UNIVERSITY OF CALIFORNIA

Santa Barbara

Computational Model Development and Validation on Systems of Biochemical Polymers

A dissertation submitted in partial satisfactions of the
requirements for the degree Doctor of Philosophy
in Chemistry

by

Nathaniel Morgan Charest

Committee in charge:
Professor Joan-Emma Shea, Chair
Professor Horia Metiu
Professor Irene Chen
Professor Michael Bowers

September 2020

The dissertation of Nathaniel Morgan Charest is approved.

_____

Horia Metiu

_____

Irene Chen

_____

Michael T. Bowers

_____

Joan-Emma Shea, Committee Chair

August 2020

ACKNOWLEDGEMENTS

I would like to thank my advisor Joan and all my educators who helped set me on this path and thank my wonderful partner Evan for supporting me along the way. Thanks to my friends and family and to all who inspired me to continue learning all that I could. And thank you to Melissa Woods, whose passion for physics taught me in ninth grade that theoretical science was one of the most fascinating things our world had to offer.

VITA OF NATHANIEL MORGAN CHAREST

September 2020

EDUCATIONS

Bachelor of Science in Biochemistry, University of California, Santa Barbara, June 2014

Doctor of Philosophy in Chemistry, University of California, Santa Barbara, September 2020

PUBLICATIONS

Song, B; Charest, N; Morriss-Andrews H.A; Molinero V; Shea J-E. Systematic derivation of implicit solvent models for the study of polymer collapse. *J. Comp. Chem.* 2017. 28. 1353-1361.

Tro, M; Charest, N; Taitz, Z; Shea, J-E; Bowers, M. The Classifying Autoencoder: Gaining Insight into Amyloid Assembly of Peptides and Proteins. *J. Phys. Chem. B.* 2019. 123(25). 5256-5264.

Charest, N; Tro, M; Bowers, M; Shea J-E. Latent Models of Molecular Dynamics Data. *J. Phys. Chem. B.* Just accepted. DOI: 10.1021/acs.jpcb.0c05763

ABSTRACT

Computational Model Development and Validation for Systems of Biochemical Polymers

by

Nathaniel Morgan Charest

Modeling of biochemical polymers using classical simulation and the analysis of large datasets has become important in an age when high-throughput experimentation and advanced computational resources have enabled the collection of massive bodies of information regarding these systems. Models help us understand and extract insight from experimental data, allowing us to develop simulations and predictions around their behavior that can help elucidate macroscopic behaviors of interest. The body of work applies techniques in molecular dynamics, machine intelligence and information theory to problems within the realm of biological heteropolymers with the intent of developing understanding regarding modern methods of data generation and analysis. The differing paradigms of first-principles models versus data first models are considered and examined, and work is done to show areas where the methods are complementary or applicable to helping understand experimental data sets.

Table of Contents

Chapter 1) Introduction

1.0 Forward

In 1976, the statistician George Box famously remarked in his paper "Science and Statistics"[1] that 'all models are wrong'. The aphorism was completed in a subsequent work of his, suggesting 'all models are wrong, but some are useful'. This, arguably, is one of the most concise descriptions of modeling's relationship to science, in which one strives to accurately describe the living reality.

Within the field of biochemistry, the complexity and diversity of systems invites conscious scrutiny of Box's notion. With the advances in the calculation of novel models and enhanced numerical ability, we must remain cognizant that we are will never, truly, be able to exactly model the reality, and knowing where our models can succeed and where they can be wrong is necessary as we pursue better representations of the 'true model' of the universe. Despite this, models remain one of our most powerful tools in understanding the natural universe, and so application of new modeling techniques to biochemical problems remains the prerogative of investigators in the field.

By construction, science is the process of elucidating the natural world and its attached phenomena as accurately and as precisely as possible. Many tools for this exist – experimental designs, mathematical language, peer review – but within this toolkit one of the most recurring approaches on the theoretical end is modeling. In biochemistry, modeling is the core means of describing often impossibly complicated systems in a way that allows prediction or deeper internalization of the material. These models can be derived from simple principles of physics – as in the case of molecular dynamics – or they can be constructed based on their ability to predict experimentally verifiable points of data – as in the case of neural network predictions of chemical activities. In either case, most of these models are 'wrong' in the sense that they can accurately

predict within a specific range of capability, but beyond these limits are unwise to casually interpret or apply.

Particularly in the Information Age, computation has reached the level where scientists have a seemingly limitless number of modern and exciting techniques to interrogate the secrets of the universe and build their models. Previously inaccessible problems can be answered through clever construction of models, who designs can grant insight into vast databases or highly complex systems. Despite how exciting these potentials are, work is still needed to assess the basic applicability of many of the approximations or assumptions that often are involved in achieving these intentions.

This work concerns itself with the application of models to modern biochemical problems, with the goal of scrutinizing where these applications yield helpful and actionable insights. It concerns itself with two primary domains – molecular dynamics modeling and information theoretic, colloquially machine learning, analysis techniques – that are unified by the shared subject matter of biochemical polymers. In the following four investigations, we consider the problem statement and its contemporary means of approach. We then apply a modern method of modeling the data, followed by a dissection of what it is capable of doing and final words of caution regarding where it might be 'wrong'.

This chapter introduces some of the key concepts used through the remaining work. In the next section, we consider the general system of 'biochemical polymer' and define it both mathematically and as it contextually relates to the greater field of biochemistry. We next review molecular dynamics, with a particular emphasis placed on coarse-graining or imbedding models, as it speaks directly to the heart of the useful assumptions one can make when attempting to apply classical models to gain insight into chemical systems. These assumptions are often essential and

attempting to compute predictions from wholly first principles is impossible with modern technology. Finally, we review some fundamentals of information theory and their associated application to artificial neural networks. These understandings inform the approaches that are applied and investigated in the following chapters.

1.1 Biopolymers: The Machinery of Life

In the most general sense, a biopolymer is simply a polymer – a molecular structure defined by the repetition of discrete 'monomers' that can be extended to arbitrary length – that has its synthetic origins within the chemical processes of a living organism.

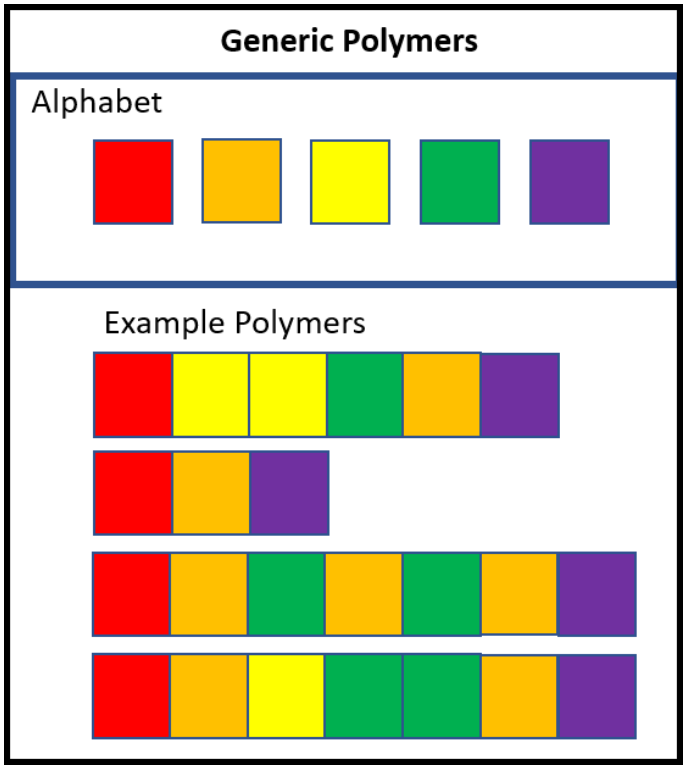

Figure 1.1: A general schematic of polymers. The red element is the beginning element, and the purple element is the end element. Orange, yellow and green are monomers that can occur in any order with any frequency in a polymer sequence, defining a set of potential polymer sequences with a variety of chemical and biological functions and activities.

Given modeling's deep connections to mathematics, we formally define a polymer as a finite mathematical sequence in which the elements are drawn from an 'alphabet' set whose elements are called monomers, with one possible exception (the terminal element). The alphabet can be of arbitrary size however it must contain at least one 'terminal' element with the property they can only occur at the beginning of a polymer sequence or at the end of the sequence. In the key systems studied with this dissertation, a given class of biopolymer has only one distinct beginning terminal and one distinct end terminal element. These terminal elements can define a directionality – in polypeptides these are called the N-terminal

element and the C-terminal element, while in polynucleotides they are called the 5'-terminal and the 3'-terminal. There is no requirement that a polymer use every element of its associated alphabet aside from the presence of the terminal elements respectively at the beginning of the sequence and the end of the sequence.

Some example general polymers using the same alphabet is shown in Figure 1.1. A polymer becomes a 'biopolymer' simply when it is a polymer whose alphabet or whose synthesis has biological origins. This mathematical definition will be helpful, as it speaks to an isomorphism between polypeptides and polynucleotides that allows many of the techniques used in this dissertation to be applied to either class.

We note that the space of associated potential polymers is countably infinite for a given alphabet based on this construction. This is the origin of biopolymer's incredible versatility in the biological context: with even a relatively small number of synthetic processes generating a modest alphabet, a countably infinite number of potential structures with varying capabilities becomes available to an organism. This ratio of potential activities to synthetic requirements is incredibly valuable and represents the core value of biopolymers to the evolution of efficient, viable lifeforms.

Due to their involvement in numerous distinct biological phenomena, modern scientific descriptions of life universally invoke these polymers. Their variability and complexity lend themselves to vast arrays of functionalities critical to sustaining an organism, all while maintain a fundamental simplicity that is extremely amicable to the limited resources of biosynthesis.

Broadly, there are three main classes of biopolymers. Polysaccharides are polymer systems made up of linked sugar monomers, and carbohydrate science has deep relevance to metabolism, organism energy storage, and the properties of biological structures. Polysaccharides are unique

amongst the three in that their terminal elements are virtual – they have no chemical relevance or manifestation and act only to denote the linear topology of a given sequence or else where the chain can connect with extraneous chemical structures. More generally, the terminal elements define a topology – the fill the mathematical role of specifying that, physically, a polymer is linear, and disruption of that topology can only occur at the terminal elements without compromising the identity of the polymer as whole.

This work concerns itself primarily with the other two classes of biopolymers, which rigorously fulfill the requirements laid out above. Three systems of study in this dissertation involve models of polypeptides – peptides and proteins – while the fourth develops a model of polynucleotides. In the next section, we review polypeptides and polynucleotides, and discuss their shared characteristics and relevance.

Polypeptides, Polynucleotides and Biopolymer Structure

The biological paradigm of biopolymers considers their structural and chemical role, biosynthetic and metabolic importance, and general involvement in the molecular evolution and execution of living processes. Due to their sheer importance in the fundamental machineries of biology, a full exposition of their background would be prohibitively vast and would require describing many decades worth of investigations into biochemistry[2]. This work focuses on modeling of their structural behavior and their activity and makes use of their shared properties in this regard. This section will discuss the structural commonalities and terminologies shared by polypeptides and polynucleotides, with a more holistic portrait their role in biochemistry left to existing literature.

Because of their shared fundamental structure as polymers – a chain of connected monomers drawn from an 'alphabet' of candidates with differing chemical properties – much of the lexicon in describing polynucleotides and polypeptides is shared. Both types can be considered structurally at similar levels: the primary structure, the secondary structure, and the tertiary and quaternary structure. It is the variabilities of these structures that constitutes the principle value of these polymers to life. Despite the space of potential monomers beings relatively quite small – only 20 proteinogenic amino acids are used in human proteins, and only 4 nucleotides constitute the 'alphabet' of human ribonucleic acid polymers – the fact they can be linked in any order and to arbitrary length confers a truly astounding amount of flexibility in function because they can adopt any number of structures and forms in the living environment. These first principles allude to a deeply useful theoretical fact: in abstract, there are significant isomorphisms between the informational structures that capture useful relationships and properties of both polynucleotides and polypeptides. This is represented in the similar means of characterizing their structures.

Primary structure refers to the exact order of monomer types in the biopolymer chain. Every biopolymer is associated with an alphabet of potential monomers, and the primary sequence captures this order.

This primary sequence interacts with itself and the environment to produce the high-level structures. Secondary structure concerns itself with the general three-dimensional form of local segments in a biopolymer, with these structures typically dominated by the hydrogen bonding patterns. Secondary structure plays an enormous role in the biochemical function of a given polymer. In functional RNA, the secondary structure typically has deep implications for its ability to catalyze or react with potential reagents. In protein chemistry, the secondary structure can

inform everything from thermodynamic stability to the availability of chemically active sites for any number of potential interactions. Characterizing secondary structure is often the initial goal of researching a new biopolymer, with many theoretical models and associated calculations done to predict the secondary propensities.

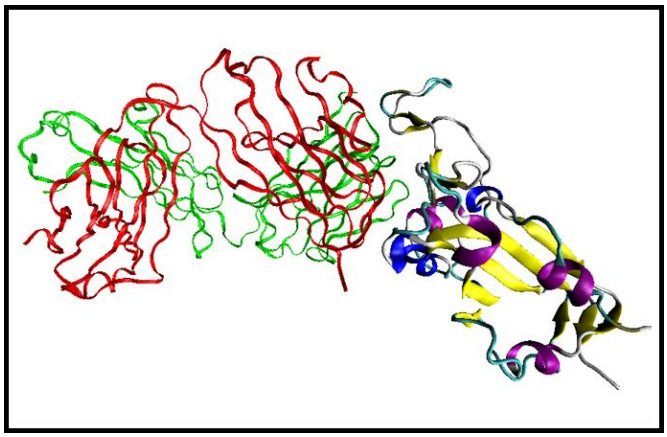

Figure 1.2 A visualized structure of COVID-19 virus spike receptor-binding domain complexed with an antibody, as a depiction of polypeptide structure. The image illustrated secondary structure in the rightmost moiety, where secondary structures are independently visualized as ribbons and coils with differing color. Tertiary structure is visualized by the three-dimension form of the sequences, while quaternary structure captures the arrangement of the separate polymer chains.

The tertiary structure considers the absolute three-dimensional structure of the polymer in its environment. The 'fold' of a protein or nucleic acid plays an essential role in its activity by enabling or disabling access of its chemically active sites to loci of metabolic importance in the living milieu. Tertiary structure is often thought of in terms of a native fold – the dominant and chemically active structure – and its misfolds. Despite this paradigm, there is evidence that flexibility in the tertiary structure facilitates the versatility of polymers in the machinery of life, as in the case of intrinsically disordered peptides and promiscuous ribozymes. The combination of interest in these flexible systems and the difficulty of capturing their behavior using classical models has drawn a considerable need for clever techniques to simulate or elucidate their behaviors.

Finally, quaternary structure concerns itself with the arrangement of multiple units of biopolymers that are not chemically bonded. This terminology is more rare with respect to the polynucleotide systems, however it does see use when considering high-level organization of

DNA[3,4]. A major system of study in two of the ensuing works is the amyloid fibril. These quaternary arrangements of intrinsically disordered peptides are defined by their preference for a secondary beta-sheet structure and are clinically associated with numerous diseases.

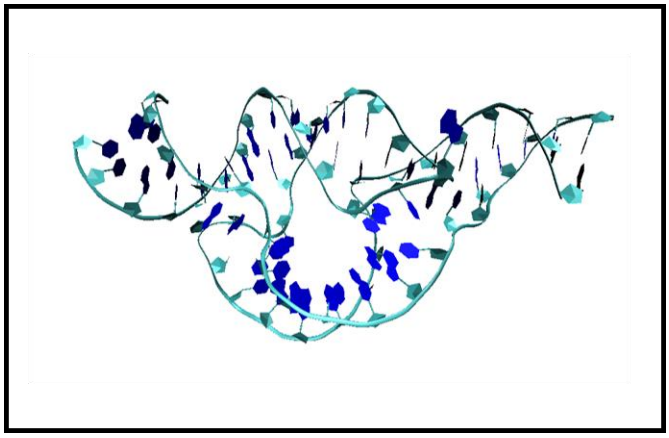

Figure 1.3 A visualized structure of a mutant hammerhead ribozyme, demonstrating the structure of functional polynucleotide polymers. The hammerhead ribozyme exhibits strong secondary and tertiary structure characteristics.

Figure 1.2 provides a visualization of a COVID-19 viral protein in complex with an antibody[5], elegantly illustrating secondary, tertiary and quaternary structure for a set of three polypeptide polymers. Image was generated with VMD[6] and the structure file was downloaded from the Protein Data Base identification code 7BZ5. Figure 1.3 shows a structure for a mutant hammerhead functional RNA polymer[7], protein database file 3ZD4. This image was also generated by VMD.

Because the structure of these polymers is so complex, much of their investigation involves determining the relationship between the primary structure – conceptually the simplest as it is restricted to L^N discrete forms, where L is the length of the monomer alphabet and N is the length in monomers of the chain – and the secondary and tertiary structures. The primary structure is immutable – it is the defining structure for a given polymer and cannot be changes without changing the fundamental identity of the chemical species. Conversely, while high level structures impact chemical activity, they are not immutably entangled with the chemical identity of the polymer. Thus, it is convenient to understand the relationship between chemical activities and behaviors of interest in terms of the primary sequence, and this necessarily requires models,

analysis techniques and theories that can relate the primary sequence to secondary and tertiary traits.

