Alzheimer’s & Dementia : Translational Research & Clinical Interventions* 4 (June): 330–43. https://doi.org/10.1016/j.trci.2018.04.008.

Do, Thanh D., Natália E. C. de Almeida, Nichole E. LaPointe, Ali Chamas, Stuart C. Feinstein, and Michael T. Bowers. 2016. “Amino Acid Metaclusters: Implications of Growth Trends on Peptide Self-Assembly and Structure.” *Analytical Chemistry* 88 (1): 868–76. https://doi.org/10.1021/acs.analchem.5b03454.

Dodart, Jean-Cosme, Kelly R. Bales, Kimberley S. Gannon, Stephen J. Greene, Ronald B. DeMattos, Chantal Mathis, Cynthia A. DeLong, et al. 2002. “Immunization Reverses Memory Deficits without Reducing Brain Abeta Burden in Alzheimer’s Disease Model.” *Nature Neuroscience* 5 (5): 452–57. https://doi.org/10.1038/nn842.

Eisenberg, David, and Mathias Jucker. 2012. “The Amyloid State of Proteins in Human Diseases.” *Cell* 148 (6): 1188–1203. https://doi.org/10.1016/j.cell.2012.02.022.

Elías-Villalobos, Alberto, Damien Toullec, Céline Faux, Martial Séveno, and Dominique Helmlinger. 2019. “Chaperone-Mediated Ordered Assembly of the SAGA and NuA4 Transcription Co-Activator Complexes in Yeast.” *Nature Communications* 10 (1): 1–17. https://doi.org/10.1038/s41467-019-13243-w.

Emily, Mathieu, Anthony Talvas, and Christian Delamarche. 2013. “MetAmyl: A METa-Predictor for AMYLoid Proteins.” *PLOS ONE* 8 (11): e79722. https://doi.org/10.1371/journal.pone.0079722.

Fitzpatrick, Anthony W. P., Galia T. Debelouchina, Marvin J. Bayro, Daniel K. Clare, Marc A. Caporini, Vikram S. Bajaj, Christopher P. Jaroniec, et al. 2013. “Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril.” *Proceedings of the National Academy of Sciences* 110 (14): 5468–73. https://doi.org/10.1073/pnas.1219476110.

Fowler, Douglas M., Atanas V. Koulov, Christelle Alory-Jost, Michael S. Marks, William E. Balch, and Jeffery W. Kelly. 2005. “Functional Amyloid Formation within Mammalian Tissue.” *PLOS Biology* 4 (1): e6. https://doi.org/10.1371/journal.pbio.0040006.

Fowler, Douglas M., Atanas V. Koulov, William E. Balch, and Jeffery W. Kelly. 2007. “Functional Amyloid – from Bacteria to Humans.” *Trends in Biochemical Sciences* 32 (5): 217–24. https://doi.org/10.1016/j.tibs.2007.03.003.

Kenney, John M., David Knight, Michael J. Wise, and Fritz Vollrath. 2002. “Amyloidogenic Nature of Spider Silk.” *European Journal of Biochemistry* 269 (16): 4159–63. https://doi.org/10.1046/j.1432-1033.2002.03112.x.

Komander, David, and Michael Rape. 2012. “The Ubiquitin Code.” *Annual Review of Biochemistry* 81 (1): 203–29. https://doi.org/10.1146/annurev-biochem-060310-170328.

Lesné, Sylvain E., Mathew A. Sherman, Marianne Grant, Michael Kuskowski, Julie A. Schneider, David A. Bennett, and Karen H. Ashe. 2013. “Brain Amyloid-β Oligomers in Ageing and Alzheimer’s Disease.” *Brain: A Journal of Neurology* 136 (Pt 5): 1383–98. https://doi.org/10.1093/brain/awt062.

Levinthal, Cyrus. 1969. “How to Fold Graciously.” *University of Illinois Press (1969)* Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois: Pages 22-24.

Makin, O. S., and L. C. Serpell. 2002. “Examining the Structure of the Mature Amyloid Fibril.” *Biochemical Society Transactions* 30 (4): 521–25. https://doi.org/10.1042/.

Nelson, Rebecca, Michael R. Sawaya, Melinda Balbirnie, Anders Ø Madsen, Christian Riekel, Robert Grothe, and David Eisenberg. 2005. “Structure of the Cross-β Spine of Amyloid-like Fibrils.” *Nature* 435 (7043): 773–78. https://doi.org/10.1038/nature03680.

“Protein Structure | BioNinja.” n.d. Accessed January 9, 2020. https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html.

Puchades, Cristina, Colby R. Sandate, and Gabriel C. Lander. 2020. “The Molecular Principles Governing the Activity and Functional Diversity of AAA+ Proteins.” *Nature Reviews Molecular Cell Biology* 21 (1): 43–58. https://doi.org/10.1038/s41580-019-0183-6.

Sakono, Masafumi, and Tamotsu Zako. 2010. “Amyloid Oligomers: Formation and Toxicity of Abeta Oligomers.” *The FEBS Journal* 277 (6): 1348–58. https://doi.org/10.1111/j.1742-4658.2010.07568.x.

“Secondary Structure Analysis - Pronalyse.” n.d. Accessed January 9, 2020. https://www.creative-proteomics.com/pronalyse/secondary-structure-analysis.html.

Sipe, Jean D., and Alan S. Cohen. 2000. “Review: History of the Amyloid Fibril.” *Journal of Structural Biology* 130 (2): 88–98. https://doi.org/10.1006/jsbi.2000.4221.

Smith, Colin A., Adam Mazur, Ashok K. Rout, Stefan Becker, Donghan Lee, Bert L. de Groot, and Christian Griesinger. 2019. “Enhancing NMR Derived Ensembles with Kinetics on Multiple Timescales.” *Journal of Biomolecular NMR*, December. https://doi.org/10.1007/s10858-019-00288-8.

Tartaglia, Gian Gaetano, and Michele Vendruscolo. 2008. “The Zyggregator Method for Predicting Protein Aggregation Propensities” 37 (7): 1395–1401. https://doi.org/10.1039/B706784B.

Thirumalai, D., and G. Reddy. 2011. “Are Native Proteins Metastable?” *Nature Chemistry* 3 (12): 910–11. https://doi.org/10.1038/nchem.1207.

Thompson, Michael J., Stuart A. Sievers, John Karanicolas, Magdalena I. Ivanova, David Baker, and David Eisenberg. 2006. “The 3D Profile Method for Identifying Fibril-Forming Segments of Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (11): 4074–78. https://doi.org/10.1073/pnas.0511295103.

Tsolis, Antonios C., Nikos C. Papandreou, Vassiliki A. Iconomidou, and Stavros J. Hamodrakas. 2013. “A Consensus Method for the Prediction of ‘Aggregation-Prone’ Peptides in Globular Proteins.” *PLOS ONE* 8 (1): e54175. https://doi.org/10.1371/journal.pone.0054175.

Walsh, Ian, Flavio Seno, Silvio C.E. Tosatto, and Antonio Trovato. 2014. “PASTA 2.0: An Improved Server for Protein Aggregation Prediction.” *Nucleic Acids Research* 42 (Web Server issue): W301–7. https://doi.org/10.1093/nar/gku399.

II. Instrumentation

A. Ion Mobility Mass Spectrometry

Mass spectrometry is one of the most widely used analytical techniques in chemistry. All mass spectrometry is based fundamentally on the force experienced by a charged particle in an electric field. Specifically, the acceleration of a charged particle in an electric field can be written by equating the force acting on a charged particle in an electric field to the algebraic version of newtons second law resulting the following equation:

$$a = \frac{q * E}{m}$$

Here a is acceleration, q is the charge of the ion, m is the mass of the ion, and E is the strength of the electric field. For singly charged particles this relationship means all we need to calculate the mass of an ion is to measure how long it takes for that ion to be accelerated through an electric field. Since we can control the electric field this should give us all we need to know about the ion to calculate its mass. However, not all ions are singly charged. Unfortunately, two ions with different masses are equally accelerated by an electric field if those ions charge differs by the same proportion as the masses. In other words if one ion has twice the mass and twice the charge of another ion they will have the exact same acceleration as each other.

This a fundamental limitation on mass spectrometry, has spurred many solutions to gain more information from a mass spectrum. Gas chromatography, couples well with mass spectrometry, a sample may be evaporated, allowed to be separated based on its interactions with some column medium, then fed into a mass spectrometer in order to be further analyzed. This extra degree of separation may separate species which happen to have the same mass to charge ratio. In this vein of thought mass spectrometry has been coupled to as many analytical methods as

possible, from relatively common methods such as, gas chromatography (as I just mentioned), liquid chromatography, and capillary electrophoresis, to more exotic methods like resonance enhanced multi photon ionization (Haggmark et al. 2018). The mass of an ion is such a good indicator of the ion's identity, mass spectrometry has proven itself to be a powerful method of analysis.

Ion mobility is yet another axis on which to perform separation prior to mass analysis, which has been highly useful in analytical chemistry, and is used in a variety of analytical environments from boarder protection, to airport security, to food safety (Liu and Hill 2015, 5; D'Agostino and Chenier 2010; Scoville 2013, 13). As with mass spectrometry we may write a relatively simple expression in order to understand the value we are measuring. Here we measure the velocity of an ion as it is pulled through a buffer gas by an electric field:

$$v_d = K * E$$

Here v_d is the drift velocity, K is the mobility of the ion, and E is the strength of the electric field. This mobility is affected by the temperature and pressure of the buffer gas and is often converted into the reduced mobility:

$$K_0 = K \frac{P}{P_0} \frac{T_0}{T}$$

Where K_0 is the reduced mobility, K is the measured mobility, P is the measured pressure, P_0 is the standard pressure (760 torr), T is the measured pressure, and T_0 is the standard temperature 273.15 K.

It is worth noting that at high drift voltages the mobility will become field dependent as the nature of the ion-gas interaction will be changed as the ions gain energy between each collision (Revercomb and Mason 1975). As such many ion mobility experiments are done at sufficiently low voltages for the above relationships to hold true.

While there are many types of ion mobility instruments, here I will focus on a drift time instrument, as it is the only ion mobility instrument I have used during my graduate studies. Generally, in this experiment, after the ions are generated, they are gated, stored, then pulsed into a drift tube with uniform electric field, while temperature and pressure are recorded. By assuming the ions reach v_d in a negligible amount of time the mobility may be calculated by measuring the drift time, t_d, it takes for the ions to cross that electric field. Since ion mobility instruments often have further analysis (namely mass analysis) after the drift tube, the arrival time at the detector, t_a, is often convoluted by the time spent outside the drift tube, t_0 (ie. $t_a = t_0 + t_d$). By using the drift velocity v_d and the length of the drift region L. One may replace t_d in the expression for the arrival time ($t_a = t_0 + t_d$) with the appropriate values based on the reduced mobility and the pressure to voltage ratio to arrive at the following expression for arrival time.

$$t_a = t_0 + \frac{L^2 T_0}{K_0 P_0 T} \frac{P}{V}$$

Here L is the length of the drift region. In this way by plotting the arrival time against the pressure to voltage ratio one may calculate the reduced mobility independently of t_0. Note, that the pressure to voltage ratio is only changed in the drift region, and thus t_0 will be constant over the course of the experiments.

While this is useful in its own right for analytical chemistry and is the basis for ion mobility being so widely used for chemical identification. We may further learn something about the structure of the analyte by exploring the kinetic theory.

B. Experimental Collison Cross Sections

While it is useful to measure the mobility of an ion, to me this has only been a means to an end. That end is calculating the collision cross section of the ion. This way we can get some

measure of the physical size of the ion in units of Å^2. To do this, again we must be in the low field limit as described by Mason (Revercomb and Mason 1975). In this case we can relate the mobility to the collision cross section with the Mason-Schamp equation (Mason and Schamp 1958):

$$K_0 = \frac{3e}{16N_0}\left(\frac{2\pi}{\mu k_B T}\right)^{\frac{1}{2}}\frac{1}{\Omega}$$

Where K_0 is the reduced mobility, N_0 is the buffer gas number density at P_0, e is the elementary charge, μ is the reduced mass between a buffer gas molecule and the ion, k_B is the Boltzman constant, and Ω is the momentum transfer collision integral. It is this value which we can measure and which we call our experimental collision cross section.

C. Theoretical Collison Cross sections

While the ability to measure the experimental collision cross section is interesting, the power of this technique is derived from comparing that experimental value to theoretical values. There are many methods to obtain the structure of a molecule including quantum calculations, molecular dynamics, x-ray crystallography, or NMR. The ability to calculate a theoretical collision cross section and compare it to the experimental value allows us to infer atomic scale structural information about our system with ion mobility measurements.

Calculating this value from a structure, however, is not trivial, and a variety of methods to approximate this value exist. To a first approximation you may think of the cross section as the average "shadow" of the molecule as it tumbles through the buffer gas, or in other words the spherically averaged cross section. This method to calculate the cross section is used and is known as the projection approximation (PA). In this method, the collision cross section is estimated by projecting the molecule onto a randomly chosen plane and recording the area of the projection. This is done until the average value of this projection converges within an acceptable error

(Wyttenbach et al. 1997). This approximation works well for spheres, but more exotic shapes (bowls for example) are particularly problematic.

To conceptually understand why a bowl shape is so problematic it helps to consider the maximum momentum transfer possible, that is where collisions of the buffer gas hit a surface perpendicular to the direction of motion. That is if a bowl is traveling through a buffer gas bottom first, a sphere reasonably approximates this surface. The area where a buffer molecule collides with a surface perpendicular to the motion is only the very center of the bowl or the sphere. However, if the bowl is now traveling with the open side in the front, the surface area that is perpendicular to the direction of motion is similar in the center of the bowl, but in addition the rim of the bowl contributes as well. While this is an oversimplified explanation, it does simply make the case for a more sophisticated model which can account for concavity leading to more efficient momentum transfer between the buffer gas and the ion.

While PA is convenient and conceptually simple to understand, it is the least accurate of the commonly used methods to approximate the collision cross section of a given molecular structure. Two other methods commonly used are exact hard-sphere scattering (EHSS) (Shvartsburg and Jarrold 1996) and trajectory method (TJM) (Shvartsburg, Schatz, and Jarrold 1998). While each of these methods represent a step forward in being able to accurately model experiment, they each take a step backwards in calculation time. I will not go into the specifics of these models, because the next model, projection superposition approximation (PSA), is the model we use most.

In projection superposition approximation (PSA) the three-dimensional shape of the molecule is better taken into account. The method is based on the projection approximation, but that value is adjusted by a shape factor. This shape factor is calculated by comparing the surface area of the analyte to the surface area of the same molecule but with all the concave surfaces filled

in to make a purely convex shape. It has been shown that PSA generally performs comparably to TJM, but is 100-1000 times faster in computation time (Bleiholder, Wyttenbach, and Bowers 2011).

D. The High-Resolution Ion Mobility Mass Spectrometer

The instrument I used the most is a high-resolution ion mobility mass spectrometer informally known as the "high-res". This instrument features a 2 meter drift cell resulting in an ion mobility resolution of 109, and able to distinguish between ions with a 1.01% difference in mobility (Kemper, Dupuis, and Bowers 2009). In addition to the instrument's remarkable mobility resolution, it is exceedingly gentle in its treatment of ions. It was designed to minimize energizing regions in order to better look at clusters.

In general, I have tried to leave this instrument in a better then I found it. In addition to describing the details of the instrument here, I will also document the various ways I have optimized it, and my diagnosis on existing problems.

i. Ionization

The ionization source is nano electrospray ionization. In conventional electrospray ionization (which interestingly, was inspired by a car painting technique), a syringe (typically around 100 μL in volume) filled with sample and slowly injects that sample into a needle near in the inlet of the instrument. The flow rate of the syringe is often controlled by some external controller in order to improve reproducibility. An electric potential is held between the instrument and the needle. The sample dispensed from the electrospray tip as a fine mist. Inevitably some of the mist droplets will have an overall charge and will be pulled towards the instrument by the electrospray voltage. As the solvent in the droplets evaporates the high density of charge will eventually cause the droplet to burst or eject the ion from the droplet. In either case this results in soft ionization of the analyte, and can often result in multiple charges deposited on the analyte without fragmentation of the

analyte itself. Nano electrospray ionization follows a similar process, but instead of a syringe filled with sample, a glass capillary is pulled to form a needle tip and filled with sample. The whole capillary is coated in a metal to conduct the electric potential. This needle typically holds between 5 to 10 μL of sample. This offers an advantage over electrospray in the amount of sample needed, but tends to have higher variance from tip to tip. In our case we often use precious samples, where 100 μL can not be spared, nano electrospray is ideal. In addition, nano spray tends to be better at ionizing the biological molecules we typically study.

The tip on this instrument is typically held between 800-1300 V above the entrance to the instrument. Higher voltages tend to result in higher charge states, while lower voltages tend to favor cluster formation. I've spent some time trying to add some amount of back pressure by attaching a hose attached to a nitrogen take to the back of the tip. This has been able to improve signal on other instruments. While I was successful in setting up the back-pressure apparatus, the improvement in signal was negligible. This may be due to the lack of pressure controls I had available. On other instruments the pressure applied to the sample is controlled with a small pressure regulator which holds some constant pressure, and can be finely tuned. Here I used a series of valves and vents in order to control the flow of the pressurized nitrogen. I suspect this did not give adequate control on the backing pressure. Perhaps a better pressure regulator, similar to our other instrument (the "ESI") would have allowed for more precise optimization and resulted in improved signal.

The ions then enter the instrument through a small capillary. The spray voltage is with respect to this capillary. The voltage of the capillary with respect to the next segment of the instrument, the ion funnel, may be changed. I have not found this voltage to be largely important for signal optimization. In the paper describing this instrument (Kemper, Dupuis, and Bowers 2009) it says that a slightly lower voltage then the ion funnel (0-10 V) for positive ions results in a 10-20%

signal increase. While I have found this to be the case in some situations, generally I've found it has little effect on signal intensity.

ii. Source Region

The ions then reach the region source region. This region is held at about 10 torr but is precisely held slightly below the pressure in the drift region. This is done in order to make sure gas flows exclusively from the drift cell towards the source region but minimizes that flow. This ensures the no impurities flow from the source region to the drift region which must be pure helium in order to calculate cross section. Minimizing this flow makes sure that ions are not impeded by this head wind. This is done by pumping on the region with a ~5 L/s roughing pump while at the same time flowing helium into the region. The flow of the helium is controlled using a feedback loop which measures the pressure difference between the drift tube and the source region and adjusts the helium flow to maintain a specific pressure difference.

This differential was set to 0.3 Torr when I started working on this instrument. I observed that had not measured ion mobilities lower than about 2 $cm^2(Vs)^{-1}$. I suspected that this pressure differential, might be inhibiting low mobility ions from entering the drift tube as well as neutral impurities. To measure the contents of the gas in the drift tube a residual gas analyzer (RGA) is placed in in the main chamber analyzes the gas. It is assumed that the gas leaving the drift tube into the main chamber at the detector side is of the same chemical composition of the gas in the whole drift tube. I lowered the differential until I saw gases other than helium show up on the RGA (essentially just nitrogen), then raised the differential until that peak was no longer appreciable, resulting in a new differential of about 0.25 torr. In the end I did not observe any new peaks in the samples I was analyzing, but perhaps in marginal cases this may help with signal of low mobility ions.

Low mobility ions also take longer to traverse the drift cell allowing for more time for the ion cloud to diffuse radially which may also hinder detection of low mobility ions.

iii. Entrance Funnel

In the source region the ions are pulled into an hourglass shaped entrance funnel. The entrance funnel serves several purposes. The first is to radially focus the ions through a small orifice which serves to limit gas flow from the drift tube to the source region. In the second part of the ion funnel, when it opens back up on the other side of the hourglass, stores ions for pulsing into the drift tube during ion mobility experiments.

The electronics in the entrance funnel are some of the most complicated electronics on the instrument. There are two constant voltages that are always present. A DC voltage that moves ions from the source region into the drift cell, and a radio frequency (RF, 1.3 MHz) AC voltage that radially confines the ions. Higher DC voltages give better ion transmission but will eventually lead to discharge resulting in no usable signal. This voltage is largely left untouched, since discharge leaves the instrument unusable for some time. Higher AC voltages also lead to better transmission but may also cause heating of the ions resulting on more nonnative state protein structures or breakage of clusters.

The entrance funnel is also responsible for storing and pulsing the ions in ion mobility experiments. In an ion mobility experiment a wire mesh at the detector side of the entrance funnel serves as a gate to trap ions. An electric potential is applied to that gate to store ions. A pulse generator signals the electronics to lower that voltage briefly to allow an ion packet into the drift tube for a mobility experiment. In a later installment, a “trap” voltage was also introduced just on the detector side of the entrance funnel. This voltage serves to flatten the voltages on the storage side of the ion funnel so that the ions are not all clustered near the gate. When the gate is opened,

this trap voltages also pulses to push ions out of the entrance funnel. The pulses are typically set to be about 0.1 s apart and are typically 250 μs long.

One problem with this instrument I have not been able to fix is that at times there may be noise in the ATDs, and in the worst cases this results in no appreciable peaks in the ATD spectrum. Since the entrance funnel is the region that controls ion pulsing, I suspect the problem is in this region. The condition is fairly rare, making diagnosis difficult. I had suspected the gate pulsing to the problem, since poor gating would seem to cause leakage or failure to trap the ions. The gate pulsing is controlled by electronics devoted to that task. I have measured the gate voltage being fed to the box, and the pulsing being fed to the box. Both of these signals appear to be the same with or with out the ATD noise issue showing up. I also measured the voltages inside the box itself to determine if perhaps a component of the box had broken. The box actually has symmetrical halves for positive and negative ions. By comparing both halves of the box I have determined that the two components most likely to fail, the transistors and the optoisolator appear to be functioning the same. While it is possible that both components on each side of the box broke at the same time, it is unlikely, and I suspect they are working correctly.

Since the only other pulsed voltage here is the trap voltage, I suspect this to be the problem. The trap voltage was added after the paper on the instrument, so it is not documented there, but is documented in schematics of the instrument in our group files. The trap voltage pulses are controlled separately along with the other entrance funnel electronics on the instrument rack, which I have not yet been able to measure. The next time this issue arises, that is the place I would look.

iv. Drift Tube

The start of the drift region is defined with a second wire mesh directly after the gating mesh. This serves to ensure the gate voltages and the RF voltages of the funnel do not penetrate into the drift region, the field of which must be uniform to ensure accurate calculation of collision cross sections. The field in the drift tube is maintained by a series of rings connected by resistors. The drift voltage is typically held at 3000 V but is brought down to 1500 V during ion mobility measurements. Voltages higher then 3000 V are not viable with a pure helium buffer gas due to discharge.