1.2 The Importance of Models

It warrants consideration of why one might be interested in developing models of these biopolymer systems, as the process of abstracting, developing rules and validating assumptions about physical structures can be nearly as resource intensive as experimentally investigating biopolymer properties to begin with.

An immediate answer is that the process of modeling enables these structures to be studied as a whole, exploiting the abstract similarities noted in the prior section to draw conclusions about the biopolymer class without independently investigating each system on a case-by-case basis. These abstractions are especially important in the Information Age, when the paradigm of object-oriented programming requires investigators to have well-articulated classes of objects and their interactions in order to be translated into languages that best enable modern computational tools to further our investigation efforts. Developing and investigating these models and abstractions enables collaborating programmers to fully utilize key principles of good software design[8] without necessarily requiring detailed expertise in chemistry or biophysics. This is an extremely valuable end for the purposes of effective cooperation between software engineers and physical scientists.

Models also possess certain advantages depending on the type of model. As discussed in the next section, certain types of models can allow for conclusions to be drawn from first-principle simulations by exploiting high-level similarities in the descriptions of vast spans of unique biopolymer species. Alternatively, while high-throughput methods for experimentally

interrogating enormous libraries of specific species are being discovered, the ability for human researchers to synthesize those collections into useful, actionable insights is significantly limited.

Models are a key step in the development of better artificial intelligence and digital research platforms, as they are essential to the encapsulation of necessary description and theory for a machine intelligence to access the material in a useful way. Much in the way a model is a helpful pedagogical tool for teaching a new researcher concepts and structures within a field, model development and exploration is essential for a successful transition of research obligations onto digital intelligences.

Thus, modeling represents an extremely important step in the futurization of scientific progress, acting as a critical bridge between artificial intelligence and human understanding of biopolymer theory.

1.3 Classical Modeling: Molecular Dynamics & Simulation

The foundational question of elucidating the relationship between primary sequence and high-level structure for biopolymer systems is: ‘given a primary sequence, can we methodically predict the ensemble of structures it will occupy for a given environment?’ The immediate resolution to this question, at least in equilibrium, can be answered using the rigorous machineries of statistical mechanics. Equilibrium, however, neglects the importance of kinetic factors and the role of non-equilibrium in facilitating biochemical processes, and the prediction of the Hamiltonian energy descriptions essential to statistical mechanics theory is nontrivial given the limitations of modern computational ability and the complexity of the quantum calculations that would yield accurate first-principles energy descriptions. Because of this, simply parameterizing and minimizing

structural energy is limited in the degree of insight it can provide or else requires the use of assumption to limit the scope of the calculation to a tractable degree.

Molecular dynamics (MD) strives to address this by explicitly parameterizing trajectories of a given system through time, through the iterative integration of the classical equations of motion. Enormous numbers of algorithms and methods have evolved to suit the needs of various questions, involving the control of environmental factors, the need to sample rare events, and the generation of accurate probability measures for ensembles of structures. MD is an alluring modeling approach to calculating the relationship between primary and high-order structure because it tends to be extremely intuitive. At its root, it is subjecting molecular structures to the same forces and dynamics that characterize the classical world we experience. It offers detailed descriptions of dynamic processes frequently inaccessible to direct or indirect empirical observation, and it can eliminate the need for expensive or dangerous experimental setups.

Despite these advantages, it is not without drawbacks. MD simulation is an inherently discrete technique applied to a fundamentally continuous phenomena: time evolution. Consequently, improper choice of temporal discretization can create erroneous results or numerical instability. It is also limited by a need to explicitly compute interactions. In a holistic MD setup, every component that can potentially affect the analyte system must be represented as a point particle, with its interactions explicitly parameterized.

This ‘all-atomistic’ approach to molecular dynamics can cause systems with high scale or complexity to be excessively expensive. Solvent interactions, for example, are known participants in the structural preferences of many biopolymers – each solvent molecule, in these cases, requires explicit representation and computational treatment to properly capture system behaviors. This can

create situations where effective sampling of even relatively small systems will either take extremely specialized hardware, or else require unreasonable computational times.

Putting aside the computational sensitivity to scale, an explicitly represented molecular system also requires the parameterization of its numerous interactions. Casually, for a system of N distinct atoms, parameterization requires (N choose 2) energy potentials of interaction. This is a forgiving approach – in truth the chemical environment of a given atom will impact its interactions with others, and thus simply parameterizing based purely on atomic identity will likely result in poor predictions. This process of parameterization is broadly referred to as the development of a force field.

In practice, well-designed force fields are the fruit of years of effort and dissertation, and they are always associated with success on some subset of systems and failure in another. A truly general force field – one capturing the 'true model' of molecular interactions – remains beyond the reach of current investigatory efforts. Cautious work is required to properly choose a force field and, if a system of interest has not been previously parameterized and validated for appropriately similar situations, the results of any simulation using a presumed forcefield is inherently in question.

By embracing the notion that 'all models are wrong, but some are useful', we can make an argument for simplifying these representations. We will know more exactly how our models are wrong by construction but take advantage of the simplifications to preserve the model utility. All-atomistic techniques come with the inherent risk that, by doing one's best to reflect the 'true model' of reality, one necessarily lets go of being able to make calculated reductions in detail to better facilitate the overall research goal. Mindful use of this exchange between resolution and

tractability transitions the conversation to coarse-grained models within MD, which address weaknesses of all-atomistic approaches at the cost of structure resolution.

Coarse-Grained Molecular Dynamics Models

Coarse graining attempts to deal with the sensitivity of molecular dynamics methods to scale by attempting to answer a simple question: how can a model preserve its ability to provide useful insight while reducing the absolute number of particles that must be considered at each time step? There a few resolutions to this.

Systematic coarse graining[9] attempts to treat groups of atoms as single particles by developing a repeatable framework that can be applied to all-atomistic structures, reducing the number of particles that need explicit consideration. This can involve the development of rigorous rules for combining the interactions of grouped particles – this is a common approach for modeling polymers[10,11] – or the parameterization of properties based on recreating experimental observables (the radial distribution function, for example, or the force distribution). One such model used in Chapter 2 is the Molinero water model[12], which was parameterized to reproduce thermodynamic and geometric properties of water, and will be discussed in more detail in the relevant chapter.

A final approach is helpful when assessing general consequences of essential principles[13–16], and is particularly robust to when there is an issue of scale. So-called 'phenomenological' models take a top-down approach to capturing behaviors of interest and will use low-resolution structural representations with relatively simple interactions. The reduction of the number of parameters can lack enough chemical detail to relate to any particular system, however in cases where similar behaviors are attributable to numerous disparate chemical species (e.g. the aggregation behavior of peptides[17] or the hydrophobic packing behavior of protein chains[18]) this reduction in specificity

can be an advantage. These coarse grained models allow the appreciation of the broad implications of low-resolution chemical similarities, and by tuning the relatively small parameter space researchers can use these models to study the process' relationship to the shared characteristics of species exhibiting the behavior of interest.

Chapter 3 involves the analysis of the Shea peptide model for amyloid formation[19] using a latent model of the molecular dynamics ensemble. Because the model is coarse-grained, two major mechanisms of amyloid formation can be easily explored by tuning one of the model parameters, easily providing diverse testing cases for using neural network methodology to characterize the system evolution and structural ensemble.

1.4 From The Data Modeling: Information & Fitting Approach

The other side of modeling in this dissertation that is so far undiscussed is to approach from top down rather than classical principles. Molecular dynamics represents a paradigm in which a well-validated fundamental description of the universe – the laws of classical motion – are integrated to explore the energetic landscapes of a biopolymer's conformations. It attempts to parameterize the systems of interest at an elemental level, and explicitly generate insight into the systems by iterative application of what essential physics tells us will happen at each time step. This is opposed to the other paradigm of modeling, in which one uses vast quantities of data and clever processing tools to find embedded information.

This paradigm has been facilitated in the modern decades by enormous advances in computational technology and theory, which have allowed the development of so-called 'machine learning' techniques for maximizing predictive model performance based on vast sets of data. The argument for these techniques is that because the information of interest is embedded in the

datasets, there is no reason to attempt to build a model that can simulate the informational trajectory. This approach thus builds its models by taking flexible mathematical structures and fitting them to the existing data.

These vast datasets have become available to those studying biopolymers relatively recently, with the advent of high-throughput techniques enabling the measurement of huge bodies of experimental data on relatively quick timescales[20]. Because of these techniques, machine learning and information models do not stand to lose relevance to the field any time soon and can be expected to continue rising in dominance. Chapter 5 applies one of the most basic principles of information theory to extract detailed insight into the structure-activity relationship of a functional RNA sequence from its related fitness landscape.

In the chapters concerning the use of information approach models, it will be important to understand the critical point of difference in the approaches. As previously stated, a first principle model tries to reproduce the true model of reality, as in the case of capturing the solvent-driven behaviors of polymers in water covered by Chapter 2. This is subtly different than the approach of the information-driven model, in which an amorphous mathematical structure is 'trained' to better reflect the information content of the data such that a researcher can easily isolate and analyze the relevant information as it pertains to the research question – this approach is exemplified in Chapter 4's analysis of a database of peptides labeled as forming amyloids or not. The articulation of whether a model generates insight or extracts insight speaks to the compatibility of the two paradigms – if a suitable means of generating data from classical simulation is found then its analysis can often be facilitated by the informational extraction tools. This is the purview of Chapter 3, in which molecular dynamics data is analyzed using the information tool of variational

autoencoders, resulting in a means of analyzing the ensembles of data associated with amyloid forming processes.

Artificial Neural Networks

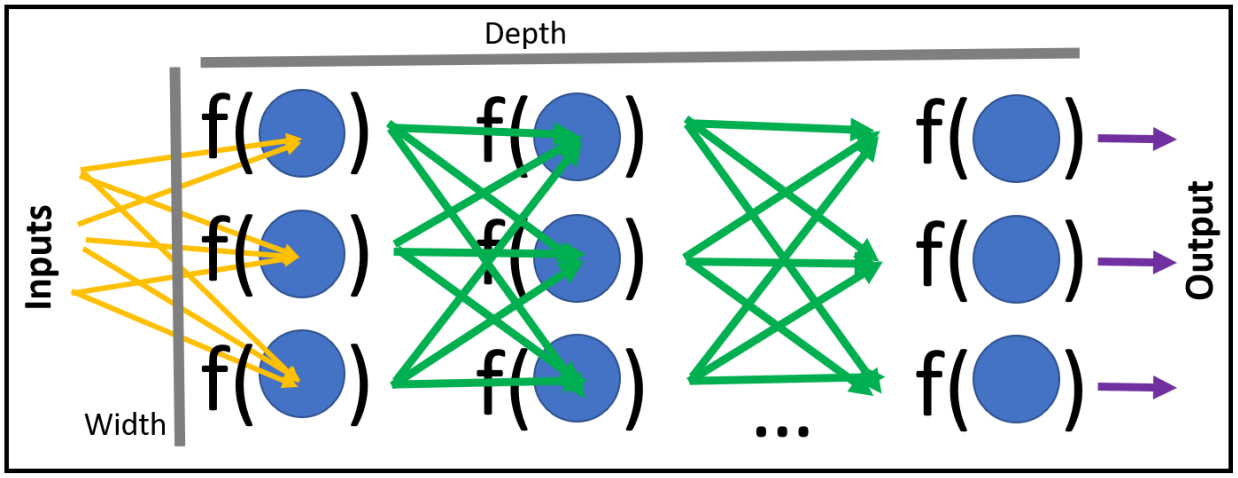

Figure 1.4: A schematic of a neural network, showing the organization of fundamental units, neurons, into layers. A neuron is associated with a weight vector, the blue dot, and an activation function, f(), which manipulates the values flowing through them to approximate some mapping. This particular connectivity and directionality is known as a feed-forward, densely connected neural network. Other topologies are associates with differing performance and ability, and this is an active area of research within the field.

The family of methods known as Artificial Neural Networks is a monolithic subfield of machine learning techniques, the scope of which is well-beyond this dissertation. However, they are used in two different ways in Chapters 3 and 4, and so are concisely introduced here.

A schematic of a neural network is shown in Figure 1.4. The artificial neural networks used in Chapter 3 & 4 are always feed-forward, representing the interest in testing the fundamental properties associated with the theoretically simple structures as opposed to the more sophisticated variants that have been engineered for their performance over the past few decades. A feed-forward neural network consists of at least two ‘layers’ of neurons, whose values are defined by an activation function and an associated weight. The width of the layer is the number of neurons per

layer, while the depth of a given network is its number of layers. While the term is so ubiquitously used as to be lacking in firm meaning, it is generally thought that a given neural network is considered to be a 'deep-learning' tool if it possesses at least three layers, where the central layers – those unassociated with input or output – are the so-called 'hidden layers'. In a feed-forward scheme, information flows one-directionally through the network, beginning at the input layer as input values, being transformed by the hidden layers, and eventually exiting as the output of the final layer.

It can be shown through the universal approximation theorem[22] that a feed-forward neural networks with an arbitrarily bounded, continuous and nonconstant activation function is a universal approximator. Simply put: a feedforward neural network satisfying certain conditions can represent any mapping between the input and output space. This mathematical versatility is the core of the ANN's prevalence in machine learning. Because it can learn any mapping with sufficient training data and the backpropagation algorithm, it can theoretically play the role of any nonlinear mathematical relationship a model requires to accurately capture a trend in data.

While this capability is utterly impressive, ANNs are possessed of several limitations. The universal approximation proof does not provide guidance on the number of neurons or their topology to approximate a given function, requiring cautious tuning of so-called 'hyperparameters' to produce accurate models. Despite these limitations, ANNs have been instrumental in multiple advents of machine learning, especially in the regime of machine optics, process control, the approximation of quantum mechanical interaction potentials, and numerous other fields. It is truly difficult to overstate the role ANNs have played in advancing machine intelligence. This work adopts the paradigm that one of the core abilities of ANNs' ability to capture any mapping is the flexibility it provides in changing representations of information into more accessible forms. In

Chapter 3, this property is used to automatically learn a means to change the representation of serial molecular dynamics trajectories from a highly dimensional enumeration of internal coordinates into a single, simple value that tracks the transition of the system from disordered monomers into an ordered solid. In Chapter 4, an artificial neural network is used to develop a low-dimensional representation of the ‘amyloidogenicity’ of a given peptide, allowing for the visualization of motifs of physical properties along the peptide polymer that relate to a propensity to form amyloids.

The exact implementations of the artificial neural networks will be delved into within the relevant chapters. Here it is emphasized that, despite the caution necessarily in training and usage, ANNs are one of the most flexible and powerful tools for manipulating the representation of data that have come out of recent advances in machine learning.

We now dive into the next chapters, in which more detailed work regarding the concepts covered in this section are delved into.

Chapter 2) Validation of Solvent Models in Coarse Grained Simulation

2.0 Forward

The following work was published in the Feb. 25 2017 issue of *Journal of Computational Chemistry* under the title "Systematic derivation of implicit solvent models for the study of polymer collapse"[23] and is reprinted with minor changes here with the permission of the publisher.

In the context of this dissertation, this work acts a convenient starting point for our investigation of modeling techniques and their applicability to the study of biochemical polymers. It embodies the assumptions endemic to coarse-grained and implicit molecular dynamics techniques, and rigorously investigates the use of a method of representing solvent that mitigates the need for the computationally prohibitive representation of the system particles. This speaks to the heart of a significant consideration regarding the first-principle models of molecular dynamics. Because they require the repeated application of fundamental physical laws – simulation – there is a demand for balance between the computational tractability of the model and adherence to the true model of the physical system. This consideration is inescapable in the world of simulation, with analogues in any field the requires the representation of many interacting objects.

The desolvation potential is a phenomenological model of solvent interaction that embeds a specific pairwise behavior of hydrophobic particles when solvated into the pairwise Hamiltonian. Embedding the interaction in this way prevents a need for explicit representation of the solvent molecules, enhancing the computational speed of simulation by orders of magnitude.

Despite this incredible advantage, the assumption that solvent activities can be captured

through pairwise additivity is not without substantial risk. Because this risk-reward consideration is a core principle in the application of models to biochemical polymers, this work serves as an analysis devoted entirely to first-principles modeling of polymer behavior.

2.1 Introduction

A characterization of the thermodynamics of folding of short (100 amino acid) biomolecules in an explicit solvent representation is feasible using enhanced sampling methods such as replica exchange molecular dynamics simulations or metadynamics. However, the problem becomes rapidly intractable when one seeks to study the assembly of multiple proteins, as would be the case in protein aggregation. Much of the cost in simulating such systems comes from evaluating the interactions and forces acting upon the large number of water molecules needed to solvate the system. Coarse-graining the solvent, or using an implicit representation of the solvent becomes a necessity given current day computational resources. In this work, we consider polymeric chains and evaluate the effect of systematically simplifying the solvent representation on the conformational free energy landscape of the chain. We start with an explicit coarse-grained water model, then develop an implicit solvent representation that reproduces features of the free energy of attraction between monomers in solvent and finally consider the polymer in a mere vacuum environment.