The drift tube is held at a pressure of about 10 torr. The cell pressure is maintained by controlling the flow of helium into the drift cell with a leak valve. That helium flows out of the cell through the entrance and exit funnel. The end of the electric field of the drift cell is defined with a second wire mesh just before the exit funnel.

v. Exit Funnel

The exit funnel serves to focus the ions through a small orifice, and is thus a standard funnel shape (as opposed to the entrance funnel which is more of an hourglass shape). This serves two purposes, the first is to radially focus the ions which have spread due to diffusion as the ions go though the drift cell. This creates a conveniently behaved ion source in the main chamber which allows for modeling and design of the post drift cell components. The second function of orifice is to limit the gas flow from the drift cell to the main chamber.

In addition, it should be noted that this is the single pressure drop from the 10 torr drift cell region to the high vacuum (collision free, 1×10^{-5} torr) main chamber. This single drop in pressure minimizes the time the ions spend in a pressure region where the ions can accelerate enough to have high energy collisions with the gas. This preserves any clusters that traveled across the drift

cell. This is important because any fragmentation after the drift cell would cause misleading data. Ions which traveled as a cluster with one mass may get fragmented into two new ions. These new ions would appear to have the mobility of the parent ion but would have a new mass. This would severely convolute data and make any analysis near impossible. This instrument is remarkable in the size of ion clusters that it is capable of measuring.

Finally, I suspect this region of the instrument currently is not preforming well. I have spent a significant amount of time to come to this conclusion. The symptoms of this problem are intermittent, but problematic when manifest. This is a mass cutoff that is fairly sharp, perhaps over 10 or so mass units. At masses higher than mass cutoff there is no signal above background. Most commonly this mass cutoff will center around 1050 m/z, but can shift, even from scan to scan (over about 10 seconds). I have observed this mass cutoff reach down to about 300 m/z. The mass cut may also shift towards higher masses as well, but seldom goes away completely once it shows up. In addition, I have noticed this problem is less often observed when the instrument has been off for some time, and it is more common once the instrument has been on for some time. However, there is no consistent time frame I have observed (typically it may work for a couple hours if it has been off for a few days, but there is significant variance in this). I have also observed that the mass cutoff can be manipulated. The mass cutoff may be lowered by lowering the exit funnel DC voltage or reducing the exit funnel amplitude voltage. The mass cutoff may be raised by raising the exit funnel DC voltage or raising the exit funnel amplitude voltage. While raising these voltages can, at times, be a temporary solution, raising them too high will eventually result in discharge and all signal will be lost. In addition, high AC amplitudes in the exit funnel are undesirable since they may heat ions, and cause fragmentation of clusters after the ion mobility measurement.

I have performed numerous measurements in order to try understand this problem. Initial measurements of voltages being fed to the instrument did not show any obvious problems.

Previous experience with mass cutoffs, first lead me to suspect that this could be caused by oil on the quadrupole or faulty RF feedthroughs on the quadruple or exit funnel. Oil can, over the course of the experiment charge and can have unpredictable effects on signal. In addition, experience told us that RF feed throughs were liable to fail without a measurable change. To remedy this, I took apart the back end of the instrument from the detector flanged to and the pre quadrupole lenses. All RF feed throughs were replaced. Each piece was thoroughly washed in hexanes, toluene, and methanol and the instrument was reassembled. This unfortunately had no effect on the mass cutoff. This procedure did actually fix an issue that we had been having where we would get signal even when the source was off, commonly called dark current. This issue had been something that could be worked around as long as there was enough signal, but after cleaning we have almost no dark current. This was likely dude to oil on the detector flange.

After cleaning had no effect on the mass cutoff, I performed more detailed diagnostics to try to determine the cause of the mass cutoff. This included measuring all voltages into the instrument, during operation, but without the high drift voltage. In order to determine if these voltages were the cause of the mass cutoff, I measured each voltages with and with out the mass cutoff. I still was not able to measure and difference between the problematic state of the instrument and when the instrument is working perfectly.

During this time I also measured how each voltage effected the mass cutoff itself. This is when I noticed the dependence on the exit funnel electronics. The funnel voltages are generally not tuning voltages and are not often changed due to the risk of discharge or heating of the ions. There are not many places in the instrument where a mass cutoff as sharp as this could originate form, but the exit funnel is once of the places, and along with this observation I suspect this is the most likely source. I have not, however, been able to directly measure any problem with the funnel itself and only the evidence above leads me to suspect the exit funnel. It is possible, in my experience, that

this effect could be caused by oil on electrodes in the funnel and also that this could not be measurable.

Even without out confirmation of the exit funnel, it is still the most likely cause of the problem. The next step I would take in troubleshooting this problem, is to either clean or replace the exit funnel. Diagrams of the instrument, along with memory of its construction cannot confirm that adequate cleaning of the piece is possible. In order to best clean the funnel one would need to take the exit funnel apart, but assembly of the exit funnel requires template hardware and is tediously time consuming. Since cleaning is likely not a viable option, the next best solution is to replace the exit funnel. We have an exit funnel that will likely work. It will have to be shortened, but this is possible. Then an adapter must be constructed in order to attach it to this instrument. While this process will also be time consuming, I currently believe it is the best course of action.

vi. Main Chamber

The main chamber contains ion optics, the quadrupole, and the detector. The ion optics serve to keep ions focused and accelerate them to about 40 eV into a well-behaved beam for the quadruple to mass filter the ions. The ions then encounter another set of ion optics which focus the ions onto the detector.

The pre quadruple ion optics consist of both x and y steering, and three lenses. Each voltage may be optimized individually. While the x and y steering generally does not change much run to run, it can sometimes have small effects on signal, and they are easy to tune. Doing so is generally recommended. Each of the three pre quadrupole lenses can also be individually adjusted, but once optimized, only the third lens need be tuned. Lower voltages on this third lens favors high mass ions, while a higher voltage favors low mass ions. This lens should be tuned to optimize the for the ions of interest and should additionally be tuned for each mass during ATD collection.

After focusing the ions are mass selected with a quadrupole. The quadrupole is a commercial set up that came with the quadrupole itself and electronics to drive it (Extrel Core MS). The quadrupole is centered at ground and the mass selecting voltages are applied to it form the Extrel electronics. We interface with the Extrel electronics by supplying it a DV voltage form 0 -20 V which is interpreted by the Extrel hardware as mapping to 0-4000 amu. In addition, the capacitance of the system must be correctly tuned in order for the quadrupole to electronically resonate and function over its full mass range. Poor resonance leads to low transmission at high mass. In order to resonate the quadruple, the Extrel electronics are equipped with two variable capacitors and an electrode which gives a reading proportional to the resonance of the system. The Extrel manual contains detailed instructions on how to measure the resonance of the quadrupole. If the system cannot resonate within the range of the variable capacitors then one may lower the capacitance of the system by removing some length of wire. To increase the capacitance in the system, however, it is not feasible to add wire. While Extrel recommends adding capacitors inside the electronics box, this is inconvenient. The Extrel manual has detailed circuit diagrams of the system which show the capacitance added must be between the system and ground. By simply adding capacitors to ground at the feedthroughs we were able to add enough capacitance to the system to get in range of the variable capacitors and maximize the resonance of the system.

After mass selection the ions again encounter steering and focusing lenses. These electrodes should be optimized each run, but generally the maximum signal is quick and easy to find.

Finally, the detector is mounted on the back flange of the instrument. Two high voltage power supplies are used to power the detector. The first applies a high voltage to the conversion dynode which attracts ions and accelerates them to high energy. When the ions strike the conversion dynode, electrons are emitted. Another voltage is applied to a cascading mechanism in

the detector. The electrons emitted by the initial ion impact are accelerated to another plate where each of the high energy electrons emit yet more electrons on impact, these are then accelerated to the next plate. This is repeated several times in order to achieve a measurable current. This current is still small enough that it must quickly be amplified so that it does not get obscured by electrical noise in the room. A small preamp is attached directly to the detector flange for this purpose.

The pumping on the main chamber is relatively simple. The load on this chamber comes only from the drift cell which is normally operated at 10 torr and is separated from the main chamber by a 0.5 mm orifice. The target pressure for a mean free path on this length and time scale is about 5×10^{-5} torr. In order to do this in one step a 250 mm diffusion pumps on the chamber. An ion gauge is used to measure the pressure in this chamber and must be monitored during use. Standard operation runs the drift cell at about 10 torr which results in pressures around $2 - 3 \times 10^{-5}$ torr. Since the drift pressure is maintained by the leak valve, it is liable to drift slightly during the course the experiment. As the drift cell pressure drifts so does the load and thus pressure in the main chamber. As such it is essential that one keeps a close eye on these pressures, especially at start up, but also during the course of the experiment. The most dangerous failure mode of this instrument is if the pressure in the main chamber gets too high and the diffusion pump stalls leading to a pressure spike and causing the detector to burn out.

E. Conclusions

Ion mobility is a powerful tool. On the simplest side it a well-defined and reproduceable analytical measurement for chemical identification. More than this, however, it can give us structural information on the analytes we are studying. We have been able to apply this method to

biological systems, like the amyloid systems I described in the introduction, in order to gain a unique perspective of the assembly of these systems.

The aspect of ion mobility which makes it so uniquely suited to study aggregating systems is the ability to measure a distribution of many unique structures present in a solution, as opposed to an average structure. As I mentioned above, other common structural determination of biological molecules include X-ray crystallography and NMR. These methods are important because they give us easily interpretable atomic coordinates of the structure the molecules in the system. This is powerful when our analyte only adopts a single structure, but for dynamic systems these techniques may be problematic. In NMR only the average structure will be measured. In X-ray crystallography the sample must be crystalized into a single crystal, and a single structure is forced. As I mentioned before, these are important steps in finding starting structures. In many of the amyloid forming systems, this has led to detailed structures of the fibrils themselves but does not allow for study of the assembly process. Furthermore, there is mounting evidence that the fibrils themselves are not the toxic agent in many amyloid diseases, but the oligomers are responsible for the toxicity (Arya et al. 2019; Downey et al. 2018; Benilova, Karran, and De Strooper 2012; Sakono and Zako 2010). By using ion mobility we have been able to ionize the oligomers present in solution and isolate them for study. This allows us to measure the relative abundance of particular oligomeric species in solution and get some idea of the size of those ions. Coupled with molecular dynamics, and inferences from NMR and X-ray studies we have been able to identify and characterize the assembly and structure of the possible toxic oligomers (Downey et al. 2018; Bernstein et al. 2009; de Almeida et al. 2016; 2017).

F. References

Almeida, Natália E. C. de, Thanh D. Do, Michael Tro, Nichole E. LaPointe, Stuart C. Feinstein, Joan-Emma Shea, and Michael T. Bowers. 2016. “Opposing Effects of Cucurbit[7]Uril and 1,2,3,4,6-Penta-

O -Galloyl-β- D -Glucopyranose on Amyloid β $_{25–35}$ Assembly." *ACS Chemical Neuroscience* 7 (2): 218–26. https://doi.org/10.1021/acschemneuro.5b00280.

Almeida, Natália E.C. de, Thanh D. Do, Nichole E. LaPointe, Michael Tro, Stuart C. Feinstein, Joan-Emma Shea, and Michael T. Bowers. 2017. "1,2,3,4,6-Penta-O-Galloyl-β-d-Glucopyranose Binds to the N-Terminal Metal Binding Region to Inhibit Amyloid β-Protein Oligomer and Fibril Formation." *International Journal of Mass Spectrometry* 420 (September): 24–34. https://doi.org/10.1016/j.ijms.2016.09.018.

Arya, Shruti, Sarah L. Claud, Kristi Lazar Cantrell, and Michael T. Bowers. 2019. "Catalytic Prion-Like Cross-Talk between a Key Alzheimer's Disease Tau-Fragment R3 and the Type 2 Diabetes Peptide IAPP." *ACS Chemical Neuroscience* 10 (11): 4757–65. https://doi.org/10.1021/acschemneuro.9b00516.

Benilova, Iryna, Eric Karran, and Bart De Strooper. 2012. "The Toxic Aβ Oligomer and Alzheimer's Disease: An Emperor in Need of Clothes." *Nature Neuroscience* 15 (3): 349–57. https://doi.org/10.1038/nn.3028.

Bernstein, Summer L., Nicholas F. Dupuis, Noel D. Lazo, Thomas Wyttenbach, Margaret M. Condron, Gal Bitan, David B. Teplow, et al. 2009. "Amyloid-β Protein Oligomerization and the Importance of Tetramers and Dodecamers in the Aetiology of Alzheimer's Disease." *Nature Chemistry* 1 (4): 326–31. https://doi.org/10.1038/nchem.247.

Bleiholder, Christian, Thomas Wyttenbach, and Michael T. Bowers. 2011. "A Novel Projection Approximation Algorithm for the Fast and Accurate Computation of Molecular Collision Cross Sections (I). Method." *International Journal of Mass Spectrometry* 308 (1): 1–10. https://doi.org/10.1016/j.ijms.2011.06.014.

D'Agostino, Paul A., and Claude L. Chenier. 2010. "Desorption Electrospray Ionization Mass Spectrometric Analysis of Organophosphorus Chemical Warfare Agents Using Ion Mobility and Tandem Mass Spectrometry." *Rapid Communications in Mass Spectrometry: RCM* 24 (11): 1617–24. https://doi.org/10.1002/rcm.4547.

Downey, Matthew A., Maxwell J. Giammona, Christian A. Lang, Steven K. Buratto, Ambuj Singh, and Michael T. Bowers. 2018. "Inhibiting and Remodeling Toxic Amyloid-Beta Oligomer Formation Using a Computationally Designed Drug Molecule That Targets Alzheimer's Disease." *Journal of The American Society for Mass Spectrometry*, April. https://doi.org/10.1007/s13361-018-1975-1.

Haggmark, Michael R., Gregory Gate, Samuel Boldissar, Jacob Berenbeim, Andrzej L. Sobolewski, and Mattanjah S. de Vries. 2018. "Evidence for Competing Proton-Transfer and Hydrogen-Transfer Reactions in the S1 State of Indigo." *Chemical Physics*, Ultrafast Photoinduced Processes in Polyatomic Molecules:Electronic Structure, Dynamics and Spectroscopy (Dedicated to Wolfgang Domcke on the occasion of his 70th birthday), 515 (November): 535–42. https://doi.org/10.1016/j.chemphys.2018.09.027.

Kemper, Paul R., Nicholas F. Dupuis, and Michael T. Bowers. 2009. "A New, Higher Resolution, Ion Mobility Mass Spectrometer." *International Journal of Mass Spectrometry* 287 (1–3): 46–57. https://doi.org/10.1016/j.ijms.2009.01.012.

Liu, Wenjie, and Herbert H. Hill. 2015. “Chapter 5 - High-Performance Ion Mobility Spectrometry.” In *Comprehensive Analytical Chemistry*, edited by Yolanda Picó, 68:275–305. Advanced Mass Spectrometry for Food Safety and Quality. Elsevier. https://doi.org/10.1016/B978-0-444-63340-8.00005-4.

Mason, Edward A, and Homer W Schamp. 1958. “Mobility of Gaseous Lons in Weak Electric Fields.” *Annals of Physics* 4 (3): 233–70. https://doi.org/10.1016/0003-4916(58)90049-6.

Revercomb, H. E., and E. A. Mason. 1975. “Theory of Plasma Chromatography/Gaseous Electrophoresis. Review.” *Analytical Chemistry* 47 (7): 970–83. https://doi.org/10.1021/ac60357a043.

Sakono, Masafumi, and Tamotsu Zako. 2010. “Amyloid Oligomers: Formation and Toxicity of Abeta Oligomers.” *The FEBS Journal* 277 (6): 1348–58. https://doi.org/10.1111/j.1742-4658.2010.07568.x.

Scoville, Stanley. 2013. “Chapter 13 - Implications of Nanotechnology Safety of Sensors on Homeland Security Industries.” In *Nanotechnology Safety*, edited by Ramazan Asmatulu, 175–94. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-59438-9.00013-8.

Shvartsburg, Alexandre A., and Martin F. Jarrold. 1996. “An Exact Hard-Spheres Scattering Model for the Mobilities of Polyatomic Ions.” *Chemical Physics Letters* 261 (1–2): 86–91. https://doi.org/10.1016/0009-2614(96)00941-4.

Shvartsburg, Alexandre A., George C. Schatz, and Martin F. Jarrold. 1998. “Mobilities of Carbon Cluster Ions: Critical Importance of the Molecular Attractive Potential.” *The Journal of Chemical Physics* 108 (6): 2416. https://doi.org/10.1063/1.475625.

Wyttenbach, Thomas, Gert von Helden, Joseph J. Batka, Douglas Carlat, and Michael T. Bowers. 1997. “Effect of the Long-Range Potential on Ion Mobility Measurements.” *Journal of the American Society for Mass Spectrometry* 8 (3): 275–82. https://doi.org/10.1021/jasms.8b01013.

III. Ion mobility experiments on p53

A. Introduction

The tumor suppressor protein p53 is an important part of the body's ability to control cell growth, in fact, it is so important, it has earned the nick name of "the guardian of the genome" (Silva et al. 2014). Mutations in p53 are found in 20% of breast cancer tumors (Olivier 2006), the gene that codes for p53 was found to be the most commonly mutated gene in tumor samples by Kandoth et al (Kandoth et al. 2013), and mutations are found in half of malignant tumor samples by Silva et al (Silva et al. 2014).

When describing the function of the p53 one will often encounter an automotive analogy; that is, if one thinks of cell growth as a car then the p53 is the breaks to that car. It plays a crucial role in slowing cell growth in the case of genetic damage (Vogelstein, Lane, and Levine 2000). It does this, in part, by sliding along DNA until it detects genetic damage and then signaling the appropriate response through a diverse and complex signaling pathway. To simplify, however, the cell will try to repair damage, but if the damage is too severe p53 also plays a role in signaling for apoptosis of the cell. In this way p53 protects the body from tumors. Essentially it signals for the repair of genetic damage, but also for the destruction of the cell if that damage is too severe. When p53 is disrupted that genetic damage may go unchecked and may lead to mutations which cause uncontrolled cell growth which result in tumors.

The protein itself is a homotetramer, with each chain consisting of 393 residues and divided in to two folded domains: the DNA binding domain from residue 94-294, and the tetramerization domain from residue 323-360. The rest of the protein is intrinsically disordered, and serves a variety of functions in the cell (Tidow et al. 2007).

It has been found that mutations of p53 not only induce the amyloid state, but further, exhibit prion-like recruitment effects (Silva et al. 2014; Ano Bom et al. 2012; Levy et al. 2011; Silva et al. 2013). That is to say these mutations not only cause a protein to adopt an amyloid state, but that amyloid fibril can incorporate other p53 proteins which are not mutated and are in their native state. When in the fibril state the protein would no longer be able to protect the cell form genetic damage, and an important anticancer safeguard will be missing.

Despite experiments showing evidence of formation of amyloid fibrils and the resultant prion-like behavior of mutated p53 to induce conformational change in wild type p53, the mechanism of this fibril formation is poorly understood. In this study, a fragment (residues 231-257), and a amyloid forming mutation to that fragment (R248Q) are studied. This mutation is the most

Figure III-1 A depiction of the DNA binding domain of p53 as determined by x-ray crystallography. Highlighted in red is the arginine at the 248th position, as it inserts itself into the minor groove of the DNA. Next to it, highlighted in orange is the arginine at the 249th position.

common cancer-causing mutation, and interestingly this arginine one of the residues which makes

contact with the DNA, depicted in Figure III-1 (Cho et al. 1994). The positive charge on the arginine interacts with the negatively charged DNA backbone and inserts itself into the minor groove of the DNA. In addition, this mutation has been associated amyloid fibrils.

Ion mobility is used to study the assembly from the monomer to higher order oligomers and the relative concentrations of these species. Transmission electron microscopy (TEM) is used to determine the final states of these aggregates. Finally, replica exchange molecular dynamics (REMD) is used to determine the mutations effects on the monomer. It is found that surprisingly this mutation has little effect on these experiments. While it may seem that the removal of a charged residue must have drastic effects on the structure of the fragment, it should be noted, as shown in Figure III-1, there is a second arginine directly adjacent to the mutation. Unfortunately, there is no conclusion to this study. Here I will present what data I have and discuss my current hypotheses and what further data must be taken to find a conclusion.

Wildtype (MW: 3144.75):

Ac-TIHYNYMCNS^{240}SCMGGMNRRP^{250}ILTIITL-NH$_3$

Blue = positive charge
Yellow = beta

Mutant (MW: 3116.73):

Ac-TIHYNYMCNS^{240}SCMGGMNQRP^{250}ILTIITL-NH$_3$

Figure III-2 The sequences used in this study. Blue letters are basic residues, and the highlighted residues represent the secondary structure as calculated from x-ray crystal structures according to the DSSP definitions of secondary structure.

B. Results

The wild type of this peptides has been well characterized. While it has been found that the peptide can form amyloid fibrils. It is also remarkably disordered. Both REMD and ion mobility show a broad and complex ensemble of structures.

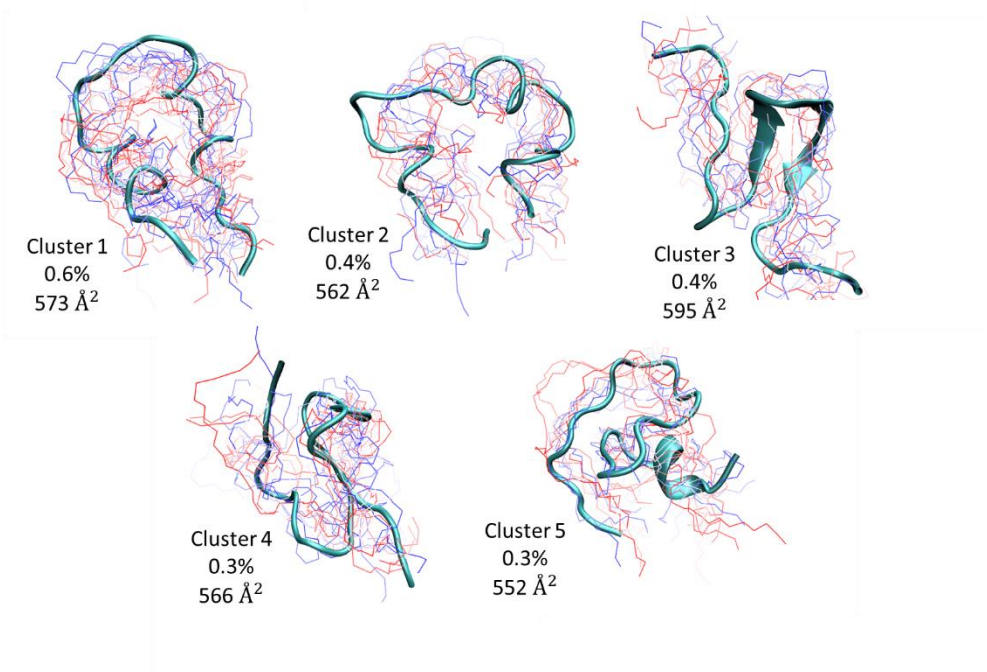


Figure III-3 The top five most populated clusters of structures from the REMD trajectory at 300 K. In teal a single structure is drawn to give an example structure, while the blue to red lines are many example structures from the cluster to give an idea of the breath of the structures in each cluster. The percentage written next to each structure represents the proportion of time the trajectory spent in that cluster throughout the course of the simulation. Finally, PSA was run on each cluster, and the average cross section is written next to each structure.