Several coarse-grained explicit water models have been developed in recent years, including the MS-CG model of Voth and co-workers[24], the relative-entropy based model of Shell and co-workers[25,26], then 3D-Mercedes-Benz model[27], the BMW model[28], to name a few. In this work we use as starting point for the development of implicit models the coarse-grained explicit water model mW[12], a single site model with no electrostatic interactions and hydrogen bonds, but which nonetheless captures the tetrahedral arrangements of water molecules, and reproduces the radial

distribution, enthalpy of vaporization, and surface tension, as well as the anomalies and phase transitions of water. Important for this work, the mW model quantitatively reproduces the free energy of association in water of a pair of M methane beads that make up our polyM polymer. In quantitative agreement with simulations of methane in fully atomistic water, the free energy of association between the pair of methanes displays three features: a contact pair (CP) global free energy minimum corresponding to the methane pairs in contact, a solvent-separated pair (SSP), and a desolvation barrier (DP) separating the CP and the SSP states.[29] The mW model has been extensively validated in studies of hydrophobic interaction of methane pair, hydrophobic solvation of methane, and cavity-ligand binding[12,29,30].

We consider two types of coarse-grained homopolymers in this work. The first one represents an alkane chain made of small hydrophobic beads with explicit bond, angle and dihedral potentials. We will refer to this model as polyC (where C stands for carbon in CH_2 and CH_3). Short alkane chains have been studied in explicit water solvent by Ferguson and co-workers, who found that the chain properties were primarily determined by ideal gas statistics, and that the role of water was quite minor, leading to a destabilization of the unfolded state and to a modest barrier (on the order of the thermal energy) between compact and extended states.[31] The second polymer model involves larger, more hydrophobic beads, similar in spirit to the ones used by Athawalle and co-workers[32], with no dihedral potentials. We will refer to this model as polyM (where M stands for methane). We use mW water simulations of the association of two "M" bead and two "C" beads as our starting point for generating two implicit solvent models, in which the solvent effects are incorporated into the effective interatomic interactions of the solute. The first implicit solvent model reproduces the free energy of contact pair (CP) minima via Lennard-Jones (LJ) interactions between the monomers. We refer to this model as implicit-LJ. The second model reproduces the

free energy of contact pair minima, desolvation barrier, and solvent-separated minima. We refer to this potential as the implicit-DP (desolvation potential) model. Finally, we consider a third model, which we refer to as the vacuum model, in which we simply model the polymer chains in vacuum without modifying the inter-bead potentials. We compare the effectiveness of the two implicit solvent models and solvent-free simulations in reproducing the free energy of folding of the hydrophobic homopolymers in explicit mW water model. The vacuum simulations were seen to outperform the implicit solvent models for homopolymers, leading us to turn to the study of heteropolymers made of hydrophobic M beads and hydrophilic beads. This heteropolymer is a closer model of a peptide, and we examined the effectiveness of solvent-free simulations in reproducing the free energy landscapes obtained with explicit solvation. In contrast to the homopolymer case, the vacuum simulations were less successful, overall leading to overcompaction of the heteropolymer, with the degree of agreement between vacuum and explicit solvation depending critically on hydrophobic-hydrophilic patterning.

2.2 Models and Methods

Molecular Models. We study the folding free energy of two types of linear homopolymers in vacuum, in explicit solvent modeled by the mW water model, and in implicit solvents and vacuum that we detail below. We similarly study the free energy of folding for four sequences of heteropolymers, in explicit solvent and in vacuum.

The mW model uses a combination of short-ranged two-body and three-body interactions to produce tetrahedrally coordinated "hydrogen-bonded" configurations. The functional form and parameters of mW are detailed in the Equation below and in Table 1. Here, r_{ij} is the distance between particles i and j; θ_{ijk} is the angle between atoms i-j-k.

$$E = \sum_{i}\sum_{j>i}\varphi_2\left(r_{ij}\right) + \sum_{i}\sum_{j\neq i}\sum_{k}\varphi_3\left(r_{ij}, r_{ik}, \theta_{ijk}\right)$$

$$\varphi_2\left(r_{ij}\right) = A\varepsilon\left[B\left(\frac{\sigma}{r_{ij}}\right)^4 - 1\right]\exp\left(\frac{\sigma}{r_{ij} - a\sigma}\right)$$

$$\varphi_3\left(r_{ij}, r_{ik}, \theta_{ijk}\right) = \lambda\varepsilon\left[\cos\theta_{ijk} - \cos\theta_0\right]^2 \exp\left(\frac{\gamma\sigma}{r_{ij} - a\sigma}\right)\exp\left(\frac{\gamma\sigma}{r_{ik} - a\sigma}\right)$$

Table 2.1 Fixed Constants for mW Model (Dimensionless)	
A	7.049
B	0.602
γ	1.2
A	1.8
$cos(\theta_0)$	-1/3

The first type of polymer we consider is a united atom model inspired by TraPPE-UA alkanes[33]. The monomer of these alkanes represents a CH_2 or CH_3 group and is referred to as C in what follows. The C-C pairwise interaction has the same Stillinger-Weber functional form and parameters as the two-body term in mW water, except for the value of ε and σ. Non-bonded interactions are excluded for C particles one of two bonds apart. The bond and angle potentials are harmonic, and the dihedral potential is of OPLS style. The C homopolymers (polyC) considered in this study have 10, 20 and 30 C beads and are denoted as C_{10}, C_{20} and C_{30}. The nonbonded parameters for the C beads and their interaction with mW water are in Table 2.2.

The second type of polymer, which we call polyM, is composed of larger, methane-like M subunits. Molecule M was previously used in studies of hydrophobic interactions, the stability, nucleation and growth of clathrate hydrates[29,30,34,34–36]. The nonbonded parameters for the M-M and M-mW interactions are from prior work and are shown in Table 2.2. PolyM has harmonic bonds and angles with no restrictions placed on its dihedrals. Nonbonded exclusions in polyM are the same as in polyC. The four lengths of polyM studied are M_{10}, M_{20}, M_{30} and M_{60}.

Table 2.2 Parameters of the Pairwise Interactions in Explicit Water

	mW-mW (mW-P)	M-M	M-mW/M-P	C-mW	C-C
ε (kcal/mol)	6.089	0.34	0.24	0.10	0.091
σ (Å)	2.4	4.08	4.05	3.166	3.95

Heteropolymers comprised of sequences of hydrophobic M beads and hydrophilic P beads are also considered. Hydrophilic beads (P) are created using the parameters of mW, however their three-body interactions are restricted to configurations in which the central atom is mW (e.g. x-mW-y, where x,y could be mW or P, but not M). This results in relatively small beads with equally strong two-body interactions between both itself and water, while the restriction on the three-body behavior yields a bead that can interact with and integrate into the water tetrahedral network while preventing strong geometries between elements of the polymers. These beads are similar to the S ions from prior studies[37–40], however they differ in that they possess two-body interactions with themselves that are equivalent in strength to their two-body interactions with water.

Pairwise interactions are modeled using the two-body term of the mW model with the same parameters except ε and σ, which are shown in Table 2.1.

The bond and angle potentials are harmonic. It should be noted that LAMMPS implements the harmonic function without the ½ factor. Angles are only constrained for atoms that are connected by one bonded atom. The parameters are shown in Table 2.3.

Table 2.3 Parameters of bond and angle potentials.				
	K_{bond} (kcal mol^{-1} Å^{-2})	r_0 (Å)	K_{angle} (kcal/mol rad^2)	θ_0
polyM	10	2.0	10	109^o
polyC	30	1.54	62.099	114^o
polyMP	15	2.0	10	109^o

The dihedral of polyC is OPLS style, with constants K_1 = 1.141 kcal/mol Å^{-2}, K_2 = -0.271 kcal/mol Å^{-2}, K_3 = -3.145 kcal/mol Å^{-2} and K_4 = 0.0 kcal/mol Å^{-2}. PolyM are not constrained by dihedrals. The bonded atoms, or atoms that interact with angle potentials, dihedral potentials do not have pairwise interactions. It should be noted that the angle, bond and dihedral potentials are identical in the explicit and implicit water simulations.

Parameterization of implicit solvent models. The parameterization of implicit water models is based on the potential of mean force (PMF) of association of monomers in explicit water at 300 K. The computation of PMF of M-M pair was performed in prior work[8], while the computation of PMF of C-C pair was conducted in this work following the same procedures in reference 35[41]. From this PMF we fitted an implicit solvent model with effective monomer-monomer interactions governed by the Lennard-Jones potential that reproduces the primary potential well of the PMF, and a tabulated implicit desolvation potential that reproduces the full extent of the free energy of interaction of two monomers in the explicit water simulations. In the vacuum simulations, the direct interactions of the polymers are the same as in the explicit water simulations. The bonding potentials including bonds, angles and dihedrals of each polymer are the same in all type of simulations.

Simulation of C-C pair association in mW water.

The association of C-C pair is simulated using LAMMPS in a NVT ensemble with temperature set at 300 K and volume corresponding to an average pressure of 1 bar. The simulation box has dimensions 25 × 25 × 25 $Å^3$ and contains 510 water molecules and two C particles. The time step was 5 fs. The temperature is kept with a Nose-Hoover thermostat with damping constant 1 ps.

Free energy calculations.

We used umbrella sampling[42] and weighted histogram analysis method (WHAM)[43] to compute the potential of mean force of the association of C-C pair as a function of the C-C distance. We simulate ten windows with separations of 1 Å between neighboring windows. The spring constant of umbrella sampling is 2 kcal/mol $Å^2$.

Table 2.4 Parameters of LJ Interactions

	M-M	C-C
ε(kcal/mol)	0.34	0.091
σ(Å)	4.08	3.95

Parameterization of Lennard-Jones (LJ) Implicit Solvent Model. As the LJ potential only reproduced the depth and location of the contact pair of the PMF of association, we tuned the values of ε and σ so that the LJ potential shares the depth and global minima of the associated desolvation curve. The parameters used are in Table 2.4.

Parameterization of Desolvation Potential (DP) Implicit Solvent Model. The desolvation PMF between monomers was divided into sections and fitted using polynomial functions. At section boundaries, the fitted functions were joined to create a continuous potential. The associated force was then smoothed by polynomial fitting followed by integration to obtain the new potential. This new potential would differ from the PMF, and in regions that differed polynomial fitting was

again conducted. The polynomial fitting was conducted in cycles of potential-force conversions through integration and differentiation. In the end, we obtained the implicit DP potential and force that were smooth and went to zero at long distances.

Simulation Settings. NVT simulations of the polymers are performed using molecular dynamic simulation package LAMMPS[44]. The time constant of the Nose-Hoover thermostat is 1 ps. Temperature is 300 K for all the simulations. The simulations with implicit water or in vacuum are run with Langevin dynamics at 300 K using a damping coefficient of 1 ps and a a time step of 10 fs. Explicit solvent models use a time step of 10 fs and a density of 0.0327 mW/A^3, which results in an average pressure of 1 bar. Heteropolymer models are run under similar conditions, however with a time step of 5 fs. All simulations are run for 400 ns total, with the first 100 ns being excluded from the WHAM computation of the PMF as the equilibration period.

Free energy calculations. We use umbrella sampling and Weighted Histogram Analysis Method (WHAM) to compute the free energy of homopolymers with respect to its radius of gyration (R_g). The sampling windows are separated by 1 Å in R_g. The restraining spring has a constant 1 kcal/mol. Heteropolymers are studied using replica exchange molecular dynamics. PMFs from these runs are compared to those obtained using the umbrella method, to verify the validity of the alternative methods. Temperatures are chosen such that acceptance ratios for exchanges ranged from 22% to 40%. Initial configurations of polymers are generated by allowing simulations to run at high temperatures for 1 ns and capturing the final structure.

2.3 Results and discussion

The potentials of mean force (PMF) for the association of M-M pair and C-C pair in mW water are shown in Figure 1, alongside the pairwise interactions of M-M pair and C-C pair in both implicit water models, and in vacuum. Consistent with similar simulations and theoretical

calculations[45,46], the PMFs have three common features: one global minimum corresponding to the contact pair, one local minimum for the solvent-separated pair, and a free energy barrier, the desolvation barrier, separating these two minima. There are two notable differences in the PMFs of the M-M and C-C pairs in water. First, the desolvation barrier for the M-M pair peaks at a free energy higher than the free energy of the fully dissociated pair, while for the C-C pair this barrier is essentially at the same level of the fully dissociated pair. Later in the chapter, we demonstrate that this difference impacts the behaviors of two types of polymers modeled with the implicit desolvation potential (DP). The second difference regards the extent to which the direct pairwise interaction contributes to the free energy of association. The association of M-M pair is noticeably influenced by the presence of water: the loci and magnitude of the minima in the free energy profile do not coincide with the one of the minimum in the direct M-M interaction. In contrast, the direct C-C pair interaction possesses a relatively large overlap with the PMF of association, suggesting a lesser role of water in the energetics of attraction.

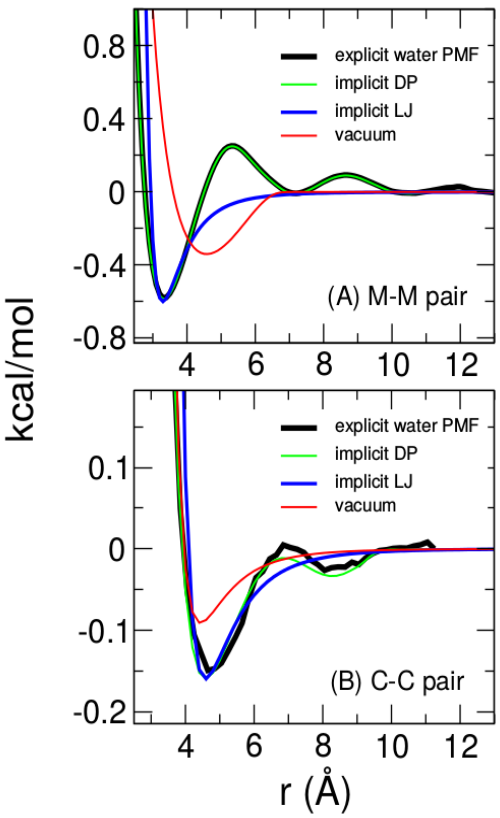

Figure 2.1 Potential of mean force of association for a pair of M (a) or C (b) beads in mW water (black lines), vacuum (red lines), implicit desolvation potential (green lines) and the implicit LJ potential (blue lines). The implicit LJ potential reproduces the first minimum of the free energy of association in mW water, while the implicit desolvation potential also reproduces the second minimum of the free energy of association besides the first minimum.

The implicit model here referred to as the DP model is parameterized such that it captures both the contact minimum and the secondary minimum at the first solvation length. Contrary to this, the implicit Lennard-Jones (LJ) potential only captures the global PMF minimum at the contact pair, omitting the features of the PMF at longer separations. This simplification naturally assumes that these features do not play a significant role in conformations of polymers. We show below these differences are extremely relevant.

Figure 2.2 shows the free energy of polyC and polyM polymers as a function of their radii of gyration. Shorter chains of both polymers display two distinct, stable states in explicit water: a compact, collapsed polymer and an extended state. As polymers grow longer, the extended state becomes less favorable and the compact state dominates, as expected for hydrophobic polymers.

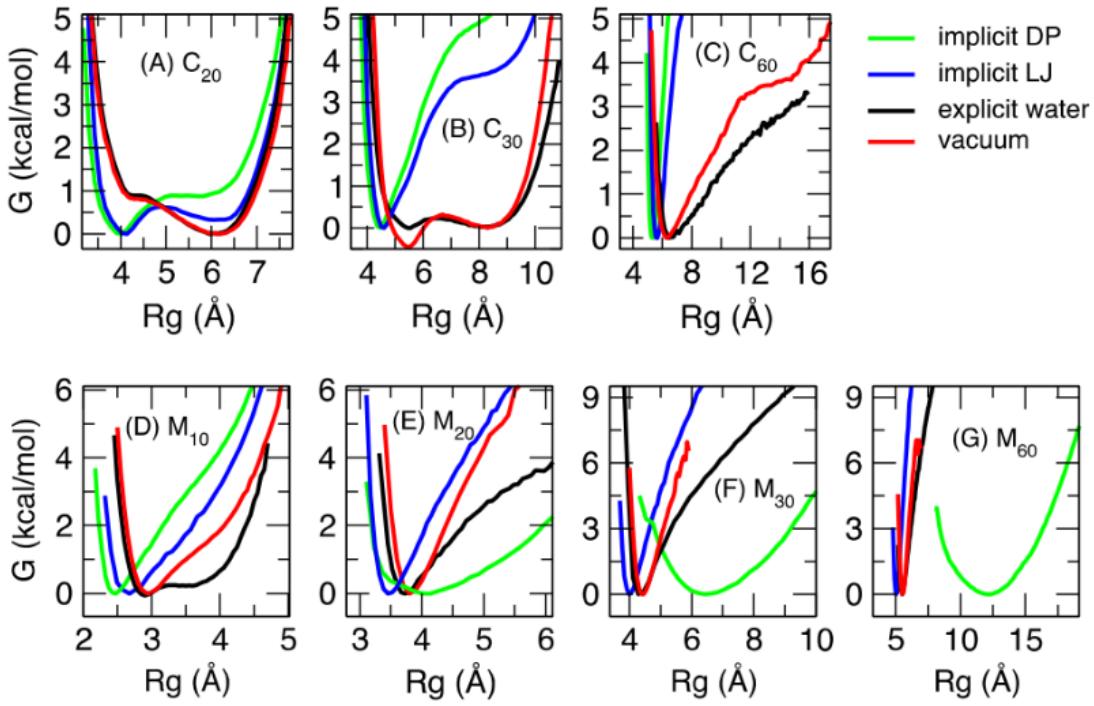

Figure 2.2: *The free energy of polyC (upper panel) and polyM (lower panel) with respect to the radius of gyration (R_g) simulated with the four water models. Following the same color-coding in Fig. 1, black represents explicit water simulations. Desolvation potential is in green, and Lennard-Jones potential is in blue. Red represents vacuum simulations. The lengths of the polymers are labeled in each figure.*

This trend was predicted for polymers with beads similar to the C beads of this work, by both extrapolations of shorter-chain simulations as well as by calculations done using a simple geometric analysis of the configuration space and energetics of solvation[31,47]. These previous works speculated n-alkane chains would begin to favor collapsed conformers at a length of about 50 carbon atoms in the direct interaction scheme; results presented here show this transition occurring at lengths above 30 carbons and below 60 carbons.

Disappointingly, neither the DP nor the LJ implicit solvent models reproduce the free energy profiles obtained in the explicit simulations with mW water. The desolvation potential fails utterly for both polymers across all the lengths, even for the shortest polymers, C_{20} and M_{10}, and is discussed in more detail below. However, consistent with prior work on similar systems, the direct interaction (i.e. vacuum simulations) does a very good job in reproducing the PMF for the various systems, particularly for shorter chains. In the case of C_{20}, the PMFs obtained in vacuum simulations and in explicit mW water very nearly overlap, except for a slightly higher transition barrier between the collapsed and extended states. Work on similar alkane chains has demonstrated the emergence of a barrier for unfolding in explicitly solvated systems with more than 10 monomers. The barrier-raising for C_{20} is smaller in mW water than that observed by Ferguson et al. using SPC/E water as solvent. This could be due to differences between parameterizations of the hydrophobic bead. Despite this difference, the general shape of the PMFs in the mW and SPC/E water models is consistent for explicit solvation of the hydrocarbon chain, describing a two-state system in which the extended conformer is favored for shorter chains and collapsed conformers are favored for longer chains.

In the case of the polyM chains, for which the dihedral potential is absent, the collapsed state is consistently favored across all length scales, regardless of the solvent model used. These results

resemble those of Athawale et al. The PMFs suggest the polyM chain consistently favors a collapsed state under the direct interaction and LJ schemes, underestimating the sampling of the more extended configurations stabilized by explicit solvent.

The simulations in vacuum come the closest in approximating the free energy landscapes produced by explicit solvation, for both polyM and polyC. The shortest chain of the M bead presents one of the biggest disagreements between the direct, vacuum interaction and explicit solvation PMFs. An explanation for this may be that without an embedded preference for trans dihedrals, the chain is overwhelmed by favorable intramolecular interactions of proximal beads. This effect is attenuated by the presence of solvent, which can stabilize the extended states through favorable water-chain interactions. This explanation is strengthened by the observation that, for all M chains, the solvated system has higher populations of extended states compared to the vacuum simulations.

The two implicit water potentials quantitatively differ from the explicit water simulations, as well as from one another. The implicit LJ potential over-collapses both polyC and polyM. This is not unexpected, as the depth of the well of the implicit LJ potential is not only much deeper than that of the direct interaction between the beads in vacuum but also encourages a short distance between particles. Thus, the distance favored by the LJ potential dominates pairwise interactions in the chain, resulting in highly compact structures. The polyC models are able to sample to some extent extended states at lower length scales due to enthalpic preference for trans conformations within chain dihedrals; in the case of longer polyC chains, it seems likely that the entropy of states with gauche defects eventually overwhelms the beneficial energetics of primarily trans conformations, resulting in favored collapsed structures.

As mentioned above, the implicit desolvation potential fails to accurately capture the behavior of the homopolymers in explicit water across all length scales, regardless of the bead used. The implicit DP potential does not capture the R_g of the most stable conformation and, in all but one case, fails to predict a two-state topology of the R_g configuration space. While the Lennard-Jones model has similar failings, the desolvation potential's shortcomings are particularly obvious when considering polyM. As the chain lengthens, the radii of gyration drift to be dramatically greater than in all the other solvent models.

This clearly unphysical behavior of the DP model in the case of polyM is almost certainly due the way the pairwise desolvation potential scales with the number of beads in the system. The use of the desolvation potential assumes that the collapsing hydrophobic polymers entrap solvent particles within them, which are subsequently expelled through some unfavorable transition states. These multimeric water traps formed by polymeric systems such as those studied here are not ably captured by a model that simply tries to account for the presence of a desolvation barrier and solvent separated pair minimum. Indeed, the DP potential predicts that a hydrophobic bead approaching two clustered hydrophobic beads characterized by an implicit desolvation potential must overcome a barrier twice that of the original desolvation barrier for the monomer. As these barriers continue to sum, the kinetics of a full collapse become consistently more insurmountable as beads aggregate. The real scenario of simply needing to expel some central solvent particle is lost as these artificial kinetic barriers continue to emerge. Further explanation for the failure lies in the reasoning that small hydrophobic systems' solvation behaviors possess considerably different physics than for large hydrophobic systems such as the hydrophobic polymers studied here. With that in mind, generalizing the energetics of two small hydrophobes to a single, considerably larger hydrophobic chain is intrinsically flawed.

The bizarre overextension behavior that the DP implicit solvent produces for polyM is not present with the DP solvation of polyC, likely because there is no appreciable desolvation barrier. In this respect the DP and LJ models are quite similar for polyC. The DP model for these chains shares the LJ weakness of greatly over-predicting the dominance of collapsed states, to the point where the magnitude of the predicted modal radius of gyration is consistently underestimated across all polymer length scales.

Further illustrating weaknesses in the DP scheme, Figure 2.3(a) shows the average R_g of polyM in mW water and with the two implicit water models. The average R_g of polyM scales with subunit number $N^{1/3}$ with implicit LJ solvation and with explicit water. In contrast, R_g is linearly proportional to $N^{0.9}$ for the chain implicitly solvated through the DP potential. Typical conformations of M_{30} with the four models are displayed in Figure 2.3(b). With the desolvation potential, M_{30} has a helical conformation; M_{30} represented with the other three models favors globular conformations. Ultimately, it is clear that the physics represented by the PMF between a

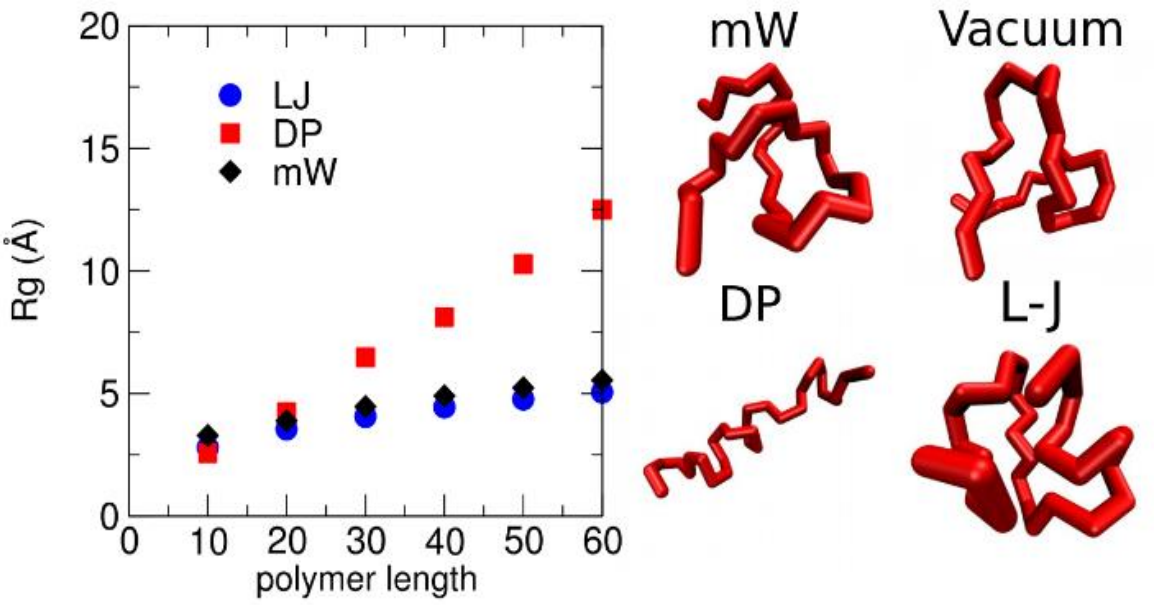

Figure 2.3: The average radii of gyration for polyM of variaed lengths obtained from the three models. On the right panel, there are typical conformations of M_{30} simulated with the four solvation models.

pair of M beads in explicit water cannot be scaled up, through addition of these effective pair interactions, to predict the physics of the whole hydrophobic chain in water.

As a final note on the success of the direct interaction scheme (i.e. vacuum simulations) against the two implicit water models, it is beneficial to consider the elements involved in the solvation of hydrophobes. Solvation of such systems can typically be viewed in two steps: the creation of a suitably-sized cavity in the solvent, followed by the activation of relevant chain-water interactions[45,47,48]. For the systems studied here, the direct interactions between the polymer beads is the primary driving force for collapse, hence the success of the vacuum simulations. Naturally, microstates exist in which solvent particles are captured within the hydrophobic cavity. However, subsequent expulsion of these particles is energetically favored[49], with the most stable globular states consisting of directly-interacting hydrophobic beads without intruding solvent particles.

Given the success of the vacuum simulations in modeling the collapse of hydrophobic

homopolymers, we turn to an assessment of the effectiveness of vacuum simulations in sampling the configuration space of heteropolymers. We introduce four models of heteropolymers possessing hydrophobic M beads and hydrophilic beads that can hydrogen bond to mW water (Figure 2.4). We compute for these sequences the PMF in explicit mW and in vacuum (Figure 2.5). The first sequence, a hydrophobic chain of M beads capped with a short hydrophilic cap, appears to replicate the results of the homopolymers in terms of both methods producing relatively similar PMFs, with the single-cap peptide essentially behaving as a shorter homopolymer. The primary difference is that the explicit solvent promotes a stronger presence of extended states, resulting in lower energy extended conformations.

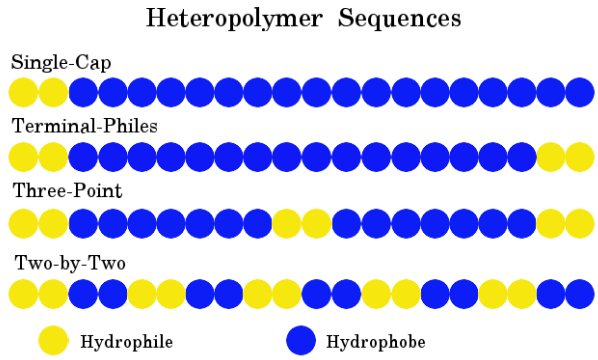

Figure 2.4 Heteropolymer models with hydrophobic (blue) and hydrophilic (yellow) beads.

The relative success of the vacuum simulations for the primarily hydrophobic Single-Cap model is not repeated for the other three heteropolymers. Two of these systems, containing more than one locus of hydrophilic sites, distinctly possess states in which they were folded (characterized by relative collapse) and unfolded (characterized by relatively linear conformations), ably described by the radii of gyration. These results are reminiscent of those in the work of Matysiak et al.[50] with similar models. The fourth and most hydrophilic model chain samples a collapsed state under these conditions. Vacuum simulation of the terminal-philes model results in a significant barrier separating the collapsed state from the extended state, stressing the difficulty of exchanging conformers at lower temperatures. This transition barrier is significantly attenuated by the presence of the explicit solvent, stabilizing transition states as the hydrophilic moieties migrate away from

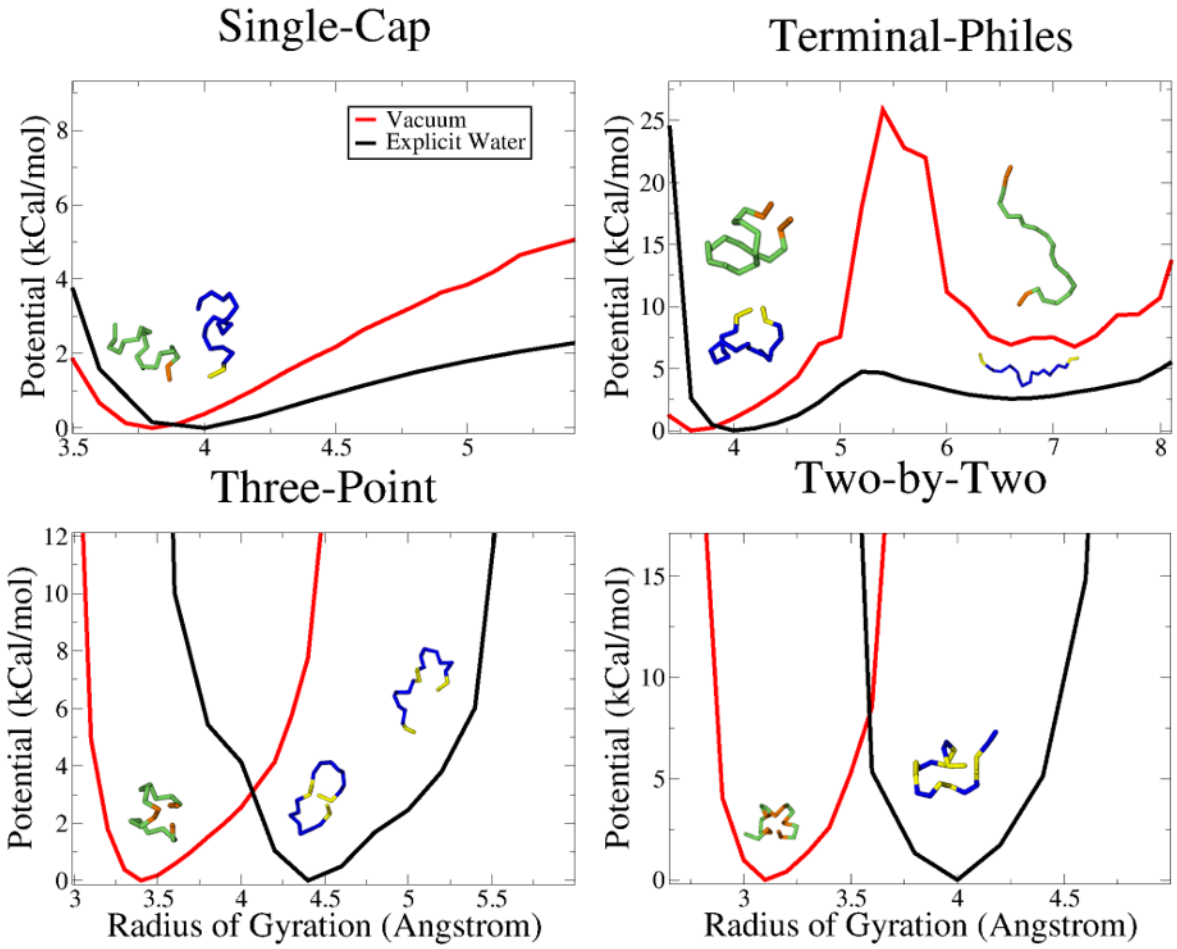

Figure 2.5 PMFs for the four heteropolymers with 20 beads of Figure 4 solvated in explicit mW water (black) and in vacuum (red). Structures in blue-yellow are for explicit solvation, structures in green-organe correspond to vacuum.

another and allowing more extended structures within explicitly solvated trajectories. The Three-Point model similarly exhibits these differences, with the folded state being the only conformer appreciably present in the direct interaction scheme, while extended conformers were enabled by the presence of solvation of the hydrophilic groups.

The Two-by-Two sequence is the most hydrophilic polymer studied. It possesses the strongest tendency for the chain to self-interact, and further illustrates the point that the direct interaction scheme is insufficient for reproducing the distribution of conformations of the heteropolymer in water. In both the explicitly solvated and direct interaction schemes, the polymer remains extremely collapsed, driven by the attractions between hydrophilic moieties. The extent of

collapse, however, is clearly overstated by the direct interaction scheme, which yields a PMF with both a noticeably smaller average radius of gyration as well as a conspicuously smaller dispersion in allowable conformers. The presence of explicit solvent appreciably stabilizes the wider more extended states, reducing the drive of the hydrophilic moieties' self-attractions.

In the case of sequences with folded and unfolded states, it is clear that the strength of the interaction with hydrophilic beads and solvent particles plays an important role in determining the landscape of accessible conformers. Without explicit solvent molecules to compete with the force of hydrophilic self-attraction, direct interaction overwhelmingly favors states in which the hydrophiles are very close. The chain in the explicit solvent explores competing configurations, in which hydrophiles are stabilized by interacting with nearby water. This results in a relaxing of the drive of the hydrophilic beads to cluster. The attenuation of transition barriers within the conformational landscapes of explicitly modeled systems clearly supports the idea that solvent-hydrophile interactions competing with hydrophile self-interactions are crucial elements of producing accurate PMFs for these chains.