The molecular dynamics run shows a variety of structures with no single structure strongly preferred over any other structure. Figure III-3 shows some of the breath of structures identified with REMD. In order to generate clusters with amber a RMSD distance cutoff must be chosen to define the diversity of structures which can be clustered together. Smaller cutoffs result in clusters which are more closely related in structure, but fewer structures in the cluster, while larger cutoffs will allow structures which are less related to eachother, but each cluster will represent a larger portion of the total structures. In Figure III-3 this cutoff has been chosen to be arguably too large. Each structure shows one example structure in teal, and a sampling of other structures in blue and red lines. Each cluster clearly shows a large diversity of structures, which one could argue are

actually unrelated and should not belong to the same cluster. Despite setting this cutoff to a value in which each cluster includes as many structures as possible. No cluster can account for even 1% of the peptide's simulation time. In other words, even using a generous definition of what a single structure is, this peptide never favors any single structure, nor even a few structures.

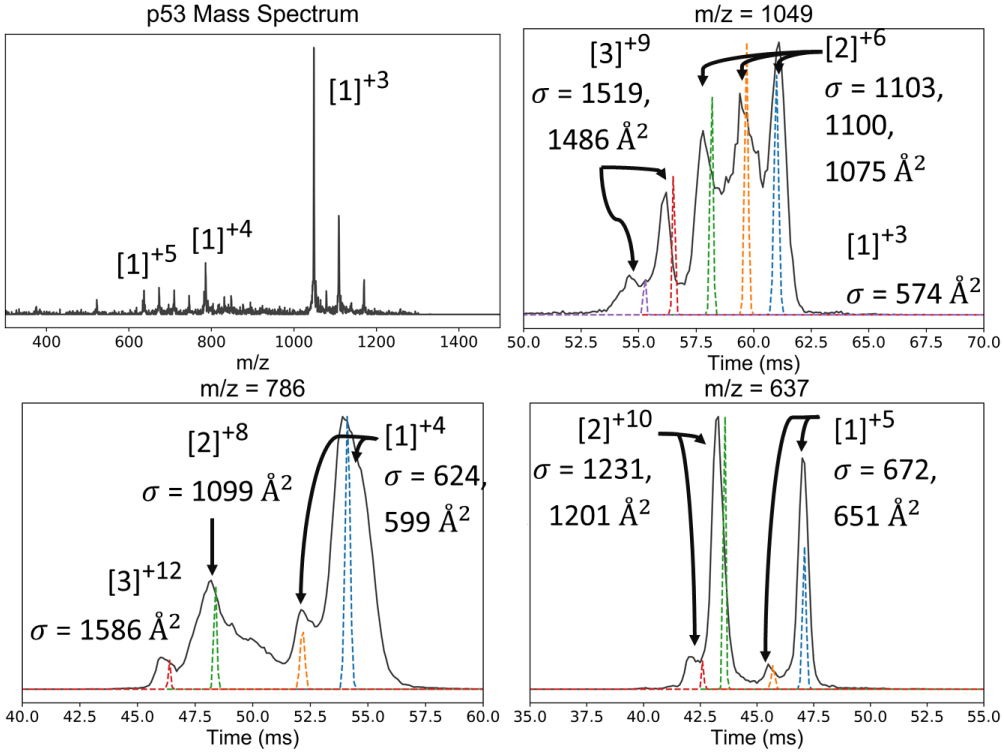


Figure III-4 Mass spectrum and arrival time distributions for each of the labeled peaks in the mass spectrum. The mass spectrum is labeled as such, and each ATD is labeled above with the mass to charge ratio that the ATD was measured at (eg. m/z = 1049). Each peak is labeled according to the convention $[n]^z$ where n is number of monomers in the cluster and z is the charge. For the mass spectrum this label has been chosen to represent the minimum possible value for n, but a single peak in the mass spectrum may represent several oligomers with the same mass to charge ratio. For ATDs experimental data is represented with a black line, while colored lines represent what a peak of a single structure with the same cross section would look like. Each peak has been labeled with the experimentally determined cross section. In some cases, a single oligomer adopts a variety of structures, in this case, the most compact structure arrives first.

This result is also reflected in the ion mobility data as well. At all three possible charge states, we can see that the smallest oligomer adopts a wide range of structures. This is most

drastically the case for the dimer six charges. The ATD with m/z = 1049 shows a triplet of peaks which are poorly resolved. The fits, the colored lines, represent what a single structure at that cross section would look like. For these fits show that this instrument could fully resolve these three peaks if they represented three distinct structures. If however these three peaks represented three structures which could easily interconvert we would expect to see the three peaks to run together as shown in this figure. For the monomer with four charges (m/z = 786) we also see that the monomer is much broader then the fit, again representing a diverse set of structures centered at that region. It is not until we reach the monomer with five charges (m/z = 637) that the data starts to approach the resolution of the fit. This is expected since intramolecular coulombic repulsion, would restrain the diversity of structures accessible to the peptide.

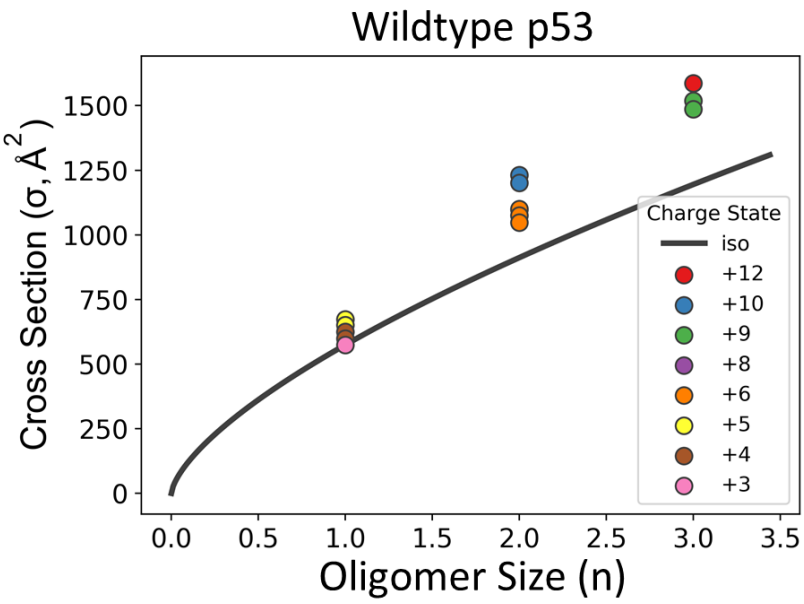


Figure III-5 Cross section dependence on Oligomer size. Each dot on the plot represents a cross section, oligomer size, and charge associated with the labeled peaks in Figure III-4. The black line represents isotropic growth based on the lowest charged monomer's cross section. Higher order oligomers above the line suggest extended growth typical of amyloid systems.

We also find that the peptides are highly prone to aggregate, and that the aggregates are fibrillar. At lower charge states (m/z = 1049) there is almost an undetectable amount of monomer,

while at higher charge states (m/z = 637) aggregates continue to be are observed. In addition, higher order oligomers appear to have cross sections larger than the isotropic prediction based on the monomer. The isotropic prediction is the cross section of an oligomer if the aggregation process was followed a globular pathway. Experimental cross sections exceeding the isotropic prediction suggest structure to the oligomers and is typical of amyloid systems. In addition, TEM data of aggregated peptide shows a fibril morphology (Figure III-7).

Interestingly, the R248Q mutation does not seem to have a strong effect on the structure. Despite the replacement of a charged residue with a neutral residue, resulting in an overall charge reduction of the peptide, the PMF plots (Figure III-6) of the simulations suggest the two peptides are sampling a similar conformational space. While it may be surprising that the removal of a charge had such a small effect of the simulation, it should be noted that there is another arginine directly adjacent to R248. Perhaps the preservation of the charge center (despite the halving of the value) preserves the structure space that the peptide samples from. Additionally, there is little preference of either peptide towards any single structure, and these two plots represent the space of a peptide randomly sampling a diverse conformational space.

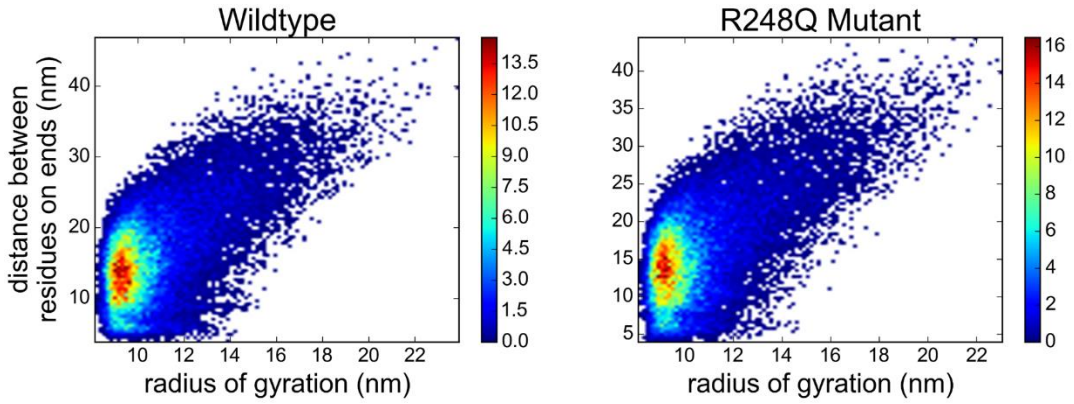


Figure III-6 PMF plots from simulations of the wildtype peptide and the mutant.

Transmission electron microscopy (Figure III-7) suggest that the peptide forms fibrils. While the morphology does appear to change between the two systems it should be noted that this could also be due to a change in the solvent conditions. The wild type was measured from a solution of pure water. The mutant however, was not soluble in pure water, and a 1:1 mixture of methanol:water needed to be used.

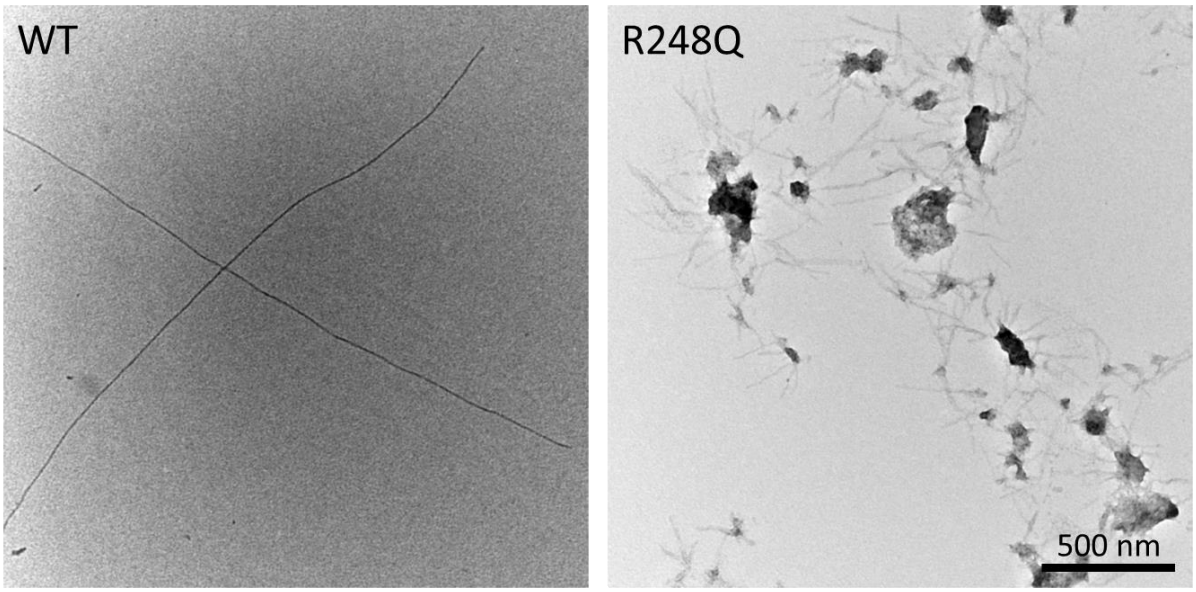


Figure III-7 Representative transmission electron microscopy images of each peptide after a week of incubation. While the morphology seems distinct, this may be due to differences in solvent. The wild type is in pure water, but the mutant was not soluble in pure water and a mixture of 1:1 methanol:water was used.

C. Methods

i. Ion Mobility Data

Ion mobility data was collected on an instrument which is described in detail elsewhere (Kemper, Dupuis, and Bowers 2009). Briefly, the instrument uses a nano electrospray ionization source. The ions then enter a source region which is held at 0.25 torr bellow the drift pressure which is around 10 torr. This pressure difference ensures that the drift tube maintains a pure helium buffer gas composition. In the source region the ions are focused through a orifice in an hour glass shaped ion funnel which separates the impurities in the source region from the pure helium in the drift region. After the orifice, the diameter of the ion funnel increases again. The drift tube side of the ion funnel serves to store ions during pulsed ion mobility experiments. The drift tube is held at 10 torr and is 2 meters long. At the end of the drift tube the ions are refocused though an orifice in the exit funnel which takes the ions straight from the 10 torr drift tube to a the main chamber region which is held at 3×10^{-5} torr. In this regions the ions are mass selected with a quadruple and detected.

To take a mass spectrum the ions on continuously introduced into the instrument and the quadruple scans a mass range to acquire a mass spectrum. To take an ATD the ions are stored then pulsed in the entrance funnel. The quadrupole is set to filter all but a single mass to charge ratio and the arrival time of the ions at the detector is recorded.

The peptides were meas

ii. Transmission Electron Microscopy

Transmission electron microscopy was carried out by a collaborator, Nikki LaPointe. The samples were adsorbed onto a 300 mesh formvar/carbon copper grids and imaged with a JEOL 123 microscope equipped with an ORCA camera.

iii. Molecular Dynamics

Temperature Replica Exchange Molecular dynamics were carried out using Amber 15 (D.A. Case et al. 2015) and the ff14SB forcefield with ff99SB backbone parameters since ff14SB is optimized for explicit solvent. The simulations were run using an implicit solvent model. Each run used 12 replicas with exchange attempts every 3 ps and with an exchange rate around 25%. Time steps for these simulations were 1 fs. Each simulation was run for about 1000 ns and convergence was checked using block analysis. Analysis was carried out on the final 600 ns of simulation time.

iv. Peptides

Peptides were custom ordered from genscript. The peptides were dissolved and aliquoted using HFIP in order to reduce any oligomers already present back to the monomer state. HFIP was allowed to evaporate, and the samples were held dry in the freezer until ready for use. They samples were dissolved to a conenctraioon of 100 μM and Ion mobility data was taken that same day. The samples were allowed to incubate for one week before TEM images were taken.

D. Conclusions and Future Work

While the data remains inconclusive without mutant and interaction data, preliminary results show that with out the mutation the peptide readily aggregates, and that while the morphology as determined by TEM shows differences between the wild type and mutant, they still both form fibrils. In addition, simulation suggest that the monomer is surprisingly unaffected by the mutation. Together this seems to suggest that the toxicity of this mutation is not necessarily that it

is much more prone to form amyloids, but that the protein is destabilized without the crucial arginine that complexes with the DNA. The result is that the already amyloid prone DNA binding domain may interact with healthy proteins, causing the suggested prion effect.

To verify this ion mobility experiment must be carried out on the mutant and the mixture of the mutant and wild type to confirm that the mechanism is unaffected by the mutant. It should be noted however that with out the extra charge the mutant is much harder to work with then the wildtype, but preliminary results show that it should be possible even in just water. In addition, molecular dynamics showing that the mutation destabilizing the protein would be a nice addition. Admittedly this would be a difficult simulation and would require some course graining and a specialized molecular dynamics group in order to achieve results.

E. References

Ano Bom, A. P. D., L. P. Rangel, D. C. F. Costa, G. A. P. de Oliveira, D. Sanches, C. A. Braga, L. M. Gava, et al. 2012. "Mutant P53 Aggregates into Prion-like Amyloid Oligomers and Fibrils: IMPLICATIONS FOR CANCER." *Journal of Biological Chemistry* 287 (33): 28152–62. https://doi.org/10.1074/jbc.M112.340638.

Cho, Y, S Gorina, P. Jeffrey, and N. Pavletich. 1994. "Crystal Structure of a P53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations." *Science* 265 (5170): 346–55. https://doi.org/10.1126/science.8023157.

D.A. Case, J.T. Berryman, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, and et al. 2015. *AMBER 15*. University of California, San Francisco.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502 (7471): 333–39. https://doi.org/10.1038/nature12634.

Kemper, Paul R., Nicholas F. Dupuis, and Michael T. Bowers. 2009. "A New, Higher Resolution, Ion Mobility Mass Spectrometer." *International Journal of Mass Spectrometry* 287 (1–3): 46–57. https://doi.org/10.1016/j.ijms.2009.01.012.

Levy, Claudia B., Ana C. Stumbo, Ana P.D. Ano Bom, Elisabeth A. Portari, Yraima Carneiro, Jerson L. Silva, and Claudia V. De Moura-Gallo. 2011. "Co-Localization of Mutant P53 and Amyloid-like Protein Aggregates in Breast Tumors." *The International Journal of Biochemistry & Cell Biology* 43 (1): 60–64. https://doi.org/10.1016/j.biocel.2010.10.017.

Olivier, M. 2006. "The Clinical Value of Somatic TP53 Gene Mutations in 1,794 Patients with Breast Cancer." *Clinical Cancer Research* 12 (4): 1157–67. https://doi.org/10.1158/1078-0432.CCR-05-1029.

Silva, Jerson L., Claudia V. De Moura Gallo, Danielly C.F. Costa, and Luciana P. Rangel. 2014. “Prion-like Aggregation of Mutant P53 in Cancer.” *Trends in Biochemical Sciences* 39 (6): 260–67. https://doi.org/10.1016/j.tibs.2014.04.001.

Silva, Jerson L., Luciana P. Rangel, Danielly C. F. Costa, Yraima Cordeiro, and Claudia V. De Moura Gallo. 2013. “Expanding the Prion Concept to Cancer Biology: Dominant-Negative Effect of Aggregates of Mutant P53 Tumour Suppressor.” *Bioscience Reports* 33 (4): 593–603. https://doi.org/10.1042/BSR20130065.

Tidow, H., R. Melero, E. Mylonas, S. M. V. Freund, J. G. Grossmann, J. M. Carazo, D. I. Svergun, M. Valle, and A. R. Fersht. 2007. “Quaternary Structures of Tumor Suppressor P53 and a Specific P53 DNA Complex.” *Proceedings of the National Academy of Sciences* 104 (30): 12324–29. https://doi.org/10.1073/pnas.0705069104.

Vogelstein, B., D. Lane, and A. J. Levine. 2000. “Surfing the P53 Network.” *Nature* 408 (6810): 307–10. https://doi.org/10.1038/35042675.

IV. The Classifying Autoencoder: Gaining Insight to Amyloid Assembly of Peptides and Proteins

Reprinted with permission from Tro, Michael J., Nathaniel Charest, Zachary Taitz, Joan-Emma Shea, and Michael T. Bowers. 2019. "The Classifying Autoencoder: Gaining Insight into Amyloid Assembly of Peptides and Proteins." The Journal of Physical Chemistry B 123 (25): 5256–64. https://doi.org/10.1021/acs.jpcb.9b03415.

A. 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.

B. 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(Chiti and Dobson 2006; Fowler et al. 2007; Zhao and Townsend 2009; Bernstein et al. 2009; Bleiholder et al. 2011). 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 (Astbury, Dickinson, and Bailey 1935;

Morriss-Andrews and Shea 2015). 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 (Economou et al. 2016; Makin and Serpell 2002; Jiménez et al. 2002; Fitzpatrick et al. 2013; Sipe and Cohen 2000). While amyloids are particularly known for association with degenerative diseases such as Alzheimer's (Jarrett and Lansbury 1993; Stelzmann, Norman Schnitzlein, and Reed Murtagh 1995; Bernstein et al. 2009), Parkinson's (Maries et al. 2003), Huntington's (Scherzinger et al. 1997), type 2 diabetes (Westermark, Andersson, and Westermark 2011), and amyotrophic lateral sclerosis (Elam et al. 2003), they may also play beneficial roles. The motif appears in spider silks, egg shells, biofilms and biomechanical scaffolds for human synthetic pathways (Fowler et al. 2007).

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 (Walsh et al. 2014; Tartaglia and Vendruscolo 2008; Thompson et al. 2006). Enough algorithms have been developed on this topic that there exists prediction algorithms which take into account as many other algorithms as possible (Emily, Talvas, and Delamarche 2013; Tsolis et al. 2013). 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. This is also common in attempts to use machine learning for understanding amyloid aggregation. 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 (Olden and Jackson 2002; White 1988; Gómez-Bombarelli et al. 2018; Brunner et al. 2018). Attempts have been made to elucidate the inner workings of ANNs (Hermundstad et al. 2011; Andrews, Diederich, and Tickle 1995), 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 (Hinton and Salakhutdinov 2006). Differing versions of the basic autoencoder, perhaps most notably the Variational Autoencoder (VAE) (Kingma and Welling 2013) have emerged, with variants typically involving goals beyond dimensional reduction and reconstruction of the data (Wehmeyer and Noé 2018). In this paper, our goal was to develop a method of classification, which we call the classifying autoencoder (CAE), based on prior algorithms(Gómez-Bombarelli et al. 2018; Brunner et al. 2018; Kingma and Welling 2013), 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(Beerten et al. 2015). There are relatively few known examples of smaller peptides which form amyloids(Reches, Porat, and Gazit 2002; Reches and Gazit 2004). 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(Beerten et al. 2015). 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(Do, de Almeida, et al. 2016) 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 (Beerten et al. 2015).

C. Methods

i. 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. Figure III-8 depicts a flow chart of the process used to generate this database.

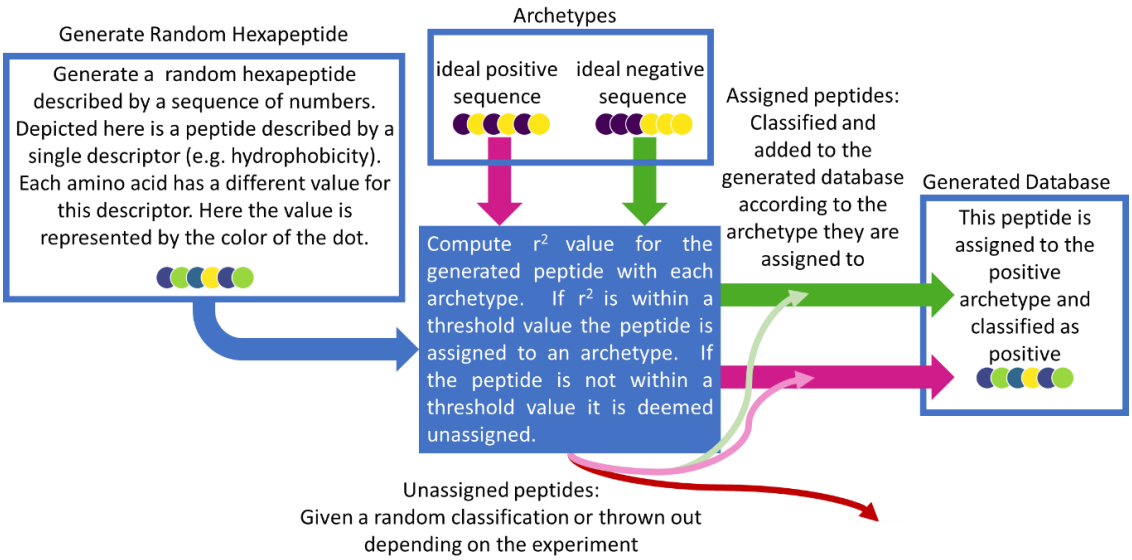


Figure III-8 This flow chart depicts the generation of the database. In the example depiction shown in the flow chart a random peptide is added to the database by being assigned to the positive archetype and classified as positive.

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.

ii. 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 (Beerten et al. 2015; Schymkowitz and Rousseau n.d.) 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 (Stanislawski, Kotulska, and Unold 2013; Kim et al. 2009). 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.

iii. Polarity Descriptor

We found the hydrophobic parameter using the AAindex database (S. Kawashima and Kanehisa 2000; Shuichi Kawashima et al. 2008; Tomii and Kanehisa 1996). 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 (Fauchere, Jean Luc; Pliska, Vladimir 1983). 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.

iv. 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 (Kemper, Dupuis, and Bowers 2009). 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(Gidden et al. 2004). 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 (Mason and McDaniel 1988).

Software

All neural nets were constructed and trained using the Keras software package(Chollet and Others 2015) with the Tensorflow backend(Martín Abadi et al., n.d.).

D. Results and Discussion

i. 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 *Figure III-9*. Inputs (a description of the peptide) are fed into the model via the input nodes. The depiction in *Figure III-9* 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.

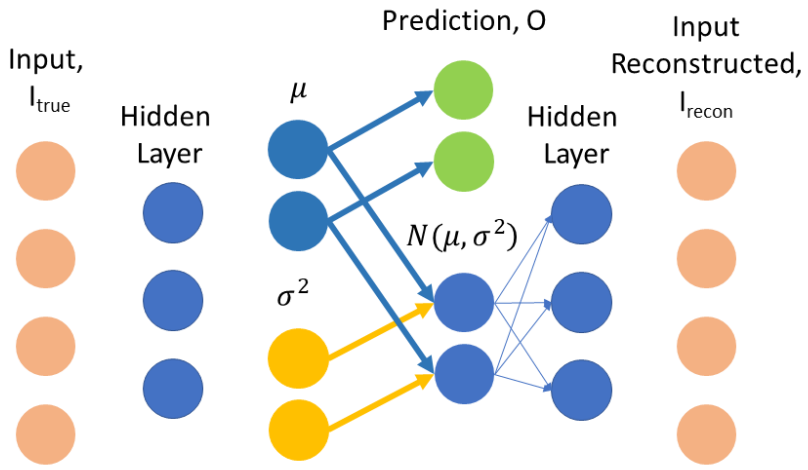


Figure III-9 Depicted is the architecture of a classifying autoencoder (CAE) with four inputs, and one hidden layer with three nodes. The latent space in this model is two-dimensional and is labeled μ. These nodes are also used as the prediction layer. Noise is added to the latent space at the nodes 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.

ii. 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 *Figure III-10* are the values in the latent space; that is the values output by the two nodes labeled μ in *Figure III-9*. 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. *Figure III-10* A, D, and E plot each peptide in the database according to where they fall in latent space. *Figure III-10* 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.

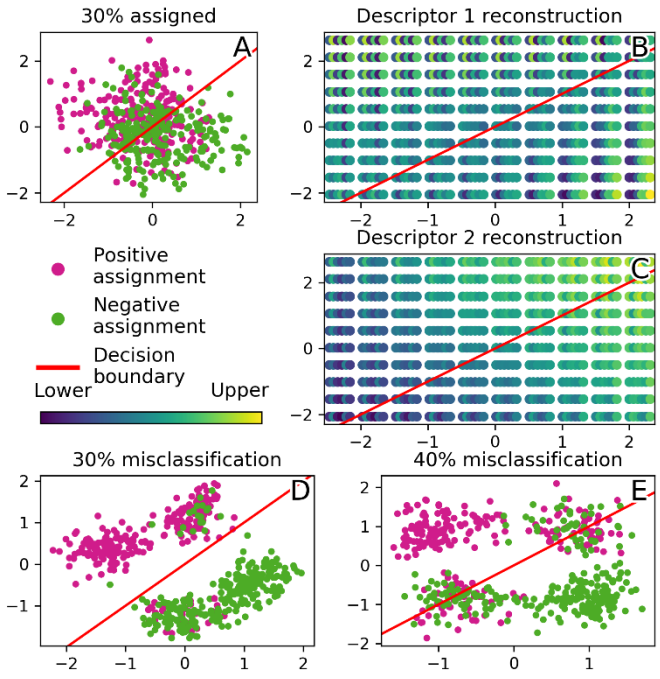


Figure III-10 For all plots, the y-axis is the positive prediction axis, and the x-axis is the negative prediction axis. A peptide is predicted positive if it is falls above the red line and is predicted negative if it falls below the red line. Plot (A) shows where each peptide is encoded by the CAE, 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.

Figure III-10 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 *Figure III-10* 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 *Figure III-10* 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.

Figure III-10 D and E depict different models than *Figure III-10* A-C. These use a database generated to mimic a set of experiments that yielded occasionally flawed results. In *Figure III-10* 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 *Figure III-10* 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, *Figure III-10* 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 *Figure III-10* B and C. The middle of *Figure III-10* A, the data's latent space distribution, shows mixed positive and negative peptides; in *Figure III-10* 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, *Figure III-10* A, we see a separation of positive classifications from the mixture of classifications; when we follow this trajectory in the reconstruction plot, *Figure III-10* 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 *Figure III-10* 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 (Figure A-I-5) correctly reflects the four archetypes. This gives us insight to how the model deals with the uncertainty in the data. In *Figure III-10* 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 *Figure III-10* 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; Figure A-1-7), 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 (Beerten et al. 2015).

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.

iii. 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 *Figure III-11* B (and later in *Figure III-12* 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.

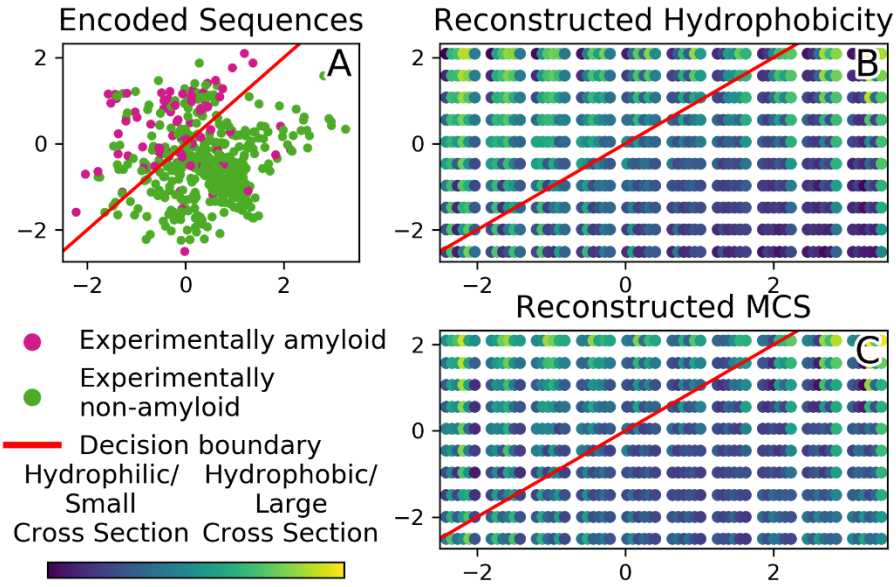


Figure III-11. Representative model trained using hydrophobicity and monomer cross section (MCS). (A) All sequences in the validation set plotted in latent space. Each axis here is the value of one of the latent space nodes. The color of the point represents the experimental classification of that peptide. (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 (West et al. 1999) 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.

iv. CAE on an Experimental Database: Monomeric Cross Section

Figure III-11 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(Valentine, Counterman, and Clemmer 1999; Dilger, Glover, and Clemmer 2017; Counterman and Clemmer 1999).

Monomer Cross Section and Dimeric Isotropic Deviation

Amino acid	Monomer Cross Section (Å±Standard Deviation)	Dimeric isotropic deviation, $\Delta i_2 \times 100$ (± Standard Deviation)
Glutamic acid	61.9 ± 0.3	-6.1 ± 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 (Δi_2) for each amino acid. The Δi_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.

v. 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 (Do, de Almeida, et al. 2016). 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 (Do, de Almeida, et al. 2016). As we collected more data on aggregation of amino acids, we found that this value was oligomer size dependent (Figure A-I-3). 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 Figure A-1-4). 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.

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

Here we probe the relationship between DID and amyloid propensity. For the most part, *Figure III-12* 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 (*Figure III-10* 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 *Figure III-10* 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 *Figure III-12* C) are most strongly characterized by a higher degree of compactness, with less distinguishing features in hydrophobicity representation.

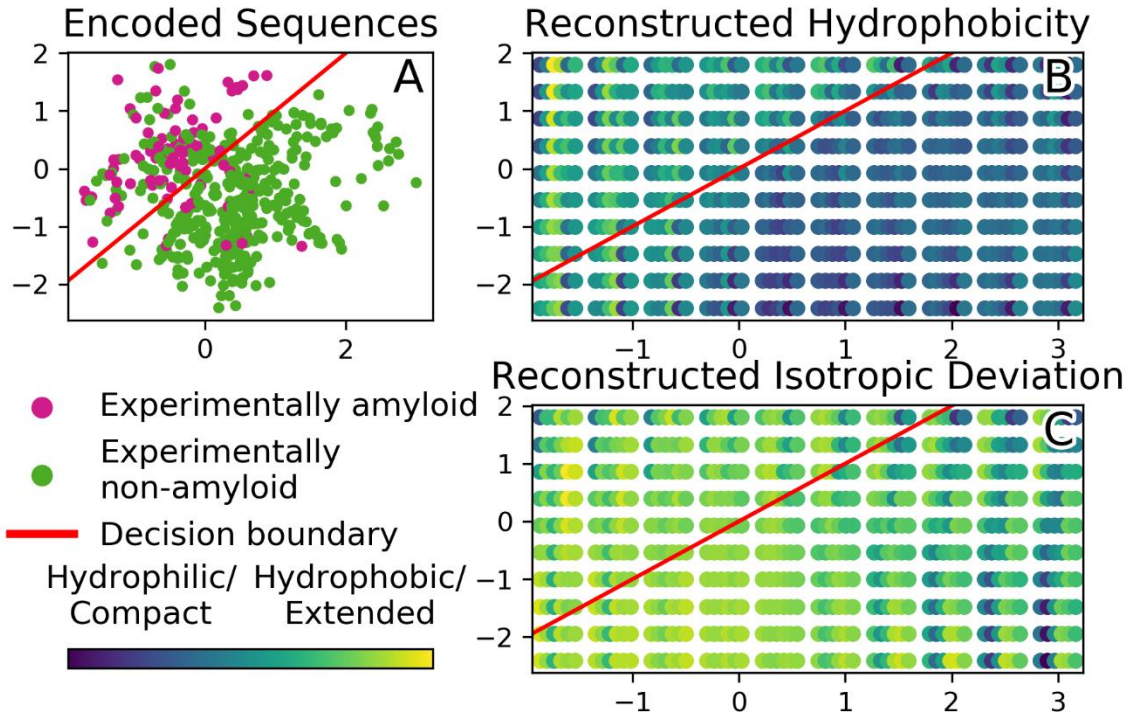


Figure III-12 Representative model trained using hydrophobicity and DID. This figure is the same representation of a model as Figure III-11 except (C) depicts the DID of each point in latent space. Here yellow is extended growth, and blue is compact growth.

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 (Fink 2005; Uversky 2010). Further, it has also recently been shown that glycine is an essential residue in cylindrin formation(Laganowsky et al. 2012; Do, LaPointe, et al. 2016; Sangwan et al. 2017), a 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.

E. 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.

F. Acknowledgments

I greatly appreciate my coautors on this paper Nathaniel Charest, Zachary Taitz, Joan-Emma Shea, and My advisor Michael T Bowers.

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).

G. References

Andrews, Robert, Joachim Diederich, and Alan B. Tickle. 1995. "Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks." *Knowledge-Based Systems*, Knowledge-based neural networks, 8 (6): 373–89. https://doi.org/10.1016/0950-7051(96)81920-4.

Astbury, William Thomas, Sylvia Dickinson, and Kenneth Bailey. 1935. "The X-Ray Interpretation of Denaturation and the Structure of the Seed Globulins." *Biochemical Journal* 29 (10): 2351-2360.1.

Beerten, Jacinte, Joost Van Durme, Rodrigo Gallardo, Emidio Capriotti, Louise Serpell, Frederic Rousseau, and Joost Schymkowitz. 2015. "WALTZ-DB: A Benchmark Database of Amyloidogenic Hexapeptides." *Bioinformatics* 31 (10): 1698–1700. https://doi.org/10.1093/bioinformatics/btv027.

Bernstein, Summer L., Nicholas F. Dupuis, Noel D. Lazo, Thomas Wyttenbach, Margaret M. Condron, Gal Bitan, David B. Teplow, et al. 2009. "Amyloid-β Protein Oligomerization and the Importance of Tetramers and Dodecamers in the Aetiology of Alzheimer's Disease." *Nature Chemistry* 1 (4): 326–31. https://doi.org/10.1038/nchem.247.

Bleiholder, Christian, Nicholas F. Dupuis, Thomas Wyttenbach, and Michael T. Bowers. 2011. "Ion Mobility–Mass Spectrometry Reveals a Conformational Conversion from Random Assembly

to β-Sheet in Amyloid Fibril Formation." *Nature Chemistry* 3 (2): 172–77. https://doi.org/10.1038/nchem.945.

Brunner, Gino, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. 2018. "MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer." *ArXiv:1809.07600*, September. http://arxiv.org/abs/1809.07600.

Chiti, Fabrizio, and Christopher M. Dobson. 2006. "Protein Misfolding, Functional Amyloid, and Human Disease." *Annual Review of Biochemistry* 75 (1): 333–66. https://doi.org/10.1146/annurev.biochem.75.101304.123901.

Chollet, François, and Others. 2015. *Keras*. https://keras.io.

Counterman, Anne E., and David E. Clemmer. 1999. "Volumes of Individual Amino Acid Residues in Gas-Phase Peptide Ions." *Journal of the American Chemical Society* 121 (16): 4031–39. https://doi.org/10.1021/ja984344p.

Dilger, Jonathan M., Matthew S. Glover, and David E. Clemmer. 2017. "A Database of Transition-Metal-Coordinated Peptide Cross-Sections: Selective Interaction with Specific Amino Acid Residues." *Journal of The American Society for Mass Spectrometry* 28 (7): 1293–1303. https://doi.org/10.1007/s13361-016-1592-9.

Do, Thanh D., Natália E. C. de Almeida, Nichole E. LaPointe, Ali Chamas, Stuart C. Feinstein, and Michael T. Bowers. 2016. "Amino Acid Metaclusters: Implications of Growth Trends on Peptide Self-Assembly and Structure." *Analytical Chemistry* 88 (1): 868–76. https://doi.org/10.1021/acs.analchem.5b03454.

Do, Thanh D., Nichole E. LaPointe, Rebecca Nelson, Pascal Krotee, Eric Y. Hayden, Brittany Ulrich, Sarah Quan, et al. 2016. "Amyloid β-Protein C-Terminal Fragments: Formation of Cylindrins and β-Barrels." *Journal of the American Chemical Society* 138 (2): 549–57. https://doi.org/10.1021/jacs.5b09536.

Economou, Nicholas J., Maxwell J. Giammona, Thanh D. Do, Xueyun Zheng, David B. Teplow, Steven K. Buratto, and Michael T. Bowers. 2016. "Amyloid β-Protein Assembly and Alzheimer's Disease: Dodecamers of Aβ42, but Not of Aβ40, Seed Fibril Formation." *Journal of the American Chemical Society* 138 (6): 1772–75. https://doi.org/10.1021/jacs.5b11913.

Elam, Jennifer Stine, Alexander B. Taylor, Richard Strange, Svetlana Antonyuk, Peter A. Doucette, Jorge A. Rodriguez, S. Samar Hasnain, et al. 2003. "Amyloid-like Filaments and Water-Filled Nanotubes Formed by SOD1 Mutant Proteins Linked to Familial ALS." *Nature Structural & Molecular Biology* 10 (6): 461–67. https://doi.org/10.1038/nsb935.

Emily, Mathieu, Anthony Talvas, and Christian Delamarche. 2013. "MetAmyl: A METa-Predictor for AMYLoid Proteins." *PLOS ONE* 8 (11): e79722. https://doi.org/10.1371/journal.pone.0079722.

Fauchere, Jean Luc; Pliska, Vladimir. 1983. "Hydrophobic Parameters π of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino Acid Amides." *European Journal of Medicinal Chemistry* 18 (3): 369–75.

Fink, Anthony L. 2005. “Natively Unfolded Proteins.” *Current Opinion in Structural Biology*, Folding and binding / Protein-nucleic acid interactions, 15 (1): 35–41. https://doi.org/10.1016/j.sbi.2005.01.002.

Fitzpatrick, Anthony W. P., Galia T. Debelouchina, Marvin J. Bayro, Daniel K. Clare, Marc A. Caporini, Vikram S. Bajaj, Christopher P. Jaroniec, et al. 2013. “Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril.” *Proceedings of the National Academy of Sciences* 110 (14): 5468–73. https://doi.org/10.1073/pnas.1219476110.

Fowler, Douglas M., Atanas V. Koulov, William E. Balch, and Jeffery W. Kelly. 2007. “Functional Amyloid – from Bacteria to Humans.” *Trends in Biochemical Sciences* 32 (5): 217–24. https://doi.org/10.1016/j.tibs.2007.03.003.

Gidden, Jennifer, Alessandra Ferzoco, Erin Shammel Baker, and Michael T. Bowers. 2004. “Duplex Formation and the Onset of Helicity in Poly d(CG)n Oligonucleotides in a Solvent-Free Environment.” *Journal of the American Chemical Society* 126 (46): 15132–40. https://doi.org/10.1021/ja046433+.

Gómez-Bombarelli, Rafael, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. 2018. “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.” *ACS Central Science* 4 (2): 268–76. https://doi.org/10.1021/acscentsci.7b00572.

Hermundstad, Ann M., Kevin S. Brown, Danielle S. Bassett, and Jean M. Carlson. 2011. “Learning, Memory, and the Role of Neural Network Architecture.” *PLOS Computational Biology* 7 (6): e1002063. https://doi.org/10.1371/journal.pcbi.1002063.

Hinton, G. E., and R. R. Salakhutdinov. 2006. “Reducing the Dimensionality of Data with Neural Networks.” *Science* 313 (5786): 504–7. https://doi.org/10.1126/science.1127647.

Jarrett, Joseph T., and Peter T. Lansbury. 1993. “Seeding ‘One-Dimensional Crystallization’ of Amyloid: A Pathogenic Mechanism in Alzheimer’s Disease and Scrapie?” *Cell* 73 (6): 1055–58. https://doi.org/10.1016/0092-8674(93)90635-4.

Jiménez, José L., Ewan J. Nettleton, Mario Bouchard, Carol V. Robinson, Christopher M. Dobson, and Helen R. Saibil. 2002. “The Protofilament Structure of Insulin Amyloid Fibrils.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (14): 9196–9201. https://doi.org/10.1073/pnas.142459399.

Kawashima, S., and M. Kanehisa. 2000. “AAindex: Amino Acid Index Database.” *Nucleic Acids Research* 28 (1): 374.

Kawashima, Shuichi, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. 2008. “AAindex: Amino Acid Index Database, Progress Report 2008.” *Nucleic Acids Research* 36 (suppl_1): D202–5. https://doi.org/10.1093/nar/gkm998.

Kemper, Paul R., Nicholas F. Dupuis, and Michael T. Bowers. 2009. “A New, Higher Resolution, Ion Mobility Mass Spectrometer.” *International Journal of Mass Spectrometry* 287 (1–3): 46–57. https://doi.org/10.1016/j.ijms.2009.01.012.

Kim, Changsik, Jiwon Choi, Seong Joon Lee, William J. Welsh, and Sukjoon Yoon. 2009. “NetCSSP: Web Application for Predicting Chameleon Sequences and Amyloid Fibril Formation.” *Nucleic Acids Research* 37 (Web Server issue): W469–73. https://doi.org/10.1093/nar/gkp351.

Kingma, Diederik P., and Max Welling. 2013. “Auto-Encoding Variational Bayes.” In *ArXiv:1312.6114*. http://arxiv.org/abs/1312.6114.