Thus, the suggestion by the homopolymers that the direct interaction might be capable of producing acceptably accurate results in comparison to explicitly solvated systems falls apart upon the introduction of beads with strong solvent interactions. Alongside a general tendency to produce minima corresponding to noticeably more compact structures, the direct interaction scheme repeatedly represents significantly higher barriers to polymer extension and collapse, thus excluding multiple conformers that would be otherwise present. Ultimately, the consideration of heteropolymers makes the point that generalization of the effectiveness of direct interaction in modeling the homopolymers begins to fall apart upon introduction of hydrophilic beads.

2.4 Conclusions

In this work we first parameterize two implicit water models based on the free energy of association of subunits in mW water for two types of hydrophobic polymers. The implicit desolvation potential reproduces both free energy minima at the contact pair and solvent-separated pair, while the implicit LJ potential only reproduces the free energy minimum at the contact pair. We then use the explicit solvent, implicit solvent models, and vacuum simulations to compute the free energy of the polymers as a function of their radii of gyration, for various length of polymer chain. We find that vacuum simulations are able to adequately capture the radius gyration of the compact state, while the implicit LJ potential over collapses the polymers. Depending on the height of the desolvation barrier between monomers, the implicit DP potential either over collapses the polymers, or prevents the polymers from properly folding. In the case of large desolvation barrier (as is the case for polyM), the polymer adopts extended (unrealistic in the context of explicit solvent) conformations as the polymer is unable to collapse while satisfying the demands of creating solvent separated pairs. This is not to say that the desolvation potential is not of utility. Indeed, when coupled to a Go-model potential (that enforces well-defined folded states by assigning attractive interactions between pairs that are known to be in contact in the collapsed structure), the desolvation potential can provide important insights into the process of water expulsion from hydrophobic cavities[46,49,51]. Overall, the implicit water models and direct interaction simulations cannot stabilize the unfolded state of polymers to the extent that explicit water can. The success of the vacuum simulations can be attributed to the fact that the direct interactions between beads in the studied homopolymers play a greater role in the collapse of the chain than water-mediated interactions.

The direct interaction schemes are tested against explicitly solvated models for four types of

heteropolymer chains. The vacuum simulations do not meet the same success as in the case of the homopolymers, with the heteropolymers populated more compact structures than in explicit solvent and encountering larger barriers between collapsed and extended states. The degree of success of the vacuum simulations in capturing the details of the explicit solvent PMFs depends heavily on the sequence patterning of the heteropolymer chain. These observations indicate that, despite the results of the hydrophobic homopolymer simulations, direct interaction fails to be a viable method for studying peptide-like chains that contained moieties with strong solvent interactions.

Chapter 3) Latent Space Representation of Molecular Dynamic Ensembles

3.0 Forward

The following work is to be published in the Journal of Physical Chemistry B. It is currently in press with the DOI 10.1021/acs.jpcb.0c05763. It is reproduced here with minor differences with permission from the publisher.

Chapter 2 was a systematic consideration of various ways of modeling the interactions of solvent with a solute, applied specifically the conformations of alkane- and peptide-like coarse-grained polymers in water. In the greater context of this dissertation, it was an entry point into considering the advantages and disadvantages of using molecular dynamics to produce ensembles of data, and from that data infer the structural tendencies of the analyte molecules of interest.

The ability of molecular dynamics to produce visualizations of the exact molecular systems of interest is one of its greatest advantages, however the volumes of data generated by even relatively modest simulations can be prohibitive to analyze frame-by-frame. This is an especially salient point when it comes to non-equilibrium simulations, such as in cases where a dynamical process is the primary subject of interest. In Chapter 2, the data was synthesized to produce equilibrium representations of the free energy landscape. The tools of statistical mechanics – manifest in the WHAM algorithm – allowed for the synthesis of the vast quantities of information from Chapter 2's simulations into easily-interpreted landscapes with physical interpretation.

In Chapter 3, we consider the issue of when a non-equilibrium process must be modeled. Specifically, we use a molecular dynamics model to generate trajectories of such a process and try

to interpret the data so that we can gain insight into the general progress of the system toward a stable end state.

For systems of scale, the phase space is enormously complex – for a system of N independent particles, the system possesses 6*N – 3 degrees of freedom. Consequently, simulations that approach the scales necessary to capture the dynamics of more detailed models of biopolymers have enormously complicated free energy landscapes.

Much theoretical work has been done in this area, including analytical approaches under the label of transition state theories as well as advanced sampling techniques that favor the generation of trajectories resulting in structures of interest. These methods attempt to synthesize actionable information from systems of potentially enormous complexity.

One primary modeling technique that pursues such tools is the use of order parameters. These constructions – intuitively or algorithmically produced – attempt to reduce the representation of a molecular trajectory from its native dimensionality to relatively few quantities, which describe broadly the degree of progress of a system toward an end state of interest. These are enormously helpful when considering systems where the number of possible pathways toward a general end state are difficult to explicitly enumerate, preventing situations where recrossing methods can develop rigorous portraits of the transitional landscape.

Order parameters, as models, are naturally not without their flaws. Outside the world of phase transitions, their definition can be argued as lacking rigor. In the most general sense, they are simply quantities that provide insight into a system's state between two endpoints. Despite this, their use is established in literature, and forms a helpful descriptive tool for quantifying the behavior of complex systems undergoing dynamic shifts.

Chapter 3 marks an intersection of the molecular dynamics modeling method and the information first modeling method. Serial trajectories of a coarse-grained model of amyloid assembly are generated, exploiting a property of the model to produce a diversity of kinetic pathways toward a similar end state. In a second layer of analysis, an order parameter of this system is automatically generated using a neural network architecture. This order parameter model is compared to an order parameter commonly used within the literature to represent similar ordering processes and is found to present a more tailored image of a given trajectory than the conventional tool.

3.1 Introduction

Amyloid fibrils are associated with the clinical pathology of numerous diseases, including Alzheimer's disease[52–54], ALS[55], Type II Diabetes[56] and Parkinson's disease[55]. They 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[57–59]. Because of their pathological connections, there has been considerable investigation into the means by which these fibrils form through both experiment and simulation for the purposes of finding insight and developing potential treatments[60–62]. Fibril formation exists at the crossroads of multiple intriguing phenomena, requiring the consideration of many contributing factors in the formation of fibrils[63–67]. The complexity and diversity of species associated with these structures suggests a comprehensive understanding of their formation would involve the consideration of a wide diversity of potential formation mechanisms, presenting an important challenge for synthesizing and interpreting information related to them.

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 fibril forms.

While fully atomistic studies of the formation of a fibrillar plaque often yield important insights[68], they remain limited by time and scale constraints in terms of offering a serial, full picture of the process. Coarse grained models have been successfully used to model fibril formation from unassembled monomers [69–74] as well as facilitating the comprehensive investigation of intrinsically disordered peptides[75], allowing for the exploration of system evolutions over large timescales and revealing a diversity of mechanisms for assembly. As computation improves, the simulation of systems of increasing complexity will become more common. This will require enhancing the researcher's capability to assess the behavior of systems with potentially many different mechanisms evolving toward common or closely related end states. Coarse graining offers a means of studying systems with common traits and allowing investigation of a wide variety of processes, including amyloid formation.

In this work, we apply a common type of artificial neural network – the variational autoencoder (VAE) [76] – 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. Order parameters, in which highly dimensional systems are reduced to a few parameters, are useful descriptors of major advances along a dynamical pathway. Here we build on work that improved molecular dynamics sampling and analysis utilizing the Ising model[77–79] by investigating the use of VAE's to find single-value parameterizations of two amyloid-associated processes. The VAE's ability to identify intermediates, characterize phase transitions and act as automated feature selection are considered.

Latent models are the result of unsupervised learning of data, in which a large bulk of information is reduced to a smaller set of values. They have been used successfully to extract information relating to protein fitness landscapes[80] as well as help characterize smaller dynamic

systems in biochemical simulations[81]. These successes suggest similar models may be a useful tool in analyzing larger systems.

In this paper we present the first application of the VAE to determine order parameters for the process of holistic protein aggregation and fibril formation, demonstrating that this approach can be used on systems larger and more dynamically complex than previously studied. We consider molecular dynamics trajectories of the aggregation of a coarse-grained peptide model developed by Bellesia and Shea [82,83]. The level of resolution of this model preserves the major common attributes of fibrillizing peptides and allows access of system sizes large enough to gain mechanistic insight into the process of aggregation. The process allows following a simulation from monomer to a final ordered solid and enabling exploration of multiple potential mechanisms. We will demonstrate how the VAE can use the information in a molecular dynamics ensemble of a serial ordering process to develop a single-valued order parameter for two aggregating systems: First, a rigid system of monomer peptides undergoing assembly into beta-rich fibrils, followed by rearrangement of these solid moieties into a greater fibril stack, and second a system of flexible peptide units first assembling into a liquid globule and subsequently undergoing a liquid-to-solid phase transition. The automatically learned order parameter for these systems will be compared to the commonly used nematic order parameter[84–87] as a means of checking performance. The nematic order parameter has been used in numerous computational studies to follow the progression of protein aggregation. We show here that the VAE generated automatic order parameter not only outperforms the nematic order parameter, but that in certain cases the nematic order parameter incorrectly describes the progress from monomeric to aggregate state and even in one case fails to predict a fibrillization event. Because the VAE develops its latent model from data immediately associated with a specific evolution, its order parameter offers a more tailored

description of the aggregation process and enables insights into events significant to the mechanism of fibril formation. Additionally, the VAE's ability to internalize and simplify the statistics of the transient ensembles associated with the dynamical processes enables for a streamlined means of analysis. By comparing the latent model's representations of various states' ensembles one can rapidly extract the features contributing to the characterization of the process. This enables a relatively 'hands free' means of detecting quantities associated with states of the dynamic process, and avoids the need for assumptions to be made about general qualities of diverse families of mechanisms to intuit reduced model parameters

3.2 Models and Methods

Shea Amyloid Monomer Model

The peptide model is a phenomenological mid-resolution coarse-grained model developed in earlier work by Bellesia and Shea[83]. The monomeric peptide unit is depicted in Figure 1. 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, motivated by combinatorial studies of Hecht and collaborators [88,89] Hecht et al. that indicated a generic amphiphilic alternating pattern is a major indicator of beta-sheet secondary structure in self-assembling peptides. This trend was also supported by recent analysis of the Waltz database using classifying autoencoders [90], further suggesting that this pattern is

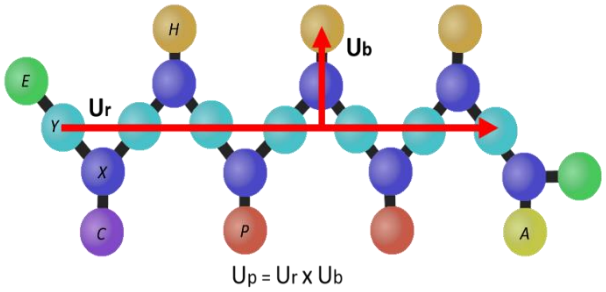

Figure 3.1 A diagram of the peptide model with important vectors labeled. The formula provided is the conventional cross product. All vectors are normalized.

favored in fibril-forming systems. A major useful feature of this model is the fact the model's preference for an amyloid-ready stacking state is controlled by a single value that modulates the dihedral. This allows for a diversity of amyloid-forming mechanisms to be explored, interpretable as the sequence's conformational preference toward beta-sheet-forming structures. In our simulations, the rigid peptides are associated with monomers that show a high degree of preference for a planar, stack-ready conformation, while our flexible peptides display greater conformational variety.

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 3.1.

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 $\varepsilon = 1.0$, $\sigma = 2.0$ and $\lambda = 0.0$. Energy constants are in kcal/mol and length constants are in Å.

Simulations & Nematic Order Parameter

Simulations were run in the molecular dynamics package NAMD[91] using Langevin dynamics with 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[84–86]. 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 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 are the two common end states for the systems studied.

Variational Autoencoder

A variational auto encoder is a method of dimensional reduction using an artificial neural network. It has been described in detail elsewhere[76], 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 has many inputs. Depending on the nature of 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 3.2. Each gray arrow has a fitting parameter associated with it. The fitting parameters are fit to the data such that when an input is fed into the model 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). 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 at the beginning of the simulation bear similarity to other frames and will encode to similar points 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. Because the variational autoencoder possesses a smooth latent space, the statistics of stable states become associated with distinct regions in the latent dimension, allowing for the single-value parameterization of the system through time. We note that, directly interpreted, the term order parameter suggests a value that determines the degree of order in a system. In this case, that term is appropriate, as the latent model shows the evolution from a disordered state to an ordered state. More generally, the latent model generated by autoencoding is simply a reduced representation of the temporal ensemble. Thus, it will discriminate between states of similar order if they are separated dynamically in the mechanistic space. This is an advantage for characterizing more general systems, as the assumption that a system evolves toward or away from order does not hold generally.

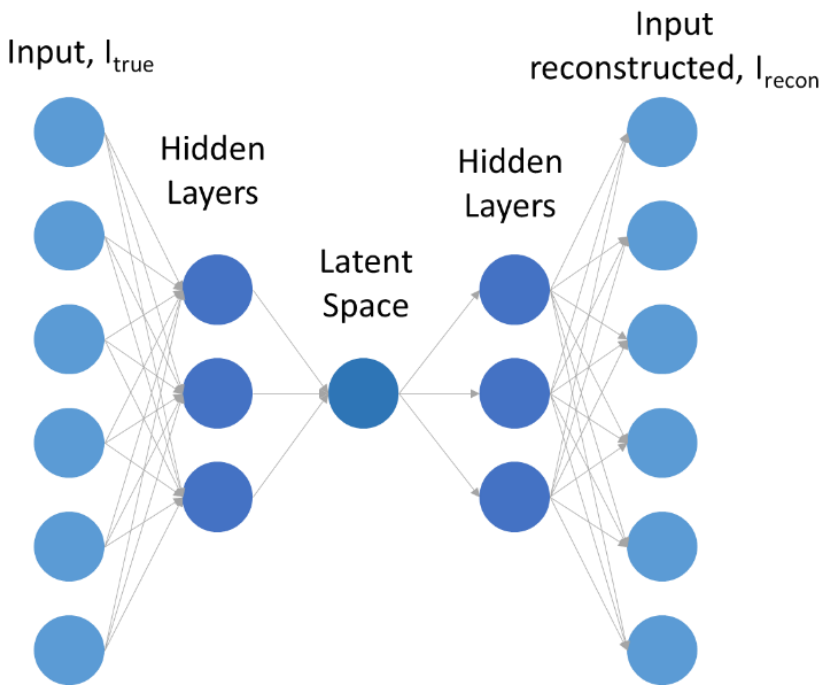

Figure 3.2 A simple depiction of a variational autoencoder. Each dot represents a node. The gray arrows show how information is passed from node to node. Inputs are fed into nodes on the left, and the network is fit to reconstruct those inputs in the nodes on the right.

Artificial neural networks were built using the Keras python package[92], using Google's Tensorflow[93] 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). The hyperparameters were tuned through a grid search and code has been posted on github at https://github.com/ncharest/Core. Molecular structure images were generated using VMD[6] and graphs were plotted using matplotlib[94].

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 associated with time points of interest. 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, the features) of 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.

3.3 Results & Discussion

This section is organized as follows. We first present results for the rigid peptides that assemble into fibrillar stacks. We compare the performance of the nematic order parameter compared to the automated order parameter (the LSV). Next, we present the same analysis for a different self-assembling system, the flexible peptide system that forms a liquid globule that then solidifies.

Analysis of Rigid Peptides

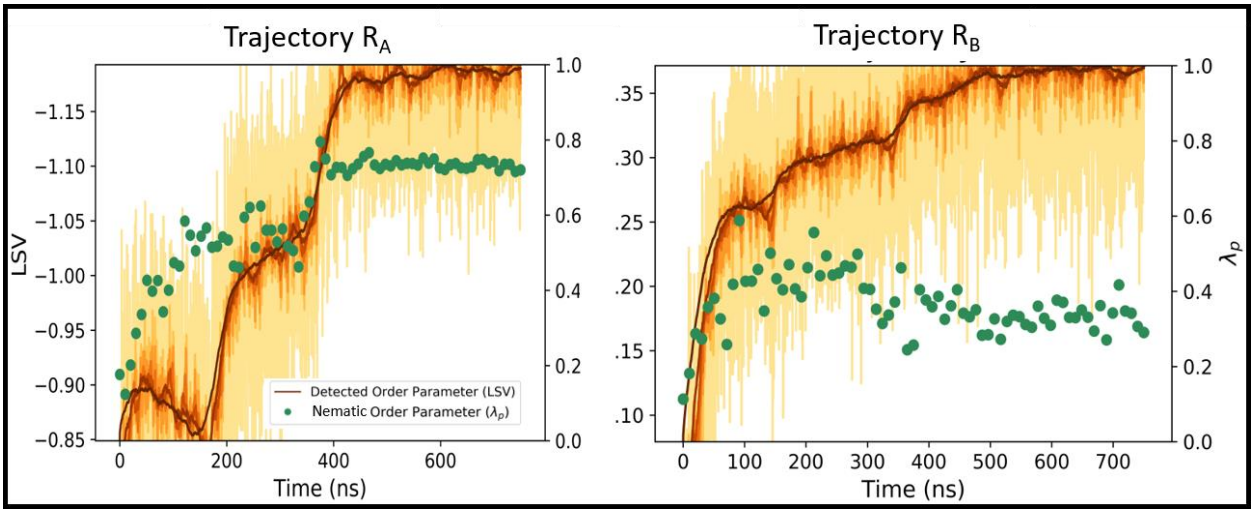

Figure 3.3 Parameterizations of representative simulations. The figure is depicted as follows. The lightest yellow line shows the exact parameterization of a frame in latent space. This fluctuates wildly, however a rolling average of these values results in the darkest red line, which functions as an effective order parameter. For clarity, differing levels of averaging are shown, indicating how the averaging process reduces the latent space to a parameter tracking system evolution. The green dots correspond to the value of the nematic order parameter. 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 relatively disordered beta-barrel-like intermediate forms, but then rearranges into a stack.