Laganowsky, Arthur, Cong Liu, Michael R. Sawaya, Julian P. Whitelegge, Jiyong Park, Minglei Zhao, Anna Pensalfini, et al. 2012. “Atomic View of a Toxic Amyloid Small Oligomer.” *Science* 335 (6073): 1228–31. https://doi.org/10.1126/science.1213151.

Makin, O. S., and L. C. Serpell. 2002. “Examining the Structure of the Mature Amyloid Fibril.” *Biochemical Society Transactions* 30 (4): 521–25. https://doi.org/10.1042/.

Maries, Eleonora, Biplob Dass, Timothy J. Collier, Jeffrey H. Kordower, and Kathy Steece-Collier. 2003. “The Role of α-Synuclein in Parkinson’s Disease: Insights from Animal Models.” *Nature Reviews Neuroscience* 4 (9): 727–38. https://doi.org/10.1038/nrn1199.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. n.d. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. tensorflow.org.

Mason, Edward A., and Earl W. McDaniel. 1988. *Transport Properties of Ions in Gases*. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/3527602852.

Morriss-Andrews, Alex, and Joan-Emma Shea. 2015. “Computational Studies of Protein Aggregation: Methods and Applications.” *Annual Review of Physical Chemistry* 66 (1): 643–66. https://doi.org/10.1146/annurev-physchem-040513-103738.

Olden, Julian D, and Donald A Jackson. 2002. “Illuminating the ‘Black Box’: A Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks.” *Ecological Modelling* 154 (1–2): 135–50. https://doi.org/10.1016/S0304-3800(02)00064-9.

Reches, Meital, and Ehud Gazit. 2004. “Amyloidogenic Hexapeptide Fragment of Medin: Homology to Functional Islet Amyloid Polypeptide Fragments.” *Amyloid* 11 (2): 81–89. https://doi.org/10.1080/13506120412331272287.

Reches, Meital, Yair Porat, and Ehud Gazit. 2002. “Amyloid Fibril Formation by Pentapeptide and Tetrapeptide Fragments of Human Calcitonin.” *Journal of Biological Chemistry* 277 (38): 35475–80. https://doi.org/10.1074/jbc.M206039200.

Sangwan, Smriti, Anni Zhao, Katrina L. Adams, Christina K. Jayson, Michael R. Sawaya, Elizabeth L. Guenther, Albert C. Pan, et al. 2017. “Atomic Structure of a Toxic, Oligomeric Segment of SOD1 Linked to Amyotrophic Lateral Sclerosis (ALS).” *Proceedings of the National Academy of Sciences* 114 (33): 8770–75. https://doi.org/10.1073/pnas.1705091114.

Scherzinger, Eberhard, Rudi Lurz, Mark Turmaine, Laura Mangiarini, Birgit Hollenbach, Renate Hasenbank, Gillian P Bates, Stephen W Davies, Hans Lehrach, and Erich E Wanker. 1997. "Huntingtin-Encoded Polyglutamine Expansions Form Amyloid-like Protein Aggregates In Vitro and In Vivo." *Cell* 90 (3): 549–58. https://doi.org/10.1016/S0092-8674(00)80514-0.

Schymkowitz, Joost, and Frederic Rousseau. n.d. "Peptide Sequences | WALTZ-DB." Accessed January 24, 2019. http://waltzdb.switchlab.org/.

Sipe, Jean D., and Alan S. Cohen. 2000. "Review: History of the Amyloid Fibril." *Journal of Structural Biology* 130 (2): 88–98. https://doi.org/10.1006/jsbi.2000.4221.

Stanislawski, Jerzy, Malgorzata Kotulska, and Olgierd Unold. 2013. "Machine Learning Methods Can Replace 3D Profile Method in Classification of Amyloidogenic Hexapeptides." *BMC Bioinformatics* 14 (January): 21. https://doi.org/10.1186/1471-2105-14-21.

Stelzmann, Rainulf A., H. Norman Schnitzlein, and F. Reed Murtagh. 1995. "An English Translation of Alzheimer's 1907 Paper, 'Über Eine Eigenartige Erkankung Der Hirnrinde.'" *Clinical Anatomy* 8 (6): 429–31. https://doi.org/10.1002/ca.980080612.

Tartaglia, Gian Gaetano, and Michele Vendruscolo. 2008. "The Zyggregator Method for Predicting Protein Aggregation Propensities" 37 (7): 1395–1401. https://doi.org/10.1039/B706784B.

Thompson, Michael J., Stuart A. Sievers, John Karanicolas, Magdalena I. Ivanova, David Baker, and David Eisenberg. 2006. "The 3D Profile Method for Identifying Fibril-Forming Segments of Proteins." *Proceedings of the National Academy of Sciences of the United States of America* 103 (11): 4074–78. https://doi.org/10.1073/pnas.0511295103.

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

Tsolis, Antonios C., Nikos C. Papandreou, Vassiliki A. Iconomidou, and Stavros J. Hamodrakas. 2013. "A Consensus Method for the Prediction of 'Aggregation-Prone' Peptides in Globular Proteins." *PLOS ONE* 8 (1): e54175. https://doi.org/10.1371/journal.pone.0054175.

Uversky, Vladimir N. 2010. "Targeting Intrinsically Disordered Proteins in Neurodegenerative and Protein Dysfunction Diseases: Another Illustration of the D2 Concept." *Expert Review of Proteomics* 7 (4): 543–64. https://doi.org/10.1586/epr.10.36.

Valentine, Stephen J., Anne E. Counterman, and David E. Clemmer. 1999. "A Database of 660 Peptide Ion Cross Sections: Use of Intrinsic Size Parameters for Bona Fide Predictions of Cross Sections." *Journal of the American Society for Mass Spectrometry* 10 (11): 1188–1211. https://doi.org/10.1016/S1044-0305(99)00079-3.

Walsh, Ian, Flavio Seno, Silvio C.E. Tosatto, and Antonio Trovato. 2014. "PASTA 2.0: An Improved Server for Protein Aggregation Prediction." *Nucleic Acids Research* 42 (Web Server issue): W301–7. https://doi.org/10.1093/nar/gku399.

Wehmeyer, Christoph, and Frank Noé. 2018. “Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics.” *The Journal of Chemical Physics* 148 (24): 241703. https://doi.org/10.1063/1.5011399.

West, Michael W., Weixun Wang, Jennifer Patterson, Joseph D. Mancias, James R. Beasley, and Michael H. Hecht. 1999. “De Novo Amyloid Proteins from Designed Combinatorial Libraries.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (20): 11211–16.

Westermark, Per, Arne Andersson, and Gunilla T. Westermark. 2011. “Islet Amyloid Polypeptide, Islet Amyloid, and Diabetes Mellitus.” *Physiological Reviews* 91 (3): 795–826. https://doi.org/10.1152/physrev.00042.2009.

White, H. 1988. “Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns.” In *IEEE 1988 International Conference on Neural Networks*, 451–58 vol.2. https://doi.org/10.1109/ICNN.1988.23959.

Zhao, Wei-Qin, and Matthew Townsend. 2009. “Insulin Resistance and Amyloidogenesis as Common Molecular Foundation for Type 2 Diabetes and Alzheimer’s Disease.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, Diabetes and the Nervous System, 1792 (5): 482–96. https://doi.org/10.1016/j.bbadis.2008.10.014.

V. Unsupervised Learning of Amyloid Aggregation Molecular Dynamics Data: Automatic Order Parameter and Nucleus Detection

Nathaniel Charest[‡], Michael Tro[‡], Michael T Bowers, Joan-Emma Shea[*]

A. Abstract

The characterization of amyloid forming peptides and processes is of interest in the biochemistry community. Two processes – the direct assembly of amyloid fibrils and the liquid-to-solid phase transition of protein droplets within the cellular milieu – are a popular target for simulation due to their association with amyloid disease pathologies. Methods for characterizing these dynamical processes make use of order parameters as single value representations of the system's state, with effective order parameters offering insight into the process in an accessible way. An unsupervised machine learning algorithm known as the variational autoencoder is applied to the internal coordinates of a coarse-grained molecular dynamics simulation of an amyloid prone peptide model that spontaneously orders from a disordered solution, representing the first application of such a technique to systems of aggregating peptides. The method is shown to be able to reduce the ensemble of data to a single variable that tracks evolution in the system and successfully characterizes large-scale system evolutions with precision exceeding or comparable to more conventional order parameters. We apply the technique to an amyloid system of scale and analyze its viability as a method for large molecular dynamic simulations.

B. Introduction

Amyloid fibrils are associated with the clinical pathology of numerous diseases, including Alzheimer's disease (Bernstein et al. 2009; Jarrett and Lansbury 1993; Stelzmann, Norman Schnitzlein, and Reed Murtagh 1995), ALS (Maries et al. 2003), Type II Diabetes (Westermark, Andersson, and Westermark 2011) and Parkinson's disease (Maries et al. 2003). Because of this connection, there has been considerable investigation into the means by which these fibrils form through both experiment

and simulation. Amyloid fibrils are distinguished by their characteristic x-ray diffraction pattern, and their structure is well-characterized as being beta-sheets whose stacking axis is perpendicular to the peptide backbone (Fitzpatrick et al. 2013; Geddes et al. 1968; Guenther et al. 2018).

Fibrillizing systems have proven difficult to explore computationally due to the large number of atoms needed to produce a fully formed fibril and the timescale over which that fibral forms. Fully atomistic studies of the formation of a fibrillar plaque are limited by time and scale constraints, and so coarse grained models have been successfully used to model fibril formation from unassembled monomers(Nguyen et al. 2019; Ilie and Caflisch 2019; Chiricotto et al. 2019; M. Chen, Schafer, and Wolynes 2018; Rojas, Maisuradze, and Scheraga 2018). Previous publications have utilized the Shea peptide model to explore possible pathways by which fibrils form (Morriss-Andrews, Bellesia, and Shea 2012; Bellesia and Shea 2009). This model was chosen because its computational tractability allows the access of system sizes that enable mechanistic insight, while preserving major common attributes of various fibrillizing peptides.

The complexity of the system lends itself to the use of relatively new methods for developing coordinates that track system transformations. In this work, we apply a common type of artificial neural network – the variational autoencoder (VAE) (Kingma and Welling 2013) – with the goal of developing a single value parameter that can indicate the degree of progression of a system approaching a final, ordered form by characterizing the underlying order of the trajectory's time series. Building on work done improving molecular dynamics sampling, facilitating analysis, developing collective variables, and analyzing the Ising model (Wehmeyer and Noé 2018; W. Chen, Tan, and Ferguson 2018; Wetzel 2017; Wang, Lamim Ribeiro, and Tiwary 2020), we the first to use VAE-based techniques to analyze large systems of fibrillizing peptides. Emphasis is placed on the use of the VAE's ability to reconstruct information in human-accessible form and provide a unique means of analysis and insight.

In this paper, we first explain the VAE and demonstrate how it can use the information in a molecular dynamics ensemble of a serial ordering process to develop its single-valued order parameter. We then use the VAE to generate a this single-valued order parameter two fibrilizing systems.. The automatically learned order parameter is compared to the commonly used nematic order parameter (Stephen and Straley 1974; Saupe 1968; Eppenga and Frenkel 1984; Ray et al. 2019) both to highlight the strengths of this method and also analyze and address potential issues. We emphasize the use of the VAE's ability to 'translate' its insights into the original representation of the system, for possible of analysis. This process represents a more general field of interest: the transition of a disordered phase of matter into an ordered phase. Our contribution is the application of this method to an amyloid assembly process, representing the use of VAE methods on systems larger and more dynamically complex than systems so-far studied.

C. Methods

Out model's monomeric unit is depicted in Figure V-1. It is a phenomenological coarse grained model based on an earlier paper (Bellesia and Shea, 2007). The backbone contains two interaction centers per residue (X and Y) along the backbone, and one interaction center on the side chain. Four different types of side chain groups are considered: hydrophobic group H, polar group P, cationic group C and anionic group A. Capping groups, E, are used at the termini. The sequence is chosen to have an alternating sequence HPHPHP, a decision initially motivated by combinatorial studies of Hecht and collaborators (Xiong et al. 1995; West et al. 1999)

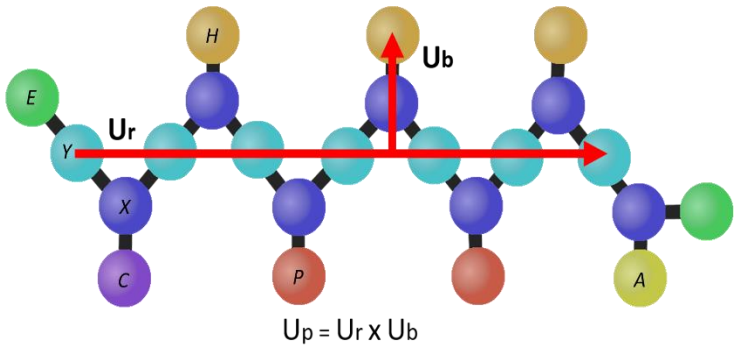


Figure V-1: A diagram of the peptide model with important vectors labeled.

Hecht et al. that indicated a generic amphiphilic alternating pattern is a major indicator of beta-sheet for self-assembling peptides. This trend was also supported by recent analysis of the Waltz database using classifying autoencoders (Tro et al. 2019), further recommending the pattern.

i. Energy Terms

The force field terms are as follows.

1. The bond potential is of the form:

$$U_{bond} = \sum_{bonds(ij)} \left(\frac{1}{2}\right) K_{b(ij)} \left(r_{ij} - r_{0ij}\right)^2$$

For which $K_{b(ij)}$ = 200.0 kcal/mol and r_{0ij} = 2.0 A

2. The angle potential is of the form:

$$\mathrm{K_{angles}} = \sum_{angles(ijk)} \frac{K_{\theta(ijk)}}{2} \left(\theta_{ijk} - \theta_{0ijk}\right)^2$$

Where $K_{\theta(ijjk)}$ = 40.0 kcal/mol, θ_{0iXk} = 120.0° and θ_{0iYk} = 180°.

3. Dihedral potentials are of the form:

$$U_{dihed} = \sum_{dihedrals(ijkl)} D_{ijkl} \cos\left(3\,\alpha - \delta_{ijkl}\right) - G_{ijkl} \cos\left(\alpha - \delta_{ijkl}\right)$$

Parameters for quadruplets are listed in Table 1.

Ijkl	D (kcal/mol)	G (kcal/mol)	Δ (degrees)
XYXY, YXYX	-0.25	-0.25	180
Sequence 1 (Rigid)			
CXXH, HXXP, PXXH, HXXA	0.0	-1.115	180.0
Sequence 2 (Flexible)			

CXXH, HXXP, PXXH, HXXA	0.0	-2.0	180.0

4. Nonbonded Interactions are of the form:

$$U_{NB} = \left(\frac{1}{2}\right)\sum_{i \neq j} 4\,\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \lambda\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{C_c q_i q_j}{r_{ij}}$$

Where $\varepsilon_{XX} = \varepsilon_{YY} = \varepsilon_{XY} = 0.5$, $\varepsilon_{HH} = 0.3$, $\varepsilon_{PP} = 0.04$, $\varepsilon_{HP} = 1.0$ and $C_c = 16.603$ kcal Å $mol^{-1} e^{-2}$. Beads A and C have charges q_A = -1 e and q_c = 1 e respectively. Parameter σ has the value 2.0 for every pair except XY, where $\sigma_{XY} = 3.0$. $\lambda_{XX} = \lambda_{YY} = \lambda_{XY} = \lambda_{HH} = \lambda_{PP} = 1.0$ and $\lambda_{HP} = 0.01$. For any pair of a side chain bead (C,H,P,A) and a backbone bead (X,Y) or else a pair involving a terminus bead (E) the parameters are ε = 1.0, σ = 2.0 and λ = 0.0. Energy constants are in kcal/mol and distance constants are in units of angstrom.

ii. Simulations & Nematic Order Parameter

Simulations were run in the molecular dynamics package NAMD (Phillips et al. 2005) using a Langevin implicit solvent. 108 copies of the peptide model were placed in a cubic box of 102 angstroms on a side. Periodic boundary conditions were applied, and the temperature brought to 305 K after scattering the peptides at a high temperature in absence of pair-wise potentials. A timestep of 10 fs was used, while randomly seeded velocity initialization promoted the evolution of multiple trajectories from the same initial conditions. Systems were then relaxed using a brief minimization to eliminate strain from scattering, and then run for 750 ns.

First-pass analysis of the trajectories was conducted using the what is referred to as the nematic order parameter, λ_p which is described in detail elsewhere (Eppenga and Frenkel 1984; Stephen and Straley 1974; Saupe 1968). Briefly, it is calculated as the highest eigenvalue of the always-diagonalizable 3-by-3 matrix Q_{ab} where:

$$Q_{ab} = \frac{1}{2N}\sum_{i=1}^{N}(3u_a^{(i)}u_b^{(i)} - \delta_{ab})$$

where a and b are x, y or z; N is the number of peptides, δ is the Kronecker delta, and u_a and u_b are the x, y, or z components of the U_p vector described above. This expression is a metric of overlap of the monomers' orientations. Eigenvalue decomposition yields eigenvectors, the primary of which is parallel with the fibril axis. The corresponding eigenvalue is the nematic order parameter, λ_p, which describes the degree of that alignment along the major axis. In other words, the nematic order parameter describes the degree of alignment of the carbon back bones in the system In a perfect fibril stack this value would approach one, however in practice fibril stacks are associated with values around 0.8. This choice of order parameter helps differentiate between two ordered but fundamentally different structures (the fibril stack and the beta-barrel-like structure) which were not appropriately discriminated between by the prior formulation of the nematic order parameter (Morriss-Andrews, Bellesia, and Shea 2012).

iii. Variational Autoencoder

A variational auto encoder is a method of dimensional reduction using an artificial neural network. It has been described in detail elsewhere (Kingma and Welling 2013), but is briefly described here. Artificial neural networks are a machine learning system inspired by biological neurons. Much like biological neurons, each node in an artificial neural network takes in many inputs, depending on those inputs the node is activated or deactivated, and the node's output along with all other nodes in its layer are used as inputs for the next set of nodes. The layout of these nodes is depicted in Figure V-2. Each gray arrow has a fitting parameter associated with it. The fitting parameters are fit to the data such that when an input, here data derived from a single frame of the trajectory, is fed into the model it reproduces that input in the reconstructed nodes. Note, a frame in this context is the state of all atomic data (coordinates, velocity and physical properties) at a specified time point of the simulation. Importantly, the input data is first reduced to a single number — encoded — before being reconstructed — decoded. That single number is termed the latent space value (LSV). In order to

decode that value with the highest fidelity possible, similar inputs will be encoded to similar points in latent space. This means that frames in the trajectory which are similar will encode to similar points in latent space.

In terms of an order parameter, frames in the beginning of the simulation are similar to other frames in the beginning of the simulation and will encode to similar point in latent space. As the simulation begins to evolve, each frame's data will encode to a different point in latent space. In this way we can plot the LSV of each frame as a function of time and this should give us an idea of when our system is changing.

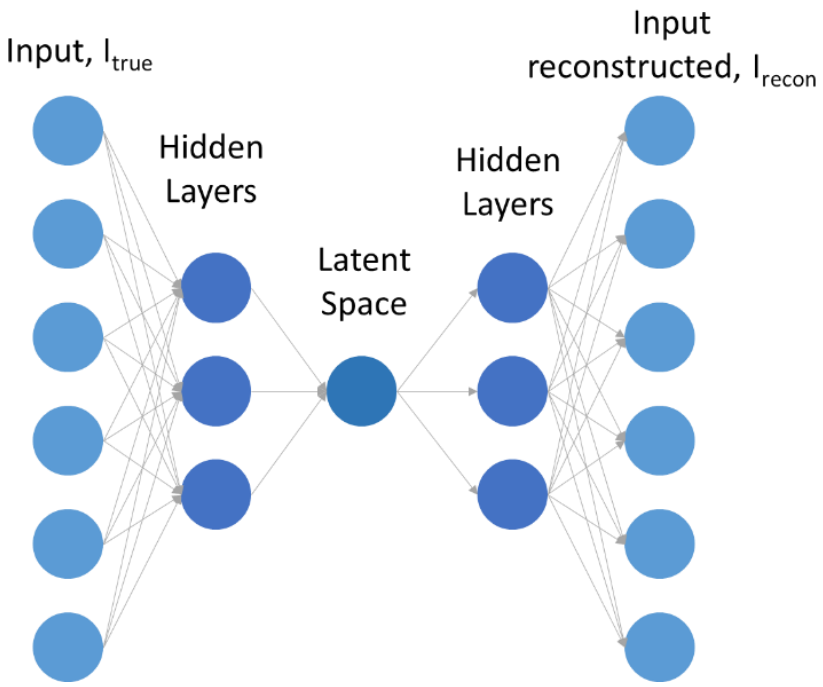


Figure V-2 A simple depiction of a variational autoencoder. Each dot represents a node. The gray arrows show how information is passed from node to node. Inputs are fed into nodes on the left, and the network is fit to reconstruct those inputs in the nodes on the right.

Artificial neural networks were built using the Keras python package (Chollet and Others 2015), using Google's Tensorflow (Martín Abadi et al., n.d.) as a backend. Learning sets were representations

of frames from the simulation. The simulation was run for 750 ns and saved every 0.1 ns resulting in 7500 frames or input data points for the model. Each frame was represented as the internal coordinates of the system (dihedral angles, molecular angles and bond lengths, 7236 total features in each frame). A full discussion of hyperparameters are in the supporting information and code has been posted on github at https://github.com/Michael-Tro/VAE_OrderParameter.

iv. Reconstruction Analysis

For the reconstruction analysis, LSVs of interest were chosen based on correlating LSV behavior with visual inspection over the course of the trajectory. For example, over the course of the simulation the LSV typically reaches a final value and remains at that value for the remainder of the trajectory. Presumably this means the system has reached some final state. Visual inspection of the trajectory confirms no large structural shifts occur once the LSV reaches its final value. Typically this final state is used as a reference state, and compared to other LSVs of interest (the beginning of a rise in LSV for example). Each LSV (including the reference) was then decoded into reconstructions.

The residual, $r = I_{ref}^{recon} - I_{int}^{recon}$, was calculated for each dihedral angle, bond length, and bond angle (or, generally, feature) in the system, where I_{ref}^{recon} represents the value of that feature in the reference reconstruction, and I_{int}^{recon} the value of that feature in the reconstruction of interest. To identify dihedral angles that contributed to changes in latent space, dihedral angles which satisfied some residual threshold were identified, where the threshold condition was varied. Thresholds presented here are ones which best demonstrate the change of interest.