Two representative trajectories (RA and RB) that we will analyze in details are shown in Figure 3.The green dots are the value of the nematic order parameter along the trajectory. The lightest yellow line shows the exact parameterization of a frame in latent space. This fluctuates wildly, however a rolling average of these values results in the darkest red line, which functions as an effective order parameter. For clarity, differing levels of averaging are shown, indicating how the averaging process reduces the latent space to a parameter tracking system evolution. Because the absolute values of the latent space parameter are arbitrary, these values are plotted such that they are on the same scale as the relative nematic order parameter. The important take away is that, when assessing a latent space model of the trajectory ensemble, it is the differences between LSVs that is important. Plateaus represent statistical states that were populated for longer lengths of time, while the connecting transitions show the system moving through transient states. The end products of each simulation are shown in Figures 4 and 5, respectively. Importantly, R_A comes

closer to realizing a single fibril stack than R_B in the time studied. R_B concludes with several stacks, oriented roughly perpendicular to one another or otherwise twisted out of alignment. Both rigid trajectories are 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. Subsequent ordering takes the form of large-scale rearrangement of largely solid stacks, resulting in the formation of a mature fibrillar arrangement. 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). By inspecting the trajectories we are able to deduce the reasons for the divergence in their behavior as described by the nematic and latent parameters.

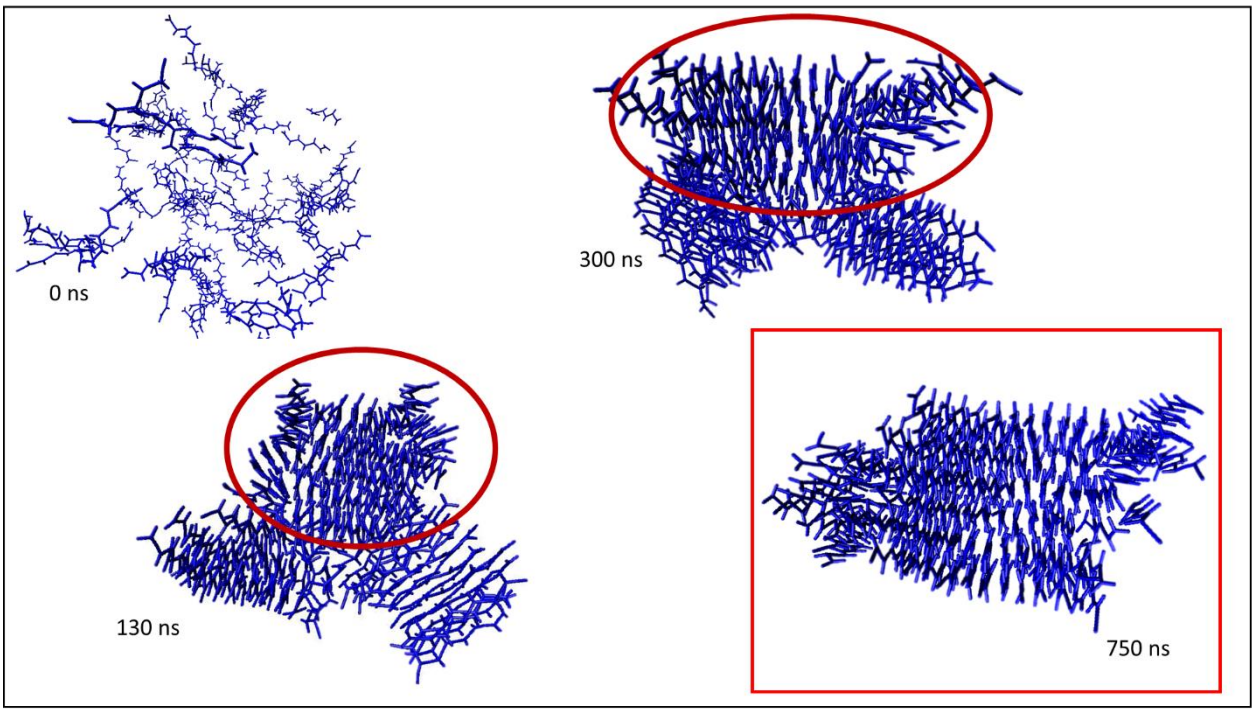

Figure 3.4 A visual summary of Trajectory (R_A). A major fibril stack capped by beta-barrel-like structures emerges (circled in red), along with two minor fibril stacks. Subsequent rearrangement results in the formation of the final state, largely a single fibril stack with only a few disordered peptides on the ends.

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 3.3) 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 plateau region during the initial simulation period. By visually inspecting the simulation, the source of the disagreement becomes clear. In the middle two images (130 ns and 300 ns) in Figure 3.3 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 apparently indicates that the fibril core condenses 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 in the system.

Trajectory R_B: Nematic Parameter Erroneously Implies Ordering With No Resulting Fibril

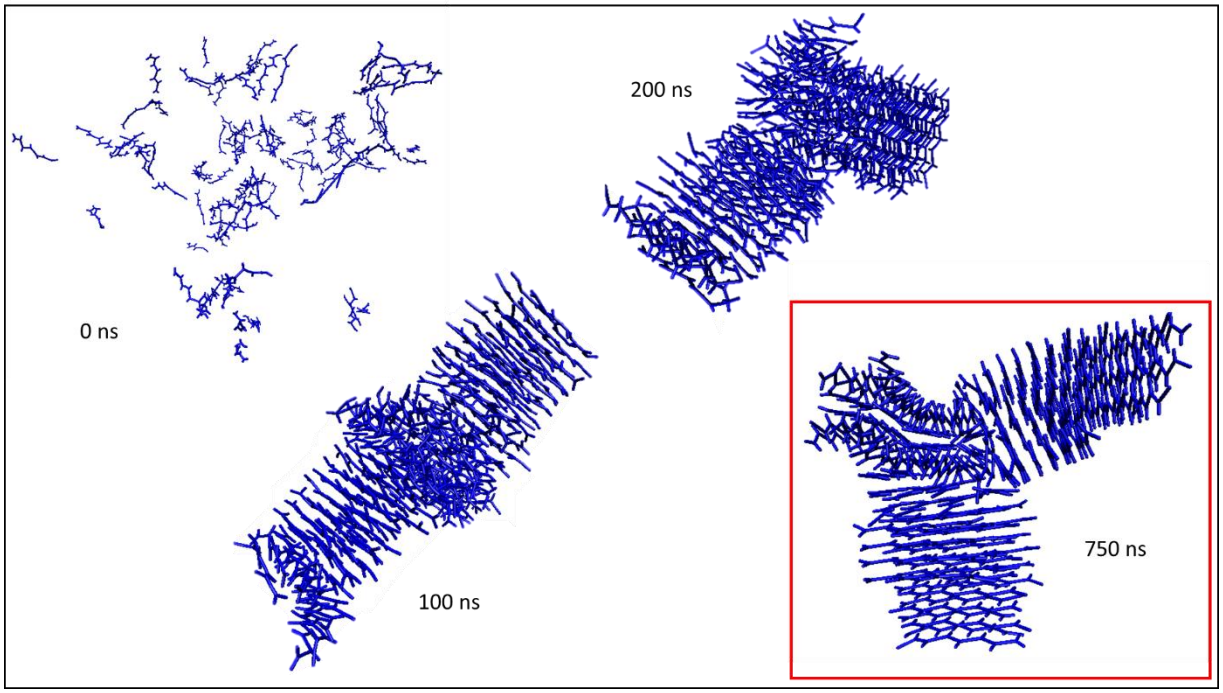

Figure 3.5 A visual summary of Trajectory (R_B). Two major fibril stacks form and are joined by a beta barrel region. Growth occurs by the absorption of smaller stacks or the reorientation of peptides from the beta-barrel. Coincidentally the two major fibril stacks shift alignment during the course of the trajectory.

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 3.5.

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 3.5). 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 3.5.

Broadly, two fibril stacks form early in the trajectory (depicted at 100 ns in Figure 3.5) 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.

Analysis of Flexible Peptides

The flexible peptides progress through a different primary mechanism of growth than the rigid peptides, with lesser diversity between different trajectories in terms of a broad mechanism. Rather than undergoing a solid-solid transition as was the case for the rigid peptides, the flexible peptides undergo a molecular liquid-to-solid phase transition. Relatively speaking, the rate at which the disordered peptides attracted into a liquid-like, internally mobile mass was faster than the rate of stacking. Consequently, 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

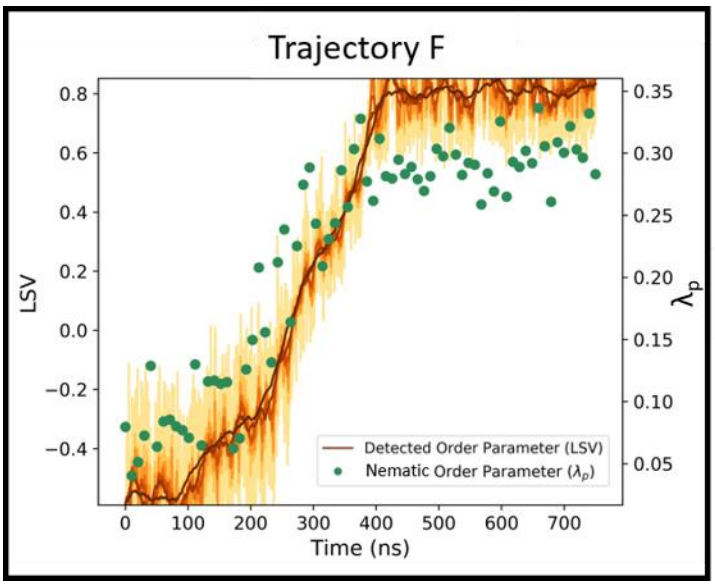

Figure 3.6 The LSV representation for trajectory F compared to the nematic order parameter. The lightest yellow line shows the exact parameterization of a frame in latent space. This fluctuates wildly, however a rolling average of these values results in the darkest red line, which functions as an effective order parameter. For clarity, differing levels of averaging are shown, indicating how the averaging process reduces the latent space to a parameter tracking system evolution. The flexible peptides undergo a liquid-to-solid phase transition with low diversity of mechanism pathways. After forming a globule characterized by high internal monomer mobility (0 – 100 ns), an initial beta-barrel-like structure spontaneously orders (~100 ns). This stable BBL templates and promotes the ordering of the globule, which spreads and solidifies the system (100 – 400 ns) in the major dynamic transition. The final BBL structure is highly stable, undergoing virtually no subsequent transitions in the length of time studied.

'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 resembles the liquid-to-solid transitions observed in literature, relating to coacervation processes and then "gelation" occurring in some peptides associated with amyloidosis.[95–98]

Trajectory F illustrates an archetypal transition for the flexible monomer peptides. The monomers rapidly condense into a liquid-like globule, characterized by a roughly spherical shape and independent internal monomer mobility. Eventually several monomers order and align their backbones into a beta-barrel-like solid structure. This promotes the ordering of nearby peptides, launching the transition from globule to stable solid BBL end phase. This process is shown schematically in Figure 3.7.

It is clear from Figure 3.6 that the LSV characterizes the phase transition very similarly to the conventional order parameter, with both parameters demonstrating a sigmoid-like transition process. The VAE's capabilities for these transitions lay in its capability to act as automated features selection, with the capacity to 'report back' the input parameters used in its

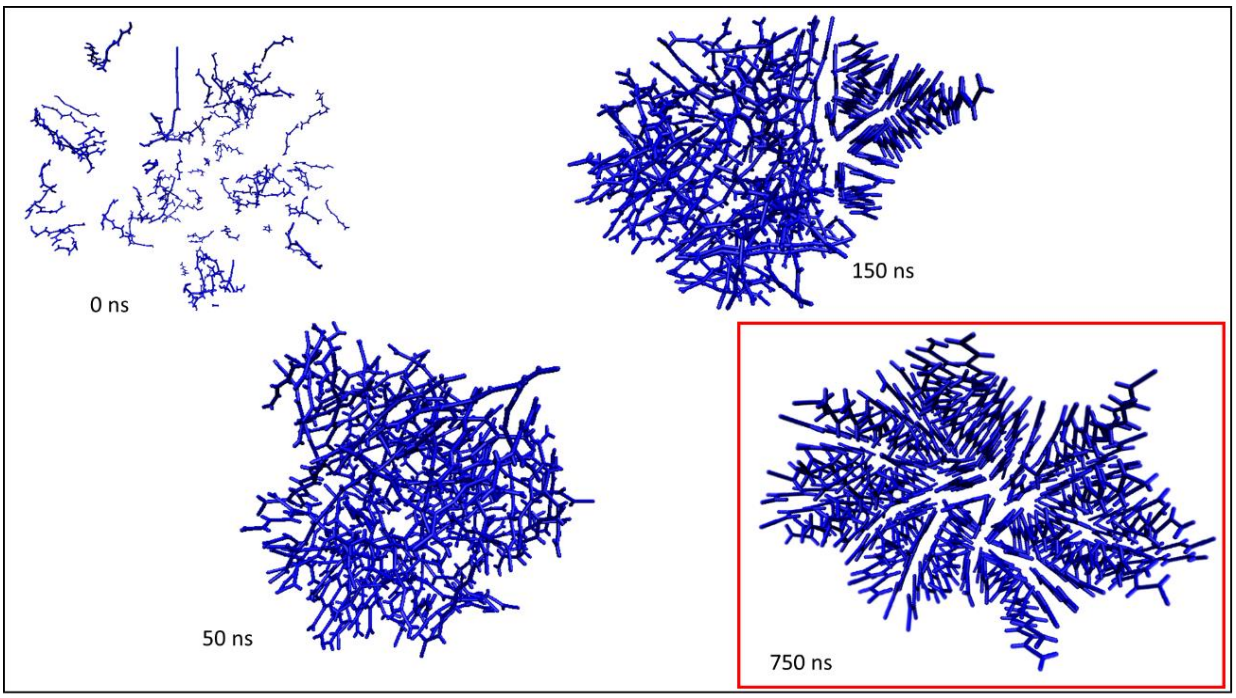

Figure 3.7 A visual summary of Trajectory (F), showing the formation of a globular, spherical disordered phase with high internal mobility. The final state is a highly immobile beta-barrel-like solid. This system forms a solid nucleus early in the trajectory and subsequently undergoes a conversion from globule to solid.

characterization through a systematic process. Because of the smoothing around points in latent space for a given set of inputs, the VAE automatically filters input values for thermal fluctuations and statistical noise in addition to detecting the input features whose statistics most effectively characterize the course of the system. We can isolate these features using the reconstructions of LSVs along the trajectory.

This isolation process proceeds as follows. A point in the LSV associated with the end state is chosen as a reference (for example, for figure 3.6, we can choose $LSV_{ref} = 0.75$). This reference point is identified with the statistics of the end state, where the statistics are smoothed by the VAE averaging and denoising. We can now scrutinize an analyte state by choosing its LSV and examining the differences between the analyte and reference states' reconstructions. This is not so trivial as comparing the frames associated with either state directly. The VAE performs automated denoising and statistical consolidation of input features as part of its characterization process.

We use this process to isolate the dihedral angles whose average values best characterize the difference between the reference (BBL end) and analyte (initial globular) states. Figure 3.8B shows a histogram of the residuals found by subtracting the reconstruction of the analyte state from the reference state. By setting a threshold at some distance from zero (no difference between reconstructed values), and scanning different thresholds, we produce the curves of Figure 3.8A. These curves are the average dihedrals of the subsets identified by the process of Figure 3.8B. The threshold is chosen by considering the fraction of residuals that are above versus below it, indicating the number of dihedrals that are significantly different between analyte and reference as a fraction of the total dihedrals. More directly: the higher the threshold, the more extreme a difference between analyte and reference reconstruction must be registered for a given dihedral to be accepted into the set of 'contributing' features, the average of which is plotted in the purple line. The darker the purple, the more exclusive the contributing subset. Each contributing subset is identified with a 'non-contributing' set, which contains the dihedrals that did not significantly change between reconstructed analyte and reference states.