D. Results

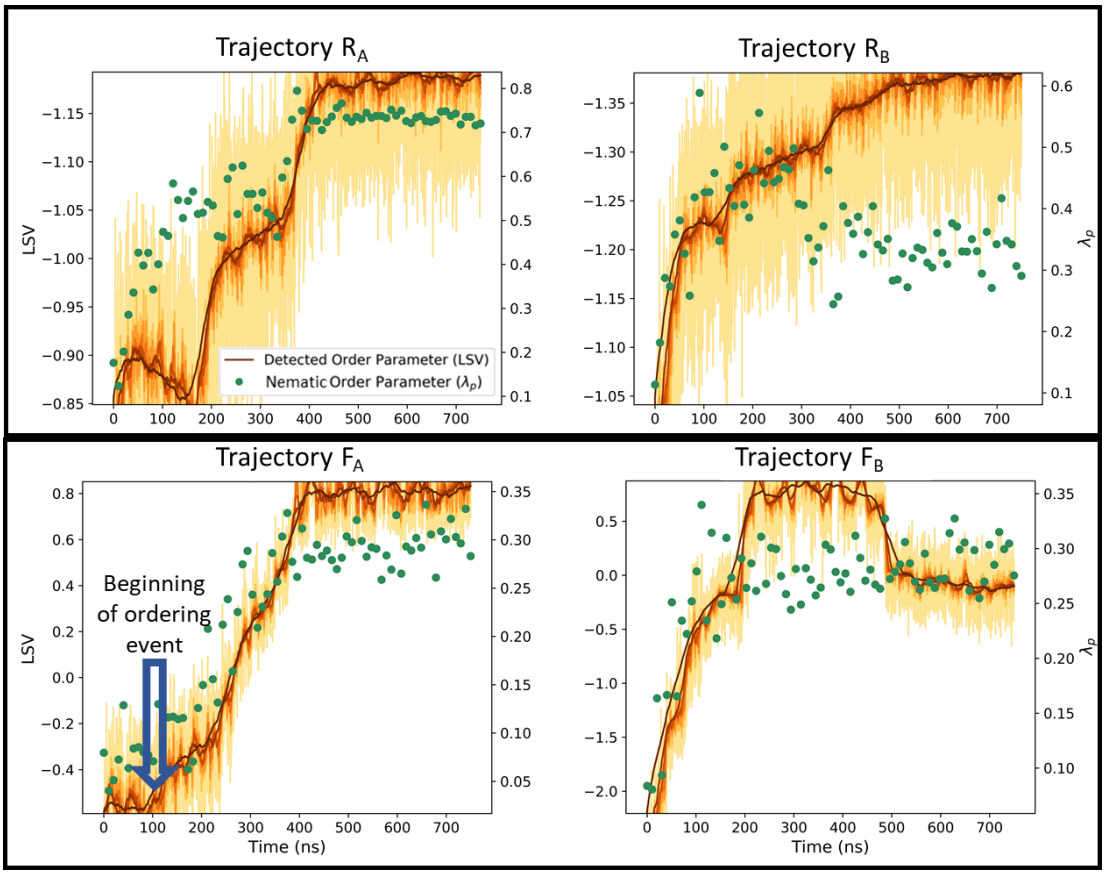


Figure V-3 Parameterizations of representative simulations. Trajectory (R_A)undergoes an evolution of rigid sequences from bulk dispersion progressing toward a final stacked beta-sheet state via the assembly of smaller stacks from bulk and subsequent rearrangement. In trajectory (R_B) the system relatively disordered beta-barrel-like intermediate forms, but then rearranges into a stack. In the flexible sequence trajectory (F_A), a liquid droplet coalesces from bulk and then undergoes a phase transition to solid. Trajectories (R_B) and (F_B) have been chosen specifically to illustrate instances where the LSV and the nematic order parameter diverged.

This section will proceed as follows. The results of representative simulations for the rigid sequence will be analyzed using traditional methods: the nematic order parameter and visual inspection. This analysis will be compared to an analysis using the automated order parameter (the LSV). Discussion will include Trajectories (R_A) and (R_B) represented by both the nematic order

parameter and LSV in Figure V-3. Emphasis will be on comparing where the either the nematic or automated order parameter fail to reconcile with visual inspection.

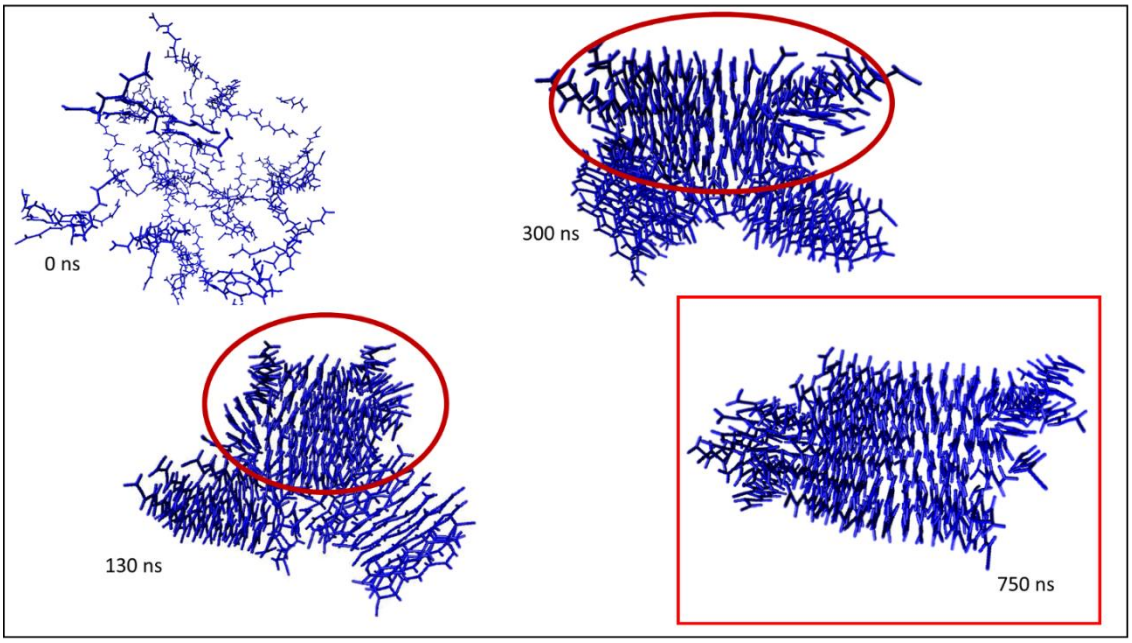
The results of a representative simulation for flexible sequence simulation (F_A) will then be treated in a similar manner, with a discussion of how the automated order parameter can be used to identify regions of interest within the simulation. Finally, a sampling issue present in a flexible simulation (F_B) will be analyzed and anomalous behavior of the automated order parameter discussed, with means of isolating the cause provided.

i. Analysis of Rigid Peptides

Simulations using the rigid sequence parameters were associated with the relatively rapid formation of states of stacked peptide sheets that resemble amyloid fibrils. These stacks are exceptionally stable, with little evolution away from these states. The primary means of assembly for these systems involves the alignment of existing stacks (seen in R_A), or the slow recruitment from nearby beta-barrels to slowly grow the stacks where possible (seen in R_B). In both cases presented here the nematic order parameter is at odds with the LSV. By inspecting the trajectories we are able to deduce the reasons for the divergences.

a. Trajectory R¬A: Nematic Parameter Implies Erroneous Growth

In trajectory (R_A) there is an apparent rapid growth of a fibril structure as evidenced by the rapid increase of the nematic order parameter (Figure V-3) in the earliest few hundred nanoseconds of the



simulation, followed by a lag period before a final burst of growth. The LSV, however, predicts a

Figure V-4 A visual summary of Trajectory (R_A). Over the length of time consider, a major fibril stack capped by beta-barrel-like structures emerges (circled in red), along with two minor fibril stacks. Subsequent rearrangement results in the formation of the final state, largely a single fibril stack with only a few disordered peptides on the ends.

plateau region during the initial simulation period. By visually inspecting the simulation, the source of the disagreement becomes clear. The middle two images (130 ns and 300 ns) in Figure V-4, Trajectory R_A shows the beginning and the end of the first major transition between plateaus in the LSV (between 130 ns and 300 ns). A noteworthy transformation is occurring during this period, in which the largest fibril stack (circled in red) is reorganizing the rest of the cluster by recruiting peptides from the more disordered regions. This is an important moment for the sequence of events – this major stack eventually becomes the template for the realization of the relatively complete fibril stack at the end of the trajectory. The nematic order parameter fails to find this transition, and seems to indicate the fibril core condenses somewhat immediately out of bulk. This occurs due to a coincidentally parallel orientation of the major stack and minor stack at about 130 ns – the nematic order parameter treats

these separated moieties as belonging to the same stack since they are in planar alignment. The matrix Q artificially diagonalizes to have an eigenvalue implying a higher degree of connected fibrilization than is actually present in the system.

b. Trajectory RB: Nematic Parameter Erroneously Implies Transient Ordering With No Resulting Fibril

Trajectory (R_B) illustrates a mechanism by which a mixed, largely immobile extended structure condenses out of bulk. This structure consists of beta-barrel-like structures or fibril stacks. Over the course of the trajectory, sections of beta-barrels or stacks get absorbed into one of two primary fibril stacks. By the end of the time studied, two major fibril stacks are formed, joined by a beta-barrel moiety. This is visually summarized in Figure V-5.

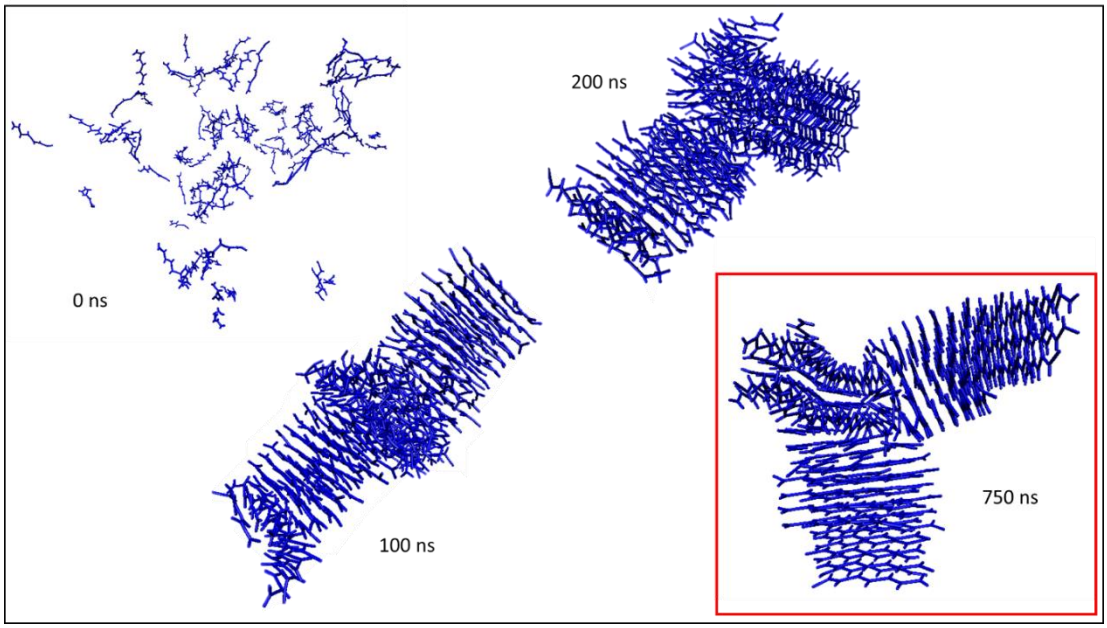


Figure V-5: A visual summary of Trajectory (R_B). Over the length of time considered, two major fibril stacks form and are joined by a beta barrel region. Growth occurs by the absorption of smaller stacks or the reorientation of peptides

Trajectory (R_B) provides a good example of how the nematic order parameter is vulnerable to oversights that the LSV detects. The nematic order parameter suggests the system initially starts by ordering and then suddenly undergoes a dramatic disordering around 200-300 ns (Figure V-5). This is at odds with the analysis of the LSV, which continues to increase over the course of the entire

trajectory. This trajectory, with images of the major structures, is presented in Figure V-5. Broadly, two fibril stacks form early in the trajectory (depicted at 100 ns in Figure V-5), resulting in both the nematic order parameter and the LSV to increase in value until about 200 ns. These two stacks have planar vectors that point roughly in the same direction, resulting in the relatively large value for the nematic order parameter. At around 200 ns, however, the two stacks twist out of alignment, resulting in an "L" shape arrangement of the two major fibril stacks. Because the planar vectors are now perpendicular, the nematic order parameter now perceives a lower degree of order despite the fact the two fibril stacks are still present, but now just oriented differently in space. This oversight is not exhibited by the LSV, which is trained on data from the monomers and therefore considers the structure from the level of the individual molecules. Because this perspective is insensitive to global rearrangement of major, multimeric structures the LSV does not consider the perpendicular arrangement of the stacks to be disordered. In this sense the LSV outperforms the nematic order parameter because it recognizes that the parallel-aligned fibril stacks are structurally quite similar to the perpendicular-aligned fibril stacks and does not register this difference as substantially impacting the overall structure.

ii. Analysis of Flexible Peptides

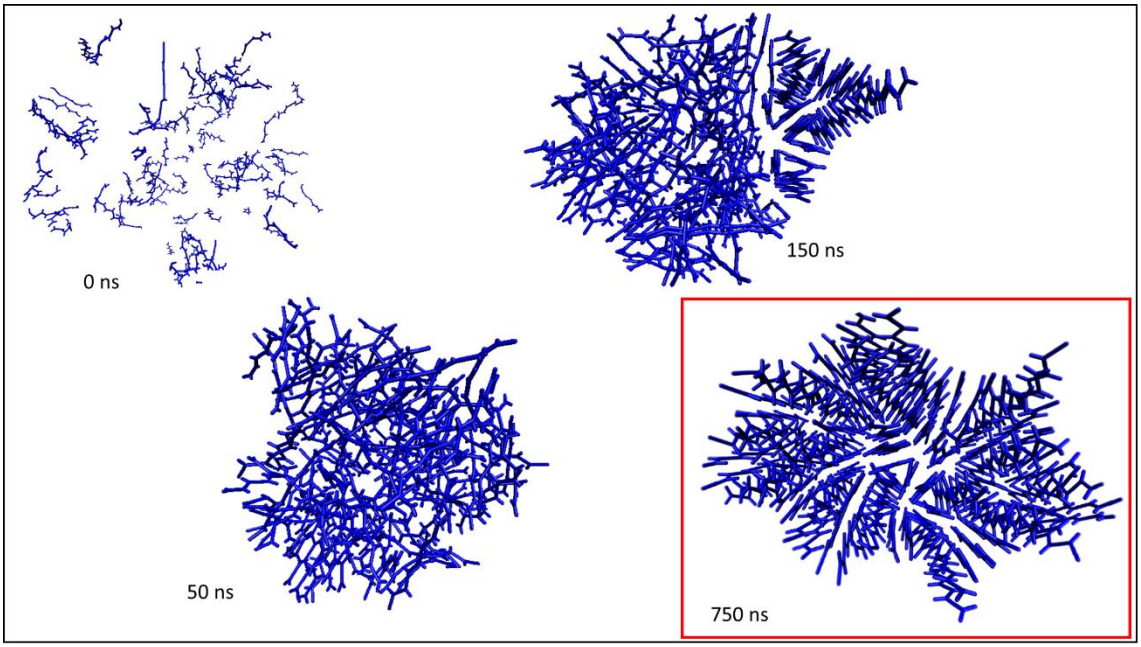


Figure V-6 A visual summary of Trajectory (F_A), showing the formation of a globular, spherical disordered phase with high internal mobility. The final state is a highly immobile beta-barrel-like solid. This system forms a solid nucleus early in the trajectory and subsequently undergoes a conversion from globule to solid.

The flexible peptides progressed through a different primary mechanism of growth than the rigid peptides. Relatively speaking, the rate at which the disordered peptides were attracted into a liquid-like mass was faster than the rate of stacking. The result was relatively few peptides aligned during the condensation of the bulk distributed peptides, with a slow transition from a liquid peptide mass, characterized by relative conformational flexibility for conformers, into a solid mass, characterized by peptides 'locked' into a beta-barrel-like conformation. The initial state will be referred to as 'dispersed', the fluid intermediate as the 'globule' and the final, solid state as 'beta-barrel-like'. This transition bears resemblance to similar liquid-to-solid transitions observed in literature, relating to coacervation processes of amyloid-forming systems.

a. Trajectory F_A: Nucleus Detection and Reconstruction Analysis

In trajectory (F_A), depicted in Figure V-6, the first stable BBL nucleates around 100 ns (see blue arrow in trajectory (F_A) in Figure V-3). During this initial phase, the nucleic dihedrals (those which have already adopted the BBL state) must be similar to those of the final state (where all the peptides are in the BBL state). Using this knowledge, the nucleus may be found by the reconstructing dihedrals from the LSV as it begins to increase (presumably due to an initial ordering event), and from the final LSV (when all peptides are in the BBL state). Then the subset of dihedrals whose reconstructed values from the early LSV which most closely resemble their values in the reconstructions from the final LSV, are most likely to belong to the nucleus. These dihedrals are highlighted in Figure V-7.**Error! Reference source not found.** for visual inspection, which confirms these dihedrals are centered in what appears to be the nucleus. Note, this is not the same as a trivial comparison of the original frames from those points in time. The VAE's reconstruction condenses the information in a way that transient obscuring factors, such as thermal fluctuations, are removed and thus comparing reconstructions automatically adds a layer of information distillation that might have otherwise required researcher intuition or manipulation to account for.

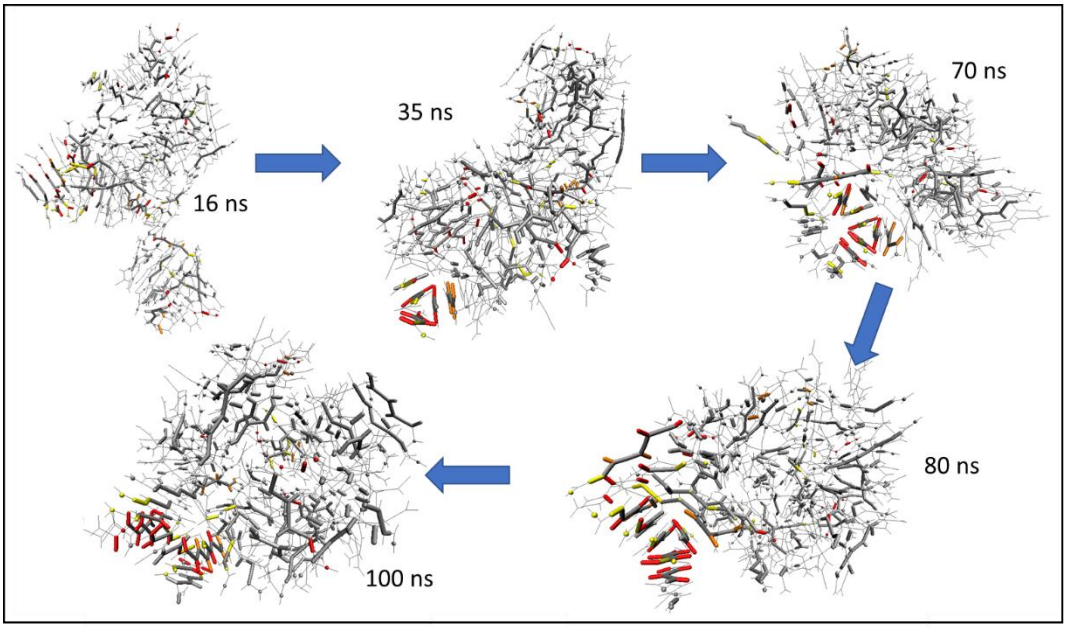


Figure V-7 The formation of the nucleus in trajectory (F_A) as labeled comparing reconstructions of the final state to reconstructions of the initial nucleation. Atoms labeled with red were found to be the most similar between the moment of nucleation and the final stable state, with orange, yellow and finally bold gray denoting increasingly weak degrees of similarity. The process successfully labels many of the atoms involved in the formation of the solid nucleus and helps identify a 'hot spot' of mechanistic behavior.

This may be taken further to look at various subsets and monitor them throughout the trajectory to find subsets that contribute to features of the LSV. Here we do this to find the set of dihedrals which correspond most to the change in order parameter over the course of the trajectory. In Figure V-8, that subset is found by comparing initial state (globule) to the final state (beta-barrel-like), and monitoring average value of that subset during the course of the trajectory. The darkest purple curve shows the subset of dihedrals that changed the most according to the reconstructions, while lighter purple indicates moving the threshold to allow a smaller amount of change. We find that by looking at just the 10% of the dihedrals that change the most according to the reconstruction, we are able to reproduce the behavior of the LSV order parameter with decent fidelity. By plotting the average value of the rest of the dihedrals (the corresponding green line for each purple line), it is made clear that

the dihedrals which were not contributing to the order parameter are filtered out by the VAE. When the threshold reaches about 50% there is little change in the dihedral values as the simulation progresses, indicating a cutoff where the dihedrals are not as affected by the fibrillization process.