We find that by looking at the 10% result of the reconstructed dihedrals that differ the most between analyte and reference, we can 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 affected by the fibrillization process. The fact we can directly compare the results of this process to the over-arching LSV is almost certainly a property of the fact that the system possesses two major dynamical ensembles. The statistics of the liquid ensemble can then be associated with

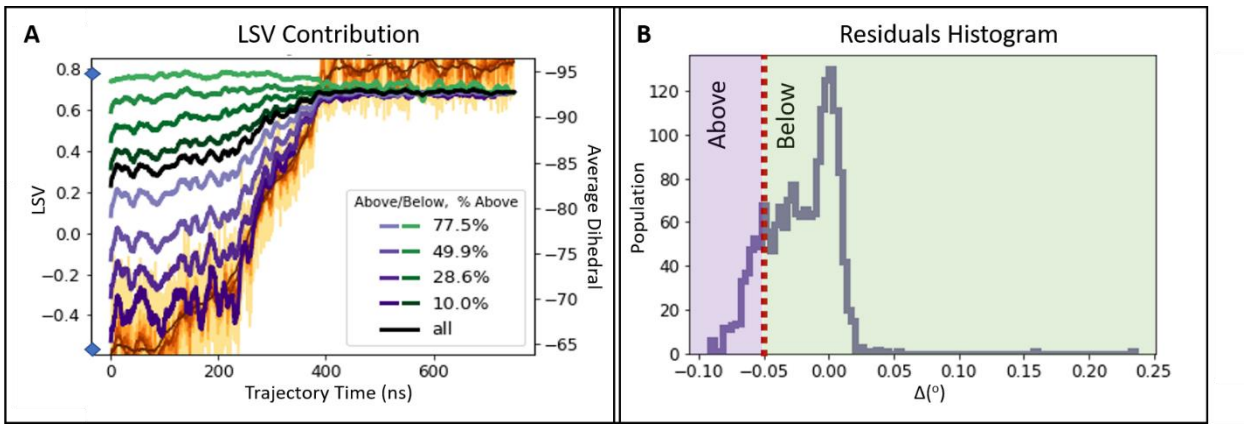

Figure 3.8 Figures illustrating the reference-analyte comparison process enabled by the VAE parameterization. LSV values associated with two states of interest (in this case the two values labeled by blue diamonds in panel A) can be reconstructed and each dihedral change (Δ) computed between reference and analyte reconstruction. These residuals are binned into a histogram (B), yielding populations with varying degrees of difference between reference and analyte. Defining a threshold allows the creation of a subset of 'contributing' features (purple), whose average values are shown to correlate with the general LSV-defined transition between states (A). The non-contributing converse subset (green) shows little mimicry of the general LSV.

one region of latent space while the statistics of the solid ensemble are associated with another, with the mechanistic pathways between them traversing between those regions. For a system with more numerous stable states and dynamic ensembles, this process would elucidate the transitional pathways between the chosen reference and analyte.

This ability to automatically isolate contributing features is useful for systems of scale, when identifying relatively small portions of large simulation volumes might require an otherwise difficult-to-intuit 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.

Sensitivity of VAEs As Feature Selection

The idea that dihedral angles of the model might serve as a good indicator of whether or not the system is in a globular liquid state is intuitive given an understanding of the model and the

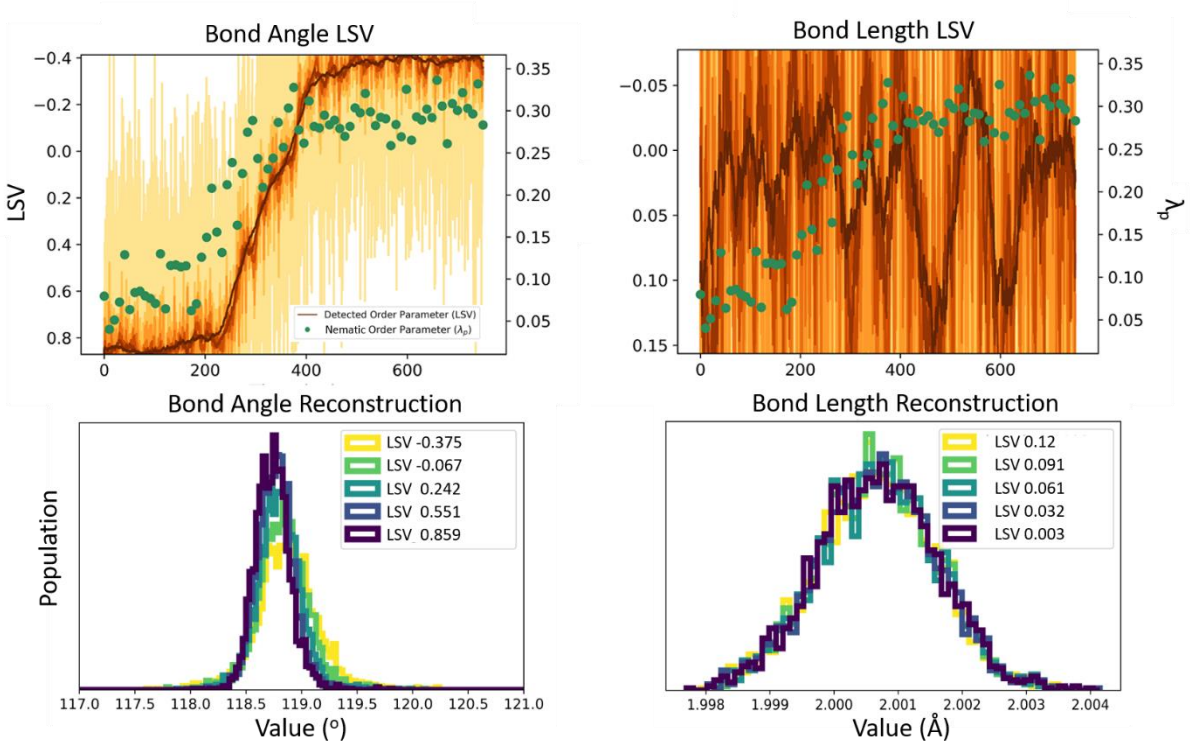

Figure 3.9 A comparison of the LSV trained on bond lengths alone versus an LSV trained on bond angles alone. Neither internal coordinate shows a relation to the difference of states like the dihedrals in Figure 6. Despite this, the VAE can parameterize a similar curve, albeit a noisier one, with the bond angles while failing with the bond lengths. Histograms of the reconstructed angles and lengths offers insight into how the statistics of the BBL state and globular state differ, and why the angles were able to parameterize when the bond lengths could not.

dynamical behavior of the liquid globular state versus the solid BBL state. The ability of the VAE to pick out these major contributing features of the system automatically is valuable. However, we can also inspect cases where the relevance of the features is not immediately obvious because it is so minutely realized in the difference between states that it might be overlooked. Variational autoencoders present themselves as highly sensitive tools for assessing the descriptive potential of such features, able to provide insight into whether a set of features is useful for differentiating between states and providing a qualitative sense of how indicative the groups might be.

Figure 3.9 presents the LSV parameterization of trajectory F as trained on only the bond angles or the bond lengths. The parameterization presented in Figure 3.6 is a synthesis of all internal coordinates – bond lengths, bond angles and dihedrals. It is intuitive that the statistics of the

dihedral angles would display most clearly the difference between the liquid and solid state, per the discussion around Figure 3.8.

It is less obvious whether the bond lengths or bond angles would be useful in describing the transition. Figure 3.9 clarifies these points. Reassuringly, the intuitively meaningless bond lengths fail to parameterize the system with any degree of useful clarity. The statistics of the bond lengths should depend little on whether the system is in the globular or BBL state, as their vibrational behavior should depend only on the temperature of the system. This is reinforced by a histogram depicting bond length populations over the course of the LSV parameterization – it is a roughly normal distribution centered essentially at 2 angstroms, the ground state length of all bonds within the system, at all reconstructed frames along the trajectory.

The relevance of the bond angles is less immediately trivial to understand. It is not clear that the BBL geometry should preferentially alter the ensemble of bond angles, however it may be that the relatively weak constraint on the angles (40.0 kcal/mol) versus the constraint on the bonds (200.0 kcal/mol) gives the bond angles enough conformational flexibility that their statistics can be identified with the general state of this system. The LSV trained from the bond angles in Figure 9 therefore helps settle this dispute. We see it parameterizes the trajectory with the same sigmoid as the full internal coordinate treatment. This time, a histogram of the bond angles as they evolve indicates how the VAE may be separating the phases: the shift of the ensemble average by a mere 0.5 ^{o} would be difficult to predict intuitively, however it is not overlooked by the precise eyes of an unsupervised algorithm. Upon consideration, the idea that the globular arrangement of disordered, liquid-phase monomers might compress the bond angles away from their preferred value of 120^{o} more than the parallel-arranged BBL solid is reasonable. It is not the sort of

conclusion that one might immediately jump to, however, given the minute affect the difference in states has, but one that the VAE is capable of identifying.

Thus, the development of single-value descriptions of the system from sets of unlabeled data with a VAE presents a sensitive means of determining the relevance of potential descriptor variables to states of interest. If it is found that a subset of descriptors can satisfactorily train a VAE to parameterize the general transitions of interest, it is an argument for investigating those variables in greater detail as they pertain to the process.

3.4 Conclusions

VAE was applied to the trajectories of a rigid and flexible coarse-grained peptide models that could respectively assemble directly into fibrils stacks, or transition from a liquid globular state to a solid fibrillar state. In all cases, the trajectories approached a solid final state with little further evolution in the time spans studied.

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) to develop latent models of system evolution, allowing for single-value parameterization of related but distinct mechanistic processes. This results in an unsupervised means of simplifying mechanisms with reduced models tailored to identify important elements of unique pathways and bypasses a need for assumptions to be made the system. In addition, the technique acts as a proof-of-concept for characterizing systems based on unlabeled features. The success of these latent

models demonstrates it is possible to characterize the state of the system by monomer features alone. Reconstruction-based methods were shown to facilitate an understanding of which features were most significant in characterization of the transition from the dispersed phase to a final solid state. Additionally, cases where the LSV more successfully characterized important features of transitional pathways than the nematic order parameter were identified.

Using the VAE's reconstructions allowed study of which elements of the simulation were contributing to its characterization. This is useful for feature selection, and the prescribed process is applied to isolating the sets of dihedrals that best characterize the difference between the liquid-like globular state and the solid BBL end state. Training VAEs using subsets of candidate input features also enables the automated detection of the sets of features suitable for differentiating the statistics of intermediate states, and therefore suitable for developing single-value parameterizations of the system.

The results reported here show the possibility of applying VAE tools to large systems with high mechanistic complexity and dimensionality. The work affirms VAEs as a useful method of information compression, filtration and denoising and provides a means of obtaining insight into the contributing factors of a given latent representation. They are unsupervised algorithms, requiring relatively little preparation of data, and they possess high potential for automation of data-processing pipelines. Ultimately, they are a promising avenue for generating useful latent representations of complex systems of scale.

Chapter 4) Extraction of Activity-Feature Relationships Using Artificial Neural Networks

4.0 Forward

The following work was published with few changes in Tro, Charest, Taitz, Shea and Bowers[90] in the *Journal of Physical Chemistry B* on Jun 27. 2019 under the title "The Classifying Autoencoder: Gaining Insight Into Amyloid Assembly of Peptides and Proteins", and has been presented here with permission from the publisher.

In the context of this dissertation, it is the first pure application of the information approach models introduced in Chapter 1 and demonstrates a relatively interesting ability of ANN techniques to shape themselves in controllable ways through their objective training functions. Unlike Chapter 3, it does not attempt to use the data generated in mass by a molecular dynamics simulation to produce insights. Instead, it uses an experimentally measured database.

Here, we use a similar artificial neural network as in Chapter 3, however the goal is no longer the purely unsupervised reduction of data dimensionality for the purposes of simplification. While the purview of Chapter 3 was the creation of simplified representations of large ensembles of noisy, potentially unrelated descriptor data, we take a more focused approach in this work by using a supervised technique that couples an ANN classifier with the VAE architecture.

This allows the structuring of the latent space, a technique generally referred to as conditioning the autoencoder. This technique is relatively novel in chemistry, though it has seen application in other fields. It represents a more refined attempt to use a latent model of data to extract insights from the dataset, however it comes with cost of requiring labeled data points. It steps away from using classical simulations to explicitly simulate insights for analysis; instead, this model indicates trends in the data in a way accessible to the researcher. Unlike the combined model of Chapter 3,

however, there is no real way to elucidate exactly how the predicted data patterns lend themselves to amyloid formation, leaving it open to interpretation.

Despite this, the Classifying Autoencoder shows a convenient means of analyzing labeled datasets and extracting trends in the primary sequence that are associated with the amyloid forming activity. It reproduces amyloid-forming motifs identified in prior literature, and the insights it offers could be used to augment first-principle models and facilitate more immediate insights.

4.1 Introduction

In this work we combine two types of machine learning algorithms by coupling two distinct artificial neural network architectures. This enables us to separate hexamer peptide chains based on whether they produced amyloid fibrils in solution, and subsequently visualize the characteristics that were associated with a tendency to produce amyloids. By employing this method, we can analyze the value of novel features (descriptors) used to describe peptides while simultaneously finding patterns in the data associated with the activity of the peptide.

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 [99–102]. Attempts have been made to elucidate the inner workings of ANNs [103,104], however these methods can still leave intuition difficult to obtain.

In this work we view classification 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 [105]. Differing versions of the basic autoencoder, perhaps most notably the Variational Autoencoder (VAE)[76] have emerged, with variants typically involving goals beyond dimensional reduction and reconstruction of the data [106]. In this paper, our goal was to develop a method of classification, which we call the classifying autoencoder (CAE), based on prior algorithms[76,101,102], 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[107]. There are relatively few known examples of smaller peptides which form amyloids[108,109]. 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[107]. 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. This process validates choices made regarding the Shea coarse grained model and supports the conclusions of its original development. 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 to illustrate a major investigatory power of the method. 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[110] 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.

3.2 Models, Methods & Proof of Concept

Given the goal of recovering visualizable patterns within primary sequence data of hexamers, we first decided to generate a toy database with an embedded structure-activity relationship and attempt to automatically recover it using the classifying autoencoder technique. 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 4.1 depicts a flow chart of the process used to generate this database.

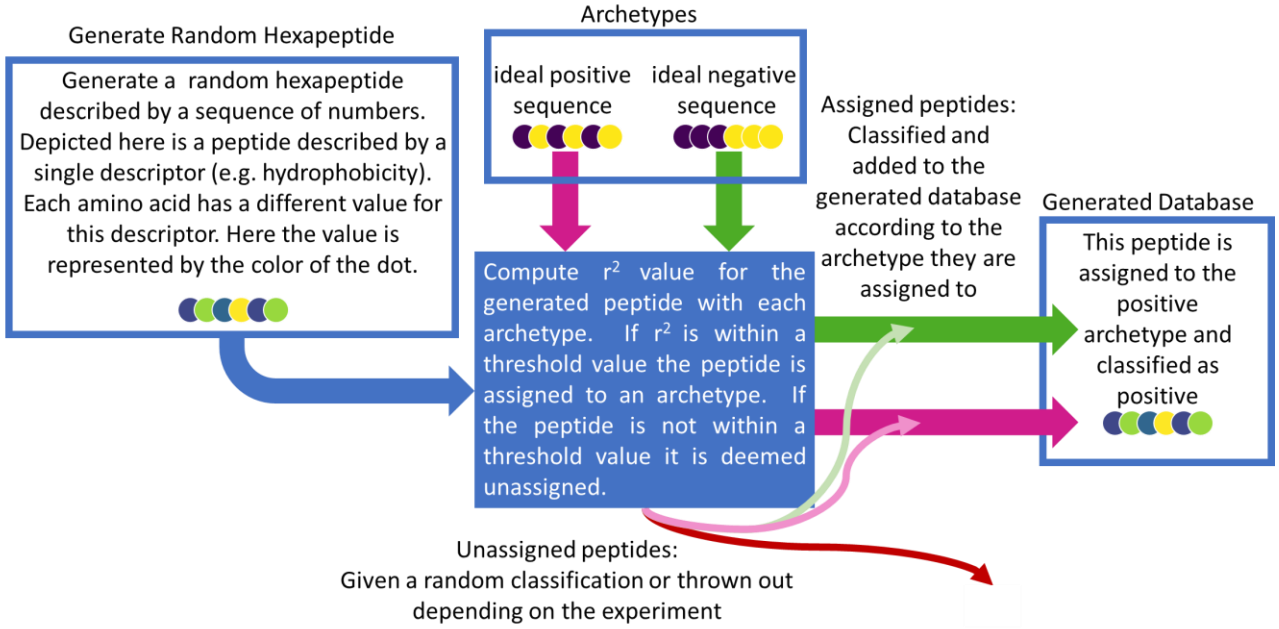

Figure 4.1 This flow chart depicts the generation of the database. In the example depiction shown in the flow chart a random peptide is added to the database by being assigned to the positive archetype and classified as positive.

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

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

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

Experimental Database

Given that our goal is to better understand how the physical descriptors of a peptide relate to amyloid activity, we used a database of experimentally-verified peptides. We use the Waltz-DB[107,111] 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. We opted for a larger validation set at the cost of a

smaller training set since the database is relatively small and we wanted to make sure there was enough data in the validation set to get a good idea of how generalizable the model is.

Polarity Descriptor

We found the hydrophobic parameter using the AAindex database[112–114] . 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[115] . We choose this metric for hydrophobicity because it correlates with many of the other hydrophobicity metrics in the database, it performs well for classification, and has a clear experimental basis and intuitive interpretation.

Cross Section Measurements

To measure the DID, amino acid samples were dissolved in water to concentrations between 1 and 12 millimolar. The cross section of the singly charged amino acid, and the cross section of the singly charged dimer cluster of the amino acid were measured using a lab-built ion mobility mass spectrometer which is described in detail elsewhere[116] . 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[117]. 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 [118].

Software

All neural nets were constructed and trained using the Keras software package[92] with the Tensorflow backend[93]. Images were generated with matplotlib[94].