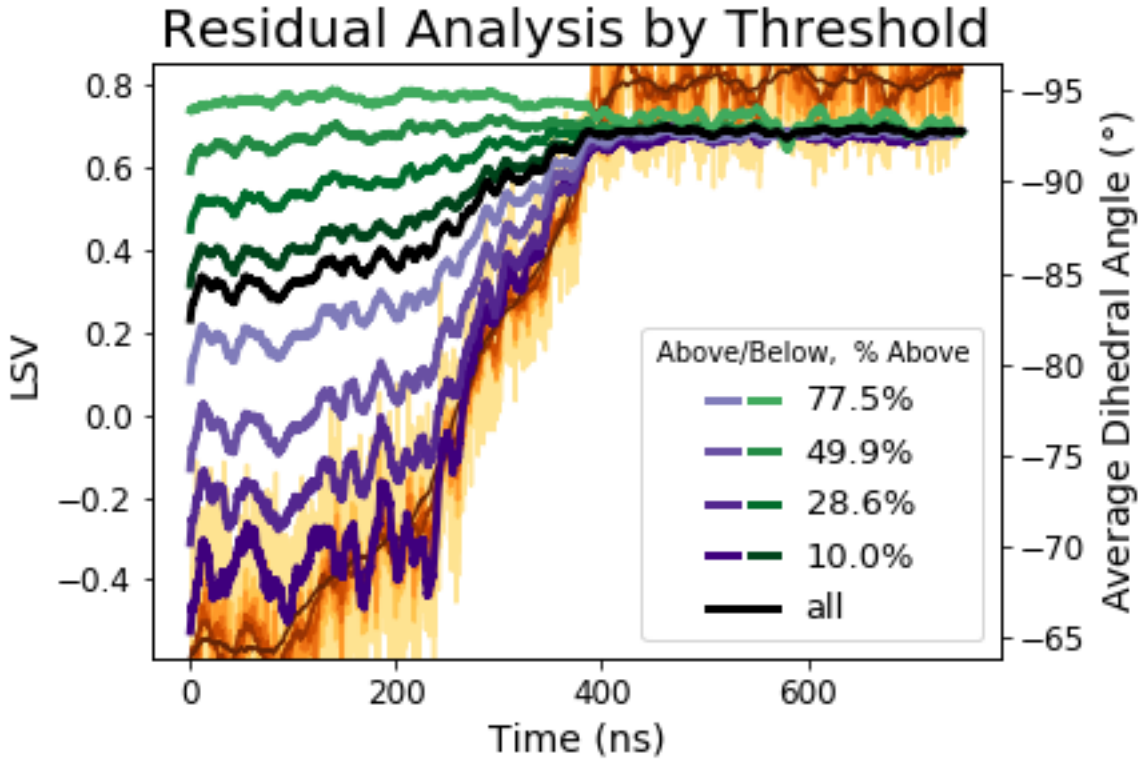


Figure V-8. Subsets of dihedrals were chosen by calculating the residual between the beginning state and the end state, then selecting the dihedrals which had the largest residuals. The threshold for including a dihedral was scanned to from including 99% of the dihedrals to excluding 99% of the dihedrals. The purple lines represent the average value of the dihedrals above the threshold (these are the dihedrals the largest residual), while the green lines represent the rest of the dihedrals. The black line is the average of all dihedral angles. These lines have been smoothed by a rolling 10 ns averaging, and they have been shifted to have the same end point.

This is useful for systems of scale, when identifying relatively small portions of large simulation volumes might require an otherwise difficult search. Furthermore, it more generally enables the comparison of any point in the trajectory with any other point, and swift labeling of the VAE's interpretation of similar regions. Thus for dynamical systems with high degrees of noise or other convoluting factors, the VAE's encoded representation may serve as an effective method of cleaning up the simulation data for analysis.

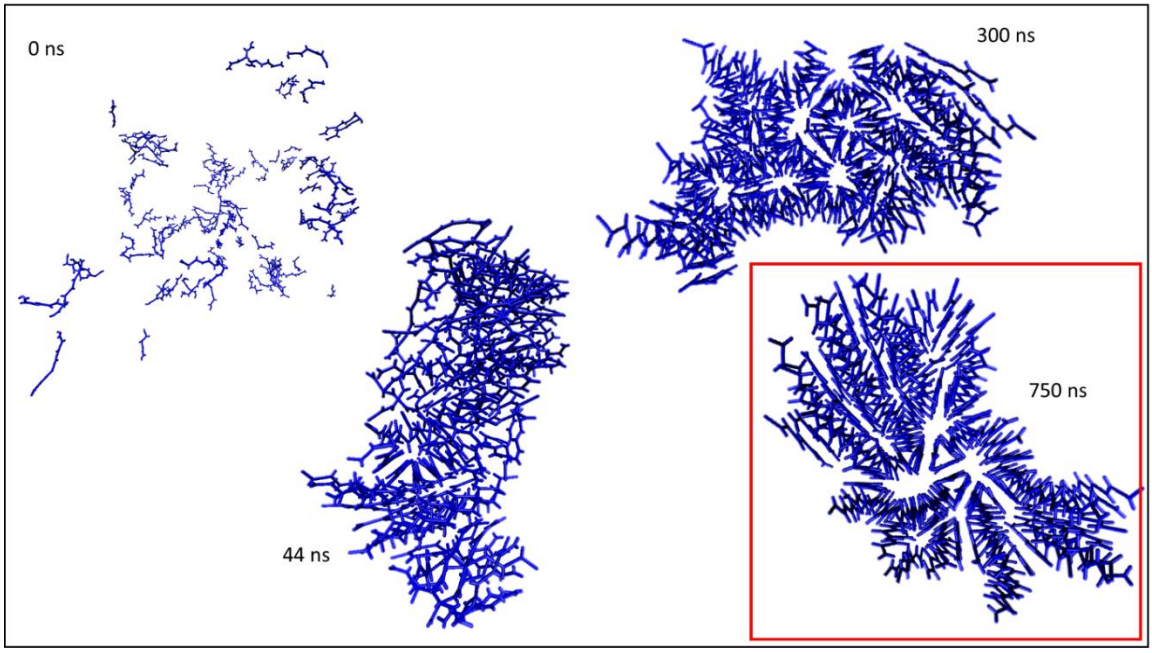


Figure V-9: Visual summary of Trajectory (F_B). The system condenses into a relatively extended, solid structure out of bulk. This solidifies into a beta-barrel-like structure quickly and remains in this state until the end of the simulation.

b. Trajectory F_B: Importance of Sampling

Trajectory (F_B) is an example of anomalous LSV behavior, and the authors emphasize this section is present to examine the flaws of applying variational autoencoding techniques to flawed data. Trajectory F_B was selected due the fact it poorly samples the liquid-to-solid ordering process of interest, and thus its analysis allows for the consideration of a potential pitfall when using an unsupervised method to study a system might be presumed to have an ordering process, but does not. Trajectory (F_B) has the most rapid ordering process of any of the flexible peptide systems,

reaching an axially aligned BBL state within 200 ns according to the nematic order parameter and by visual inspection. Consequently, its ensemble is composed entirely of BBL states, and thus its LSV depicts odd behavior that captures the shifting changes in the BBL structures but does not accurately parameterize the ordering process of interest. Using an VAE trained on trajectory (F_A) to encode trajectory (F_B) we achieve LSVs which better agree with the nematic order parameter and visual inspection (Figure V-10).

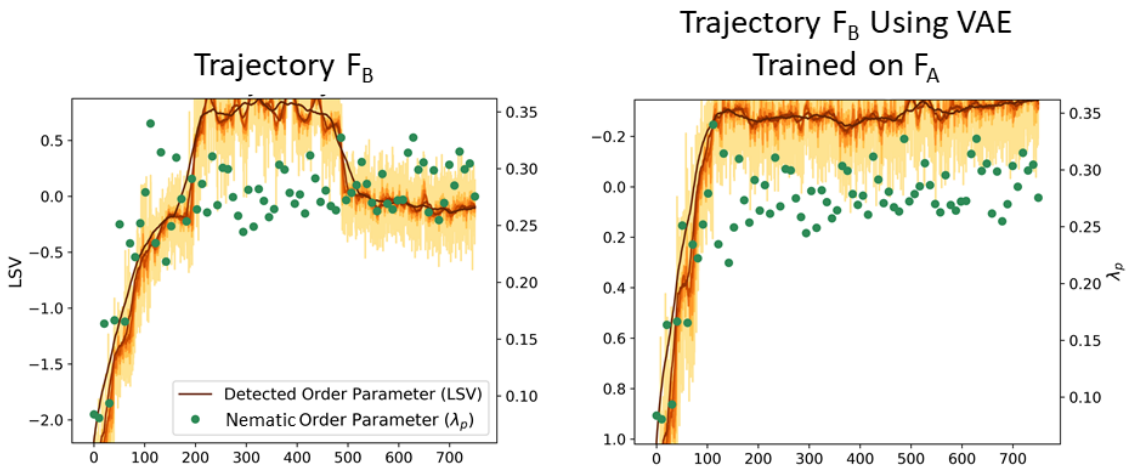


Figure V-10 Trajectory F_B's LSV anomalously shows an increase from 200 ns to 500 ns. This LSV behavior is not seen when an order parameter is generated by encoding the trajectory of F_B with an VAE trained on F_A, suggesting that this is related to training, and not a feature of the trajectory itself.

To further investigate this anomalous behavior the trajectory was examined by analyzing the reconstructions, much like Figure V-8. By reconstructing representations of the system from the LSV of the plateau between 200 to 500 μs, and the LSV after the plateau we can identify the change in dihedral values that the plateau represents. The dihedrals that changed the most in the plateau were isolated and their average value was plotted alongside the average value of all dihedrals over the course of the simulation Figure V-11. Comparing the behavior of this subset of dihedrals to the LSV we can see the analysis method successfully found the subset of dihedrals that contribute to the

anomalous behavior of the LSV. Visual inspection, in a similar manor as the nucleus detection above, of these dihedrals did not reveal behavior that was of substantial interest and acts as evidence that the LSV's characterization was not useful in providing information about general trends in the trajectory's evolution.

This evidence, along with the extreme over representation of the BBL state in in this trajectory, leads to the conclusion: if the model is trained on data with overrepresentation of a given state, as relevant when sampling a mechanistic pathway with potentially brief representation of intermediate states, then it may learn to differentiate small changes within the over-represented state not just transformations between states of interest. This is where researcher oversight is important; the algorithm must have good data to learn, or else it may focus on numerically valid but scientifically uninteresting aspects of the data. While many machine learning techniques present exciting and novel ways of analyzing data, it remains extremely important that researchers have a very clear idea of what they are trying to detect when they implement them. This allows reconciliation of the unsupervised method's result with research intuition, and that remains an important check on this evolving field of analytical approaches.

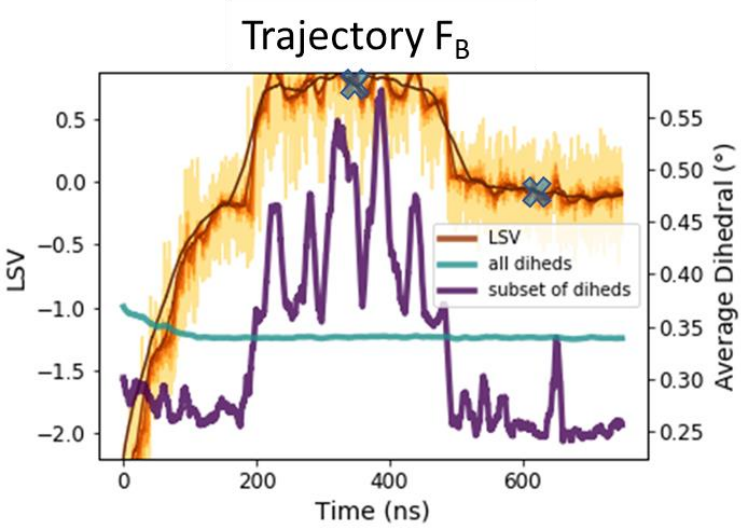


Figure V-11 Trajectory F_B with its 'plateau' region marked. This is the major consequence of its poor sampling and its contributing features (purple) are isolated for examination. Turquoise is a curve representing the behavior of all dihedrals.

The ability to perform this reconstruction analysis represents an important means to infer the VAE's process. Generally, lack of insight resulting in misinterpretation is a major, and valid, criticism of the so-called 'blackboxes' of other artificial neural network techniques. When something seemingly anomalous occurs, it is important to be able to inspect the neural network and validate whether or not it is an appropriate quantification of the system, something which we are able to do by using a VAE.

E. Conclusions

A coarse-grained model of amyloid-prone peptides was used to simulate the formation of amyloid fibrils. It was found by visual inspection that, depending on the flexibility of the monomer backbones, different mechanisms were possible for reaching the final state, supporting conclusions from previous studies on similar models (Bellesia and Shea, 2009; Morriss-Andrews and Shea, 2012). True fibril stacks were observed as being the final state of more rigid peptides, while more flexible peptides were observed to form more cylindrical intermediate structures whose kinetic stability was high. In both cases, a beta-barrel-like state and a fibril stack state were noted to have low internal mobility of the constituent monomers consistent with a solid state. However, a highly mobile liquid phase was also observed especially in the more flexible systems. In all cases, the trajectories approached a solid final state with little further evolution in the time spans studied.

The results provided here represent a survey of techniques for using Variational Autoencoder machinery in automatically learning an order parameter to characterize the ordering of a simplified model of solution-phase amyloid aggregation. In addition, this represents an increase in the size of systems treated by autoencoder methodologies and shows its potential for application even for systems that possess considerable scale and a diversity of different behaviors and processes. The order parameters (LSVs) learned from serial trajectories of aggregating monomers were compared to the nematic order parameter widely used in aggregation literature. The VAE required no implicit

enforcement of a time series to properly organize the data, nor did they require significant feature selection by the researcher. They were trained on states encoded as monomeric internal coordinates (the bond lengths, the bond angles and the dihedrals defined in the system). The success of these models demonstrates it is possible to characterize the state of the system by these monomer features alone. Reconstruction-based methods were shown to facilitate an understanding of which features were most significant to characterization of the transition from the dispersed phase to a final solid state.

Weaknesses in the generation of order parameters using variational autoencoders were addressed, including a need for well-sampled trajectories and good statistics of the transitions being characterized. Remediations in the form of strategies for elucidating how the VAE was 'thinking' were offered as a means of validating the found solutions, specifically using the VAE's reconstructions to study which elements of the simulation were contributing to its characterization. This allows for targeted study of the visualized trajectories and served as a useful means of confirming or excluding the solutions of variational autoencoders as useful to characterization of the system. The process additionally facilitates the analysis of large amount of simulation data by acting a means of filtering transient behaviors and comparing points in time with one another in a streamlined fashion.

The work presented here indicates, VAEs represent a powerful tool in the methods of information compression, filtration and denoising. Additionally, they are unsupervised algorithms, requiring relatively little preparation of data to be used. As these algorithms become more prevalent more work will be needed to characterize their effectiveness on systems of scale, their flexibility, and successfully apply them.

F. References

Bellesia, Giovanni, and Joan-Emma Shea. 2009. "Diversity of Kinetic Pathways in Amyloid Fibril Formation." *The Journal of Chemical Physics* 131 (11): 111102. https://doi.org/10.1063/1.3216103.

Bernstein, Summer L., Nicholas F. Dupuis, Noel D. Lazo, Thomas Wyttenbach, Margaret M. Condron, Gal Bitan, David B. Teplow, et al. 2009. “Amyloid-β Protein Oligomerization and the Importance of Tetramers and Dodecamers in the Aetiology of Alzheimer’s Disease.” *Nature Chemistry* 1 (4): 326–31. https://doi.org/10.1038/nchem.247.

Chen, Mingchen, Nicholas P. Schafer, and Peter G. Wolynes. 2018. “Surveying the Energy Landscapes of Aβ Fibril Polymorphism.” *The Journal of Physical Chemistry B* 122 (49): 11414–30. https://doi.org/10.1021/acs.jpcb.8b07364.

Chen, Wei, Aik Rui Tan, and Andrew L. Ferguson. 2018. “Collective Variable Discovery and Enhanced Sampling Using Autoencoders: Innovations in Network Architecture and Error Function Design.” *The Journal of Chemical Physics* 149 (7): 072312. https://doi.org/10.1063/1.5023804.

Chiricotto, Mara, Simone Melchionna, Philippe Derreumaux, and Fabio Sterpone. 2019. “Multiscale Aggregation of the Amyloid Aβ 16–22 Peptide: From Disordered Coagulation and Lateral Branching to Amorphous Prefibrils.” *The Journal of Physical Chemistry Letters* 10 (7): 1594–99. https://doi.org/10.1021/acs.jpclett.9b00423.

Chollet, François, and Others. 2015. *Keras*. https://keras.io.

Eppenga, R., and D. Frenkel. 1984. “Monte Carlo Study of the Isotropic and Nematic Phases of Infinitely Thin Hard Platelets.” *Molecular Physics* 52 (6): 1303–34. https://doi.org/10.1080/00268978400101951.

Fitzpatrick, Anthony W. P., Galia T. Debelouchina, Marvin J. Bayro, Daniel K. Clare, Marc A. Caporini, Vikram S. Bajaj, Christopher P. Jaroniec, et al. 2013. “Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril.” *Proceedings of the National Academy of Sciences* 110 (14): 5468–73. https://doi.org/10.1073/pnas.1219476110.

Geddes, A. J., K. D. Parker, E. D. T. Atkins, and E. Beighton. 1968. “‘Cross-β’ Conformation in Proteins.” *Journal of Molecular Biology* 32 (2): 343–58. https://doi.org/10.1016/0022-2836(68)90014-4.

Guenther, Elizabeth L., Peng Ge, Hamilton Trinh, Michael R. Sawaya, Duilio Cascio, David R. Boyer, Tamir Gonen, Z. Hong Zhou, and David S. Eisenberg. 2018. “Atomic-Level Evidence for Packing and Positional Amyloid Polymorphism by Segment from TDP-43 RRM2.” *Nature Structural & Molecular Biology* 25 (4): 311–19. https://doi.org/10.1038/s41594-018-0045-5.

Ilie, Ioana M., and Amedeo Caflisch. 2019. “Simulation Studies of Amyloidogenic Polypeptides and Their Aggregates.” *Chemical Reviews* 119 (12): 6956–93. https://doi.org/10.1021/acs.chemrev.8b00731.

Jarrett, Joseph T., and Peter T. Lansbury. 1993. “Seeding ‘One-Dimensional Crystallization’ of Amyloid: A Pathogenic Mechanism in Alzheimer’s Disease and Scrapie?” *Cell* 73 (6): 1055–58. https://doi.org/10.1016/0092-8674(93)90635-4.

Kingma, Diederik P., and Max Welling. 2013. “Auto-Encoding Variational Bayes.” In *ArXiv:1312.6114*. http://arxiv.org/abs/1312.6114.

Maries, Eleonora, Biplob Dass, Timothy J. Collier, Jeffrey H. Kordower, and Kathy Steece-Collier. 2003. “The Role of α-Synuclein in Parkinson’s Disease: Insights from Animal Models.” *Nature Reviews Neuroscience* 4 (9): 727–38. https://doi.org/10.1038/nrn1199.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. n.d. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. tensorflow.org.

Morriss-Andrews, Alex, Giovanni Bellesia, and Joan-Emma Shea. 2012. “β-Sheet Propensity Controls the Kinetic Pathways and Morphologies of Seeded Peptide Aggregation.” *The Journal of Chemical Physics* 137 (14): 145104. https://doi.org/10.1063/1.4755748.

Nguyen, Hoang Linh, Pawel Krupa, Nguyen Minh Hai, Huynh Quang Linh, and Mai Suan Li. 2019. “Structure and Physicochemical Properties of the Aβ42 Tetramer: Multiscale Molecular Dynamics Simulations.” *The Journal of Physical Chemistry B* 123 (34): 7253–69. https://doi.org/10.1021/acs.jpcb.9b04208.

Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. 2005. “Scalable Molecular Dynamics with NAMD.” *Journal of Computational Chemistry* 26 (16): 1781–1802. https://doi.org/10.1002/jcc.20289.

Ray, Sourav, Stephanie Holden, Lisandra L. Martin, and Ajay Singh Panwar. 2019. “Mechanistic Insight into the Early Stages of Amyloid Formation Using an Anuran Peptide.” *Peptide Science* 111 (5): e24120. https://doi.org/10.1002/pep2.24120.

Rojas, Ana V., Gia G. Maisuradze, and Harold A. Scheraga. 2018. “Dependence of the Formation of Tau and Aβ Peptide Mixed Aggregates on the Secondary Structure of the N-Terminal Region of Aβ.” *The Journal of Physical Chemistry B* 122 (28): 7049–56. https://doi.org/10.1021/acs.jpcb.8b04647.

Saupe, A. 1968. “Recent Results in the Field of Liquid Crystals.” *Angewandte Chemie International Edition in English* 7 (2): 97–112. https://doi.org/10.1002/anie.196800971.

Stelzmann, Rainulf A., H. Norman Schnitzlein, and F. Reed Murtagh. 1995. “An English Translation of Alzheimer’s 1907 Paper, ‘Über Eine Eigenartige Erkankung Der Hirnrinde.’” *Clinical Anatomy* 8 (6): 429–31. https://doi.org/10.1002/ca.980080612.

Stephen, Michael J., and Joseph P. Straley. 1974. “Physics of Liquid Crystals.” *Reviews of Modern Physics* 46 (4): 617–704. https://doi.org/10.1103/RevModPhys.46.617.

Tro, Michael J., Nathaniel Charest, Zachary Taitz, Joan-Emma Shea, and Michael T. Bowers. 2019. “The Classifying Autoencoder: Gaining Insight into Amyloid Assembly of Peptides and Proteins.” *The Journal of Physical Chemistry B* 123 (25): 5256–64. https://doi.org/10.1021/acs.jpcb.9b03415.

Wang, Yihang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. 2020. “Machine Learning Approaches for Analyzing and Enhancing Molecular Dynamics Simulations.” *Current Opinion in Structural Biology* 61 (April): 139–45. https://doi.org/10.1016/j.sbi.2019.12.016.

Wehmeyer, Christoph, and Frank Noé. 2018. “Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics.” *The Journal of Chemical Physics* 148 (24): 241703. https://doi.org/10.1063/1.5011399.

West, Michael W., Weixun Wang, Jennifer Patterson, Joseph D. Mancias, James R. Beasley, and Michael H. Hecht. 1999. “De Novo Amyloid Proteins from Designed Combinatorial Libraries.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (20): 11211–16.

Westermark, Per, Arne Andersson, and Gunilla T. Westermark. 2011. “Islet Amyloid Polypeptide, Islet Amyloid, and Diabetes Mellitus.” *Physiological Reviews* 91 (3): 795–826. https://doi.org/10.1152/physrev.00042.2009.

Wetzel, Sebastian J. 2017. “Unsupervised Learning of Phase Transitions: From Principal Component Analysis to Variational Autoencoders.” *Physical Review E* 96 (2): 022140. https://doi.org/10.1103/PhysRevE.96.022140.

Xiong, H., B. L. Buckwalter, H. M. Shieh, and M. H. Hecht. 1995. “Periodicity of Polar and Nonpolar Amino Acids Is the Major Determinant of Secondary Structure in Self-Assembling Oligomeric Peptides.” *Proceedings of the National Academy of Sciences* 92 (14): 6349–53. https://doi.org/10.1073/pnas.92.14.6349.

Appendix I. Supplementary Information for The Classifying Autoencoder: Gaining Insight to Amyloid Assembly of Peptides and Proteins

a. Optimization

To optimize the model, we first need a metric for characterizing the effectiveness of a given model. Accuracy, as given in the following equation,

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{False\ Positives\ (FP) + False\ Negatives\ (FN) + TP + TN}$$

can be a poor metric in cases where one class is much greater than the other. Consider, for example, a situation in which 90% of all peptides were non-amyloid and 10% were amyloid. Then, by trivially classifying all inputs as non-amyloid, our classifier would obtain an accuracy of 90% without any real characterization of the classifying task. Thus, we sought a better measure of success.

The Matthews correlation coefficient (MCC):

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{\big((TP + FP)x(TP + FN)x(TN + FP)x(TN + FN)\big)}}$$

is a metric of agreement (or correlation) between the predicted class and the experimentally determined class over all samples in the data base [1]. The MCC ranges from -1 to 1. 1 is perfect correlation, 0 is no correlation, and -1 is perfect anti-correlation. We use this as the metric for comparing our models' effectiveness due to its robustness to unequal populations within classes.

Using the MCC to evaluate our model's performance, we optimized the so-called 'hyperparameters' to maximize this performance. Before using most machine learning algorithms, several hyperparameters must be chosen. Hyperparameters are parameters which are not fit during

training (the number of nodes per layer for example). There is generally more than one set of hyperparameters which yield a strong model, but poorly chosen hyperparameters can detrimentally affect performance.

In the next sections each hyperparameter is detailed. The CAE has three hyperparameters (relating to the terms of the loss function), as well as the typical hyperparameters for artificial neural networks specifying its architecture (the number of layers, and the number of nodes per layer).

Since the fitting parameters in the ANN are initialized to random numbers, there is a stochastic element to the fitting process. To address this, we train each set of hyperparameters 50 times. The performance of each model is logged, and the best models are saved for later analysis. Because of the stochastic nature of training the models, the reconstructed space between different trainings with the same parameters vary from one another slightly. It is thus important to draw conclusions from consensus between models. The figures presented here represent behaviors found in high-scoring models.

b. Weights of the Loss Function

The loss function defines the goals of the training process. During the training process the fitting parameters are varied until a minimum in the loss function is found. It is analogous to the square of the residual in linear regression. The classifying autoencoder has three terms in the loss function. The reconstruction term compares the reconstructed input to the original input. The prediction term compares the model's classification prediction (amyloid or not) to the classification from the database. Finally, the Kullback-Leibler (KL) divergence term relates to how much noise is added to the model during training [2]. Interestingly, we did not find the relative weighting of these terms affected the maximum MCC score a configuration can achieve (Fig. S1). We attribute this to the model being able to independently minimize the prediction term regardless of the magnitude of the other terms in the loss function. The weights do, however, affect the frequency with which a

model can find this maximum MCC, suggesting that while the global minimum remains constant, a poorly weighted loss function tends to get stuck in local minima during training, leading to a need for more trainings and thus inefficiency in the training process.

Once we found that the maximum MCC achievable was unaffected by the loss weighting, we investigated the reconstructive capabilities of the model. As shown in Fig. S1, every weight of the reconstruction term over 5 orders of magnitude can score an MCC of around 0.5. However, we notice a strong dependence on the minimum reconstruction loss the model can achieve at each weighting. We seek to minimize the reconstruction loss (ensuring fidelity of reconstruction), while simultaneously maximizing the MCC (representative of a successful representation of the classification function). The models in the bottom right of Fig. S1 achieve this.

c. Number of Layers

This plot also informs the number of hidden layers to choose. All models have the same number of nodes. A node is a unit cell of an artificial neural network, that stores a unit of the information learned during training. The depth is the number of hidden layers and the nodes per layer is the total number of nodes (24) divided by the depth. We find that depths of 1, 2, and 3 are found most frequently in the bottom right of Fig. S 1.

d. Hyperparameters

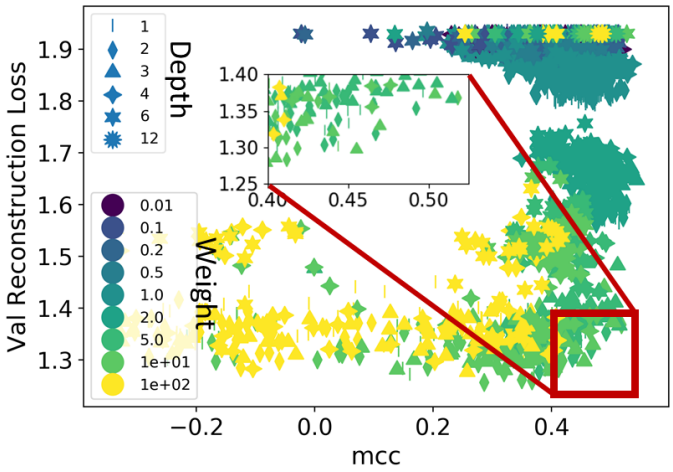


Figure S1: A scatter plot over 50 trainings with each combination of hyper parameters shown. The depth of the model is the number of hidden layers, while the total number of hidden nodes was held constant at 24. The weight is a scaler multiplied by the reconstruction term of the loss function. MCC is the Mathews correlation coefficient of the test set. The Val Reconstruction Loss is the loss of the reconstruction term on the test set before being multiplied by the weight.

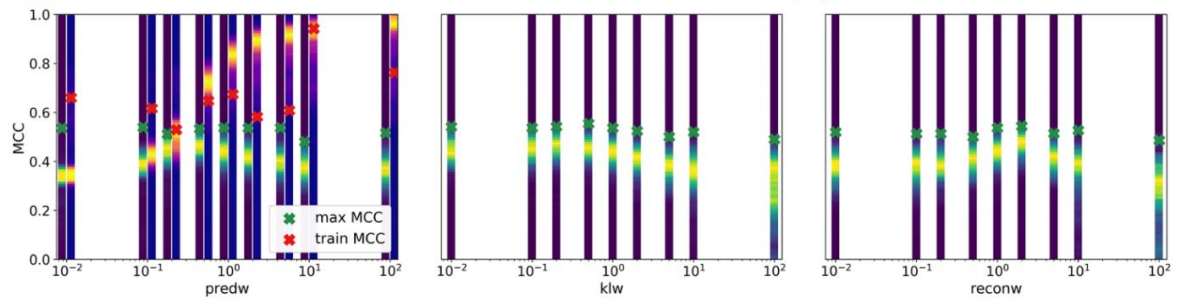


Figure S2 Histograms of models trained with differing weights in the loss function against MCC of each model. To optimize the weights in the loss function, each weight was varied over 5 orders of magnitude from 0.01 to 100, while holding the other two weights at 1. The heat bars represent the MCC distribution of models with the weight set to the value of the x axis. Light colors on the heat bars represent more models with that MCC. The green X denotes the best model at that weight—the model with the highest MCC on the test set at that set of hyperparameters; the red X is the MCC on the test set of that same model. There is little dependence in the maximum MCC (from the test set) from the change in weights. There is, however a change in the distribution of MCC depending on the weights. When the reconstruction weight (reconw) or the noise weight (klw) are raised to 100 the histogram shows a larger spread in MCC. Surprisingly as the prediction weight (predw) is increased, the peak of the green histogram (the test MCCs) decreases. This trend is explained by studying the red histogram, the training set MCC. When the prediction weight is set to high values the model overfits, and 'memorizes' the data in the training set. By doing so the model only learns a specific dataset and is not able to generalize to the test set.

e. Ion Mobility Measurements

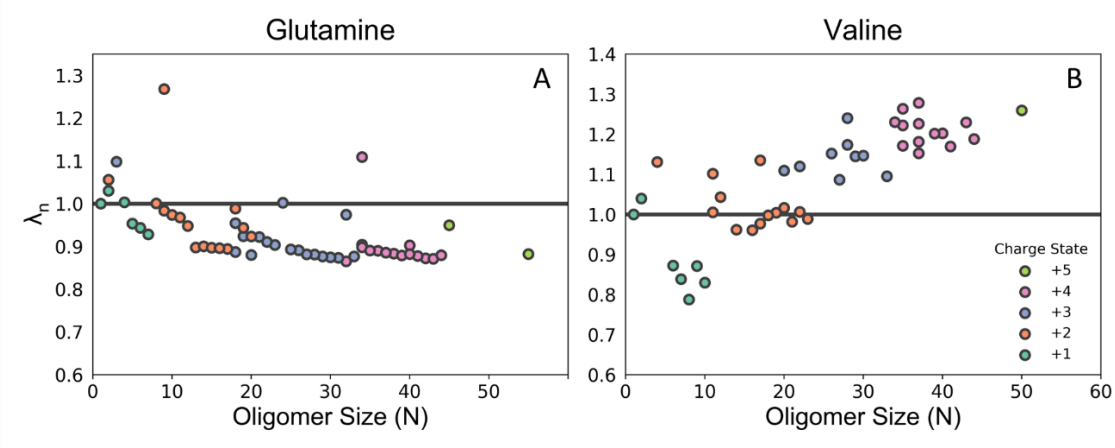


Figure S3 This figure demonstrates how isotropic deviation depends the number of amino acids the cluster — oligomer size. Here $\lambda_n = \frac{\sigma_n^{exp}}{\sigma_n^{iso}}$, and the line is where $\lambda_n = 1$. Recall this relates to isotropic deviation by the following equation $\Delta i_n = \left(1 - \frac{\sigma_n^{exp}}{\sigma_n^{iso}}\right)$. In this figure we see that isotropic deviation for glutamine is perhaps approaching some value as we reach higher oligomer sizes. We also note that the isotropic deviation is largely dependent on the charge state. If this was the case for all amino acids this would have been a good metric for isotropic deviation, however as we explored more amino acids we found that Valine remains oligomer size dependent in the oligomer range we are able to observe, thus we decided that we needed a metric which constant in both oligomer size, and charge state.

f. Arrival Time Distributions

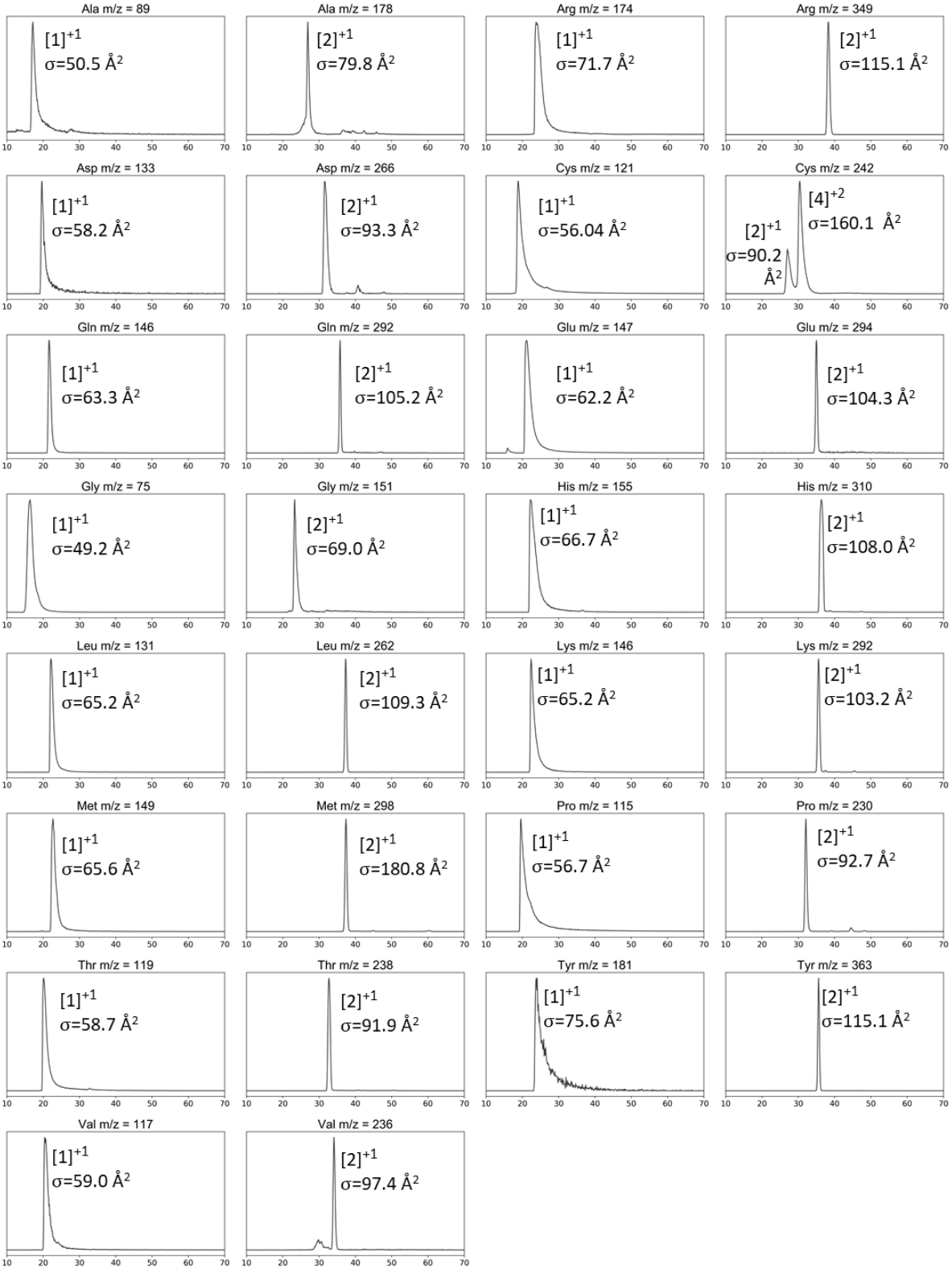


Figure S4 Arrival time distributions of each amino acid and the dimer of each amino acid. Using notation $[n]^z$ where n is the number of amino acids in the cluster, and z is the total charge of the cluster. The x axis is in units of milliseconds. Amino acids which are not in this figure (Asn, Ile, Phe, Ser, and Trp), are published elsewhere. [3,4]

g. Validation

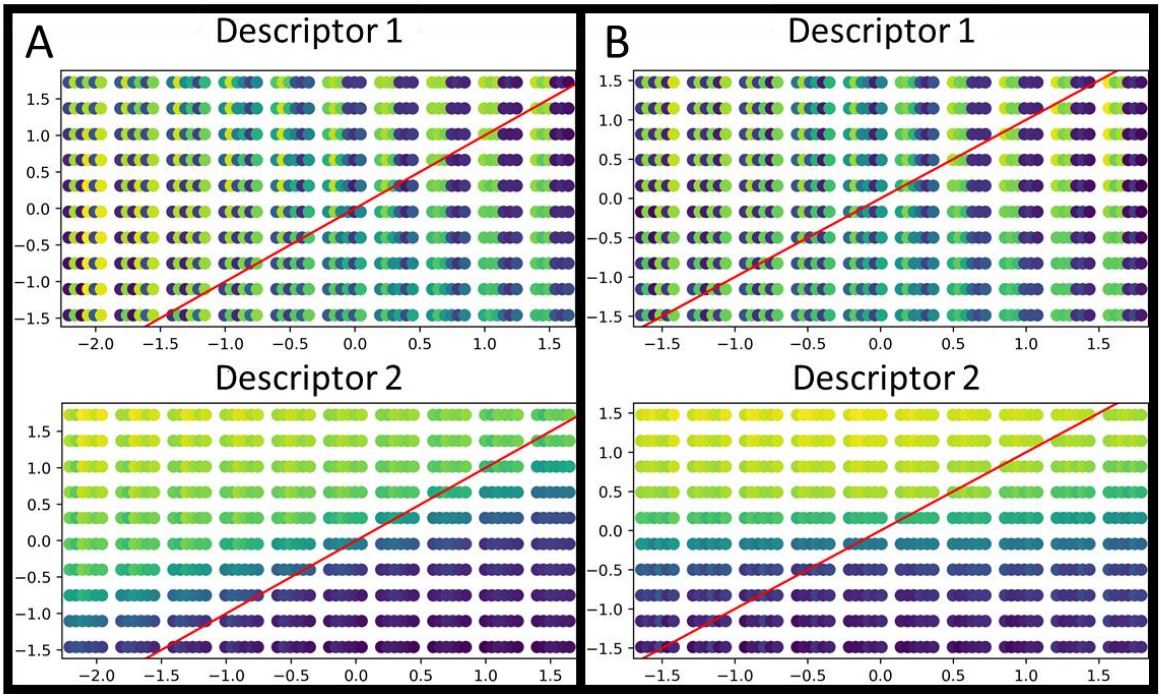


Figure S5 Reconstruction plots of Figs 4 D and E (A and B here respectively). The legend here is the same as Fig. S6. Here the positive motifs are LULULU or UUULLL in descriptor 1 and UUUUUU in descriptor 2, and the negative motif is LULULU or UUULLL in descriptor 1 and LLLLLL in descriptor 2. In top right of these figures the classification was made ambiguous. In the top right of the reconstruction plot of descriptor 2 UUUUUU penetrates the negative region more when there the classification was 60% classified positive (B) than when the classification was 80% positive (A).

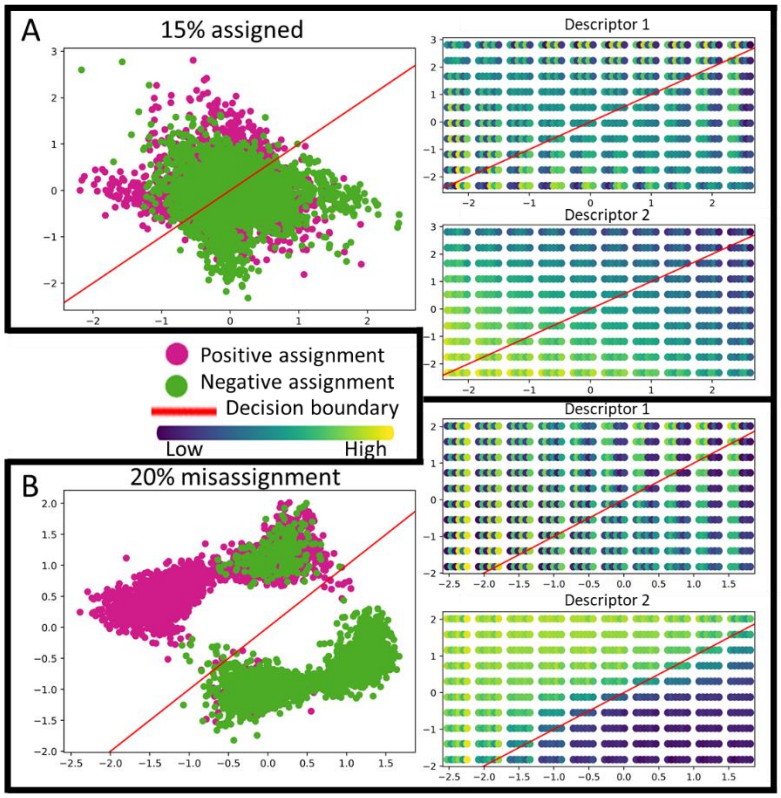


Figure S6. Shown here is our first attempts to validate this method. For these two models there are 10,000 entries in the database. We show that the positive and negative motifs are recovered despite large amounts of noise. The model A has only 15% of the data base with a positive or negative class assignment. All other points in the data base are randomly asigned a class. Here the positive motif is ULULUL in descriptor 1 and LLLLLL in descriptor 2, and LULULU in descriptor 1 and UUUUUU in descriptor 2. The negitive motif is UUULLL in descriptor 1 and LLLLLL in descriptor 2, and LLLUUU in descriptor 1 and UUUUUU in descriptor 2. This model shows that not only can the model recover the positive and negtitive motifs, it groups the related positive and negitive motif on the same side of the plot, according to descriptor 2. Model B is the same as Fig. 4 but with 10,000 data points.

h. Primer of Variational Autoencoders

The Variational Autoencoder (VAE) has been published elsewhere [2], but here we offer a simplified explination. The VAE attempts to efficiently compress data to a lower dimensionally using the encoder half of the model (which maps from the inputs to the node labeled μ) and then reconstruct the original data from the lower dimensional representation using the decoder half of the model (which starts at the node labeled $N(\mu, \sigma^2)$ and outputs when the model reaches the reconstructed input). To make this compression more robust, random noise is injected into the reduced representation (commonly called the latent space, or latent representation) during training. For a one-dimensional latent space, the encoder side of the VAE will reduce the data to two nodes labeled μ and σ^2. These values will be used to define a Gaussian distribution where $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. During training the node labeled $N(\mu, \sigma^2)$ will output values by sampling from this probability distribution. When training over the same input many times the average of the values given by node $N(\mu, \sigma^2)$ will be μ and the variance will be σ^2. When trying to understand variational auto encoders it is important to remember that traditional ANNs are deterministic. This means that if you put the same input into an ANN you will always get the same output. VAEs are unique in that during training the node labeled $N(\mu, \sigma^2)$ is not deterministic, but a single input will form a distribution over many rounds of training. In this way, the encoder finds an appropriate point in latent space to encode to and finds the amount of noise it can add before no longer being able to reconstruct the inputs effectively. This consideration of tolerance for noise is the source of the VAE's eventual

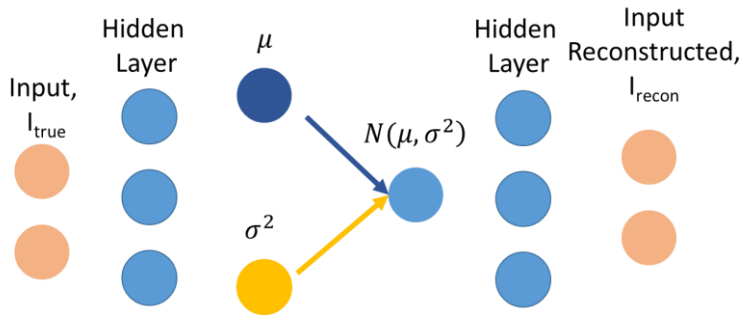


Figure S7 A representation of a variational auto encoder (VAE) with two inputs and a one-dimensional latent space. All layers can be assumed to be densely connected except layers where arrows are explicitly drawn.

resilience to variations in the input data. The tolerance also allows effective decoding of arbitrary points in latent space. In addition, the output of the autoencoder smoothly transitions between points in latent space. This is convenient because the model will group points with similar inputs in latent space and will allow us to interpolate what input would reach an arbitrary point in latent space. In the variational autoencoder, the loss function balances two objectives. 1. It minimizes the difference between the true input and the reconstructed input, and 2. it seeks to maximize the amount of noise tolerated by this reconstruction. This reconstruction capability was key in our method to understand why classification was made by an ANN.

i. References

Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta BBA - Protein Struct.* **1975**, *405* (2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

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

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.

Do, T. D.; Kincannon, W. M.; Bowers, M. T. Phenylalanine Oligomers and Fibrils: The Mechanism of Assembly and the Importance of Tetramers and Counterions. *J. Am. Chem. Soc.* **2015**, *137* (32), 10080–10083. https://doi.org/10.1021/jacs.5b05482.