Optimization of Model Hyperparameters

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 [119]. 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.

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 [76]. Interestingly, we did not find the relative weighting of these terms affected the maximum MCC score a configuration can achieve (Figure S4.1). 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 Figure S4.1, 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 Figure S4.1 achieve this.

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 Figure S4.1.

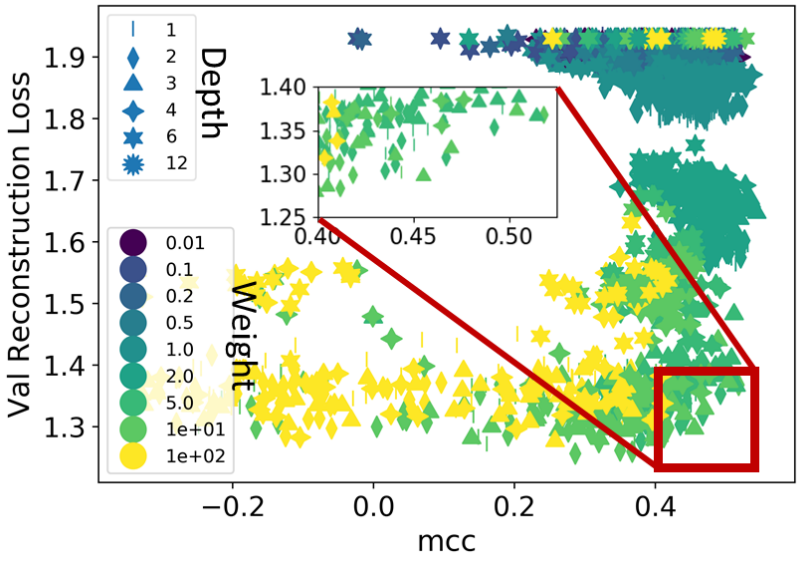

Figure S4.1 A scatter plot over 50 trainings with each combination of hyper parameters shown. The depth of the model is the number of hidden layers, while the total number of hidden nodes was held constant at 24. The weight is a scaler multiplied by the reconstruction term of the loss function. MCC is 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.

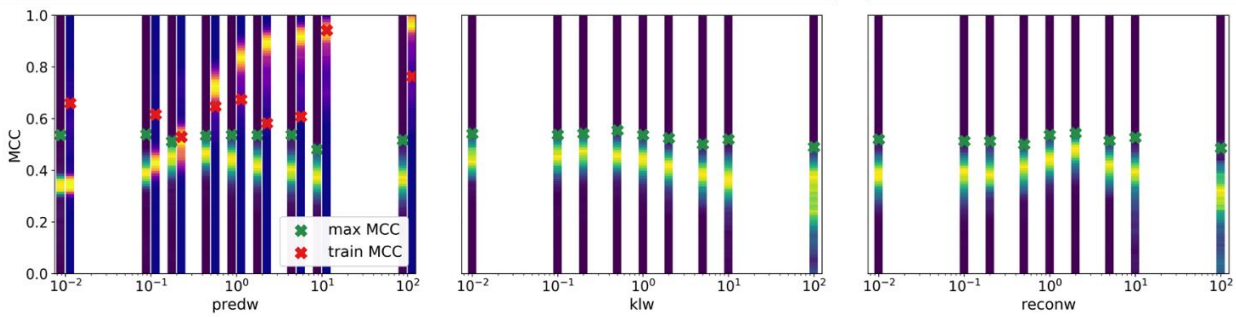

Figure S4.2 Histograms of models trained with differing weights in the loss function against MCC of each model. To optimize the weights in the loss function, each weight was varied over 5 orders of magnitude from 0.01 to 100, while holding the other two weights at 1. The heat bars represent the MCC distribution 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.

4.3 Results

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 4.2. Inputs (a description of the peptide) are fed into the model via the input nodes. The depiction in Figure 4.2 shows only four input nodes, but in the final model there will be an input node for each value that represents the peptide, i.e. the number of descriptors times the number of amino acids in the peptide. The hidden layers add more fitting parameters. The nodes labeled μ represent what is termed the latent space. Typically, the term latent space is used to refer to the space of the reduced representation. Here, these values are also used as the prediction. The latent space is two dimensional, one for the amyloid propensity and one for the non-amyloid propensity. A peptide is classified depending which node outputs a higher value. The nodes labeled $N(\mu, \sigma^2)$ inject noise into the data during training. This noise is in the form of a normal distribution centered at the reduced representation, μ, and has a standard deviation, σ^2. The nodes to the right of the nodes labeled $N(\mu, \sigma^2)$ (the decoder) attempt to reconstruct the original input. For a more detailed explanation of this please see the primer of VAEs in the supporting information.

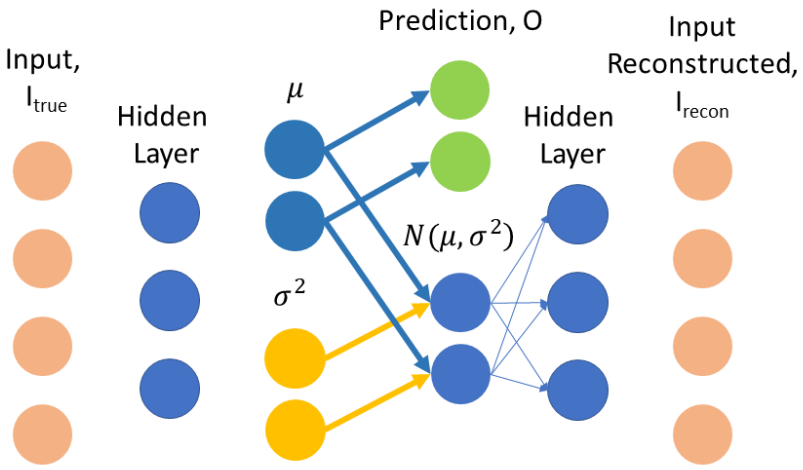

Figure 4.2 Depicted is the architecture of a classifying autoencoder (CAE) with four inputs, and one hidden layer with three nodes. The latent space in this model is two-dimensional and is labeled μ. These nodes are also used as the prediction layer. Noise is added to the latent space at the nodes labeled $N(\mu, \sigma^2)$; this noise is in the form of a normal distribution centered on the reduced representation, μ, with a standard deviation, σ^2. The decoder (all nodes to the right of the nodes which introduce the noise) tries to reconstruct the input. Dense connectivity (see primer of VAEs in Supporting Information) can be assumed for all layers not drawn explicitly. In addition, connectivity has been drawn explicitly at the latent space to highlight that the output nodes are not fed into the decoder, but the nodes labeled $N(\mu, \sigma)$ are fed to the decoder.

Validation on a Constructed Data Set

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

classifications. When a peptide was assigned to an archetype it was classified as amyloidogenic or non-amyloidogenic according to a probability. The second scenario captures errors in the experimental data in the database, or an amyloid mechanism that only partially relates to the chosen descriptors.

All models are trained on 500 peptides and validated with 500 different peptides. Peptides are described with two descriptors per amino acid. The axes for each plot in Figures 4.3 are the values in the latent space; that is the values output by the two nodes labeled μ in Figures 4.2. The y-axis is the positive prediction axis, and the x-axis is the negative prediction axis. A peptide is classified as positive if its positive prediction value is greater than its negative prediction value. Thus, if a peptide falls above the red line on the plots that peptide is predicted positive, while falling below the red line is a negative prediction. Figures 4.3 A, D, and E plot each peptide in the database according to where they fall in latent space. Figures 4.3 B and C show the reconstructed description of the peptide for equally spaced points in latent space. These types of plots will be referred as reconstruction plots.

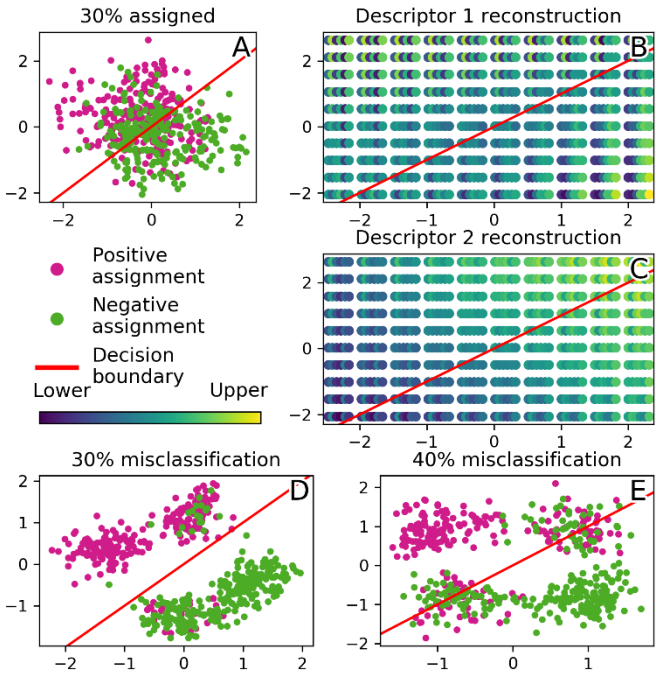

Figure 4.3 For all plots, the y-axis is the positive prediction axis, and the x-axis is the negative prediction axis. A peptide is predicted positive if it is falls above the red line and is predicted negative if it falls below the red line. Plot (A) shows where each peptide is encoded by the CAE, while plots (B) and (C) show the reconstructed description at each point in latent space for that same model. Plots (D) and (E) are each a separate model (see text) and show where each peptide in their database encodes to in latent space.

Figures 4.3 A-C are generated from the same model. Of the 1000 peptides 27% are assigned to an archetype, and subsequently classified as either positive or negative depending on the archetype. The positive archetype is LULULU (U = upper, L = lower) in descriptor 1 and LLLLLL or UUUUUU in descriptor 2. The negative archetype is LLLUUU for descriptor 1 and either LLLLLL or UUUUUU in descriptor 2. All peptides not assigned to an archetype were not within a threshold r-squared distance of any of these archetypes and were classified randomly.

In Figure 4.3A, 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 4.3 B and C, we visualize the reconstructed description of the peptide,

descriptors 1 and 2, respectively, in latent space. This representation of the peptide description arranged in latent space, the reconstruction plot, is the key to gaining intuition from the CAE, as it visualizes the different regions of positive and negative predictions that the CAE identified.

Figures 4.3 D and E depict different models than Figures 4.3 A-C. These use a database generated to mimic a set of experiments that yielded occasionally flawed results. In Figures 4.3 D and E all peptides in the database are within a threshold distance to one of four archetypes. For the two positive archetypes, one archetype was always classified positive, while the other archetype was misclassified at the rate indicated in the plot title. The negative archetypes were assigned in the same way.

These results show the models can simultaneously sort the data into the positive and negative classifications and identify the original archetypes used to generate the data. In the top left of Figure 4.3 A, the positive prediction region of the latent space, positive peptides have been separated from a mixture of positive and negative peptides, correctly predicting those peptides as positive. The corresponding region in the reconstruction plot, Figures 4.3 B and C, correctly reflects the positive archetypes. This happens similarly for the negative prediction region. We, also, learn how to interpret the reconstruction plot by examining Figures 4.3 B and C. The middle of Figures 4.3 A, the data's latent space distribution, shows mixed positive and negative peptides; in Figures 4.3 B and C, the reconstruction plot, this region shows no evidence of the positive or negative archetypes. However, as we move to the top left of the data's latent space distribution, Figures 4.3 A, we see a separation of positive classifications from the mixture of classifications; when we follow this trajectory in the reconstruction plot, Figures 4.3 B and C, the positive archetype emerges. The separation of a single class from a mixture of classes can tell us about the trend that contributed to that separation.

In Figure 4.3 D and E, the model correctly shows four clusters in the latent space, according to the four archetypes used to construct the database. The reconstruction plot (Figure S4.5) correctly reflects the four archetypes. This gives us insight to how the model deals with the uncertainty in the data. In Figure 4.3 D, the archetype which has been 80% classified positive and 20% classified negative is placed in the positive prediction region of the latent space. However, this is nearer the decision boundary (the red line) than the cluster associated with the 100% positive archetype, suggesting the model identified the ambiguous archetype. Further, in Figure 4.3 E, the ambiguous archetype was associated 60% to one classification and 40% to the other. In this case, the cluster that represents the ambiguous archetype is placed nearly atop the decision boundary, leaning slightly positive. The method can make identifications regarding how an archetype leans, in addition to characterizing archetypes that are certainly associated with activities.

We note our validations show our method works with large databases that are typically used in machine learning (N = 10,000; Fig. S7), but crucially also with the limited databases we have available for amyloid studies (N = 1000; as shown here). This suggests potential generalizability of the models to problems associated with relatively small databases, such as the Waltz database we use later [107].

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

CAE on an Experimental Database: Hydrophobicity

Metrics related to hydrophobicity were found to be the most effective descriptors, and such a metric is used in both descriptors examined here. In Figure 4.4 B (and later in Figure 4.5 B) we can see a region of peptides with yellow or green amino acids in the middle (positions 3 and 4) and dark green or blue amino acids on the ends. This means peptides in this region of the latent space tend to be hydrophobic in the middle, and more hydrophilic on the ends, suggesting this type of amyloid fibril buries the hydrophobic core by stacking while the hydrophilic ends on the outside interact with water. It should be noted in both cases much of this region is an extrapolation by the model (there are few data points in the latent space in these regions). While extrapolation must be taken with caution, this motif in this "most-likely amyloid" region is worth noting due to the intuitive sense that hydrophobic amino acids should be buried away from the solvent. This motivates further investigation on sequences capturing this motif. In other words, if a goal is to investigate the coarse forces driving amyloid formation or design new amyloid forming peptides, the CAE's extrapolation can be a hypothesis to pursue.

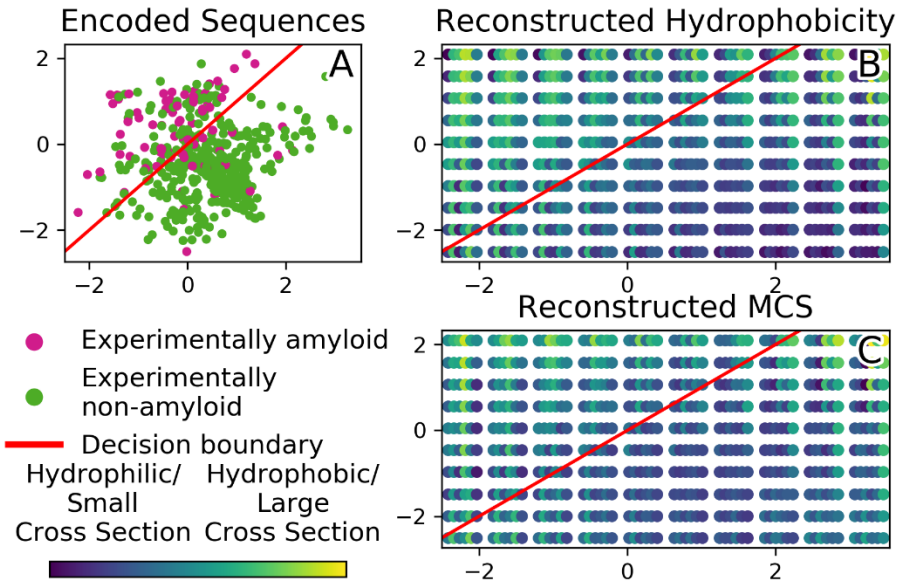

Figure 3.4 Representative model trained using hydrophobicity and monomer cross section (MCS). (A) All sequences in the validation set plotted in latent space. Each axis here is the value of one of the latent space nodes. The color of

the point represents the experimental classification of that peptide. (B) and (C) show the reconstructed descriptions. The axes here also represent values in the latent space, but the markers represent peptide descriptions. Each group of six dots represents a peptide, and the color of that dot represents the reconstructed descriptor value at that point in latent space. (B) depicts the hydrophobicity of the peptides, where yellow is hydrophobic, and blue is hydrophilic. (C) depicts the monomer cross section (MCS) of each point in latent space. Here yellow is a large cross section, and blue a small cross section.

There also exists some signs of the West et al result [89] of NPNPNP (P = hydrophobic (polar), N = hydrophilic (non-polar) within the core residues (2, 3, 4, 5) in both models. This patterning is also characterized by less extreme hydrophobicity, suggesting this motif is preferred by those residues with moderate hydrophobicity values. It is worth noting this region has been interpolated as there are many amyloid points in this region of the latent space; we can then be more confident that these motifs are well-represented within the database. Observations based on this interpolated region could also provide grounds to investigate forces driving amyloid formation or to inspire novel amyloid forming peptides.

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

CAE on an Experimental Database: Monomeric Cross Section

Figure 4.4 represents a model trained using Monomeric Cross Section (reported in Table 1) and hydrophobicity as the descriptors. The populated region on the amyloid side tends to include mid-to-large residues, while the populated region in the non-amyloid side tends to include small residues. This could suggest a preference for bulky side chains, perhaps to help drive amyloid stability through surface-area dependent forces such as van der Waals. Additionally, on the amyloid side, there is some alternation of large and small residues. We also note that hydrophobicity similarly alternates in the same region of latent space. Perhaps this alludes to a connection between the size of a side chain and its potential for stronger hydrophobic-related

forces resulting in a preference for sequences that alternate large, hydrophobic residues and small, hydrophilic residues[120–122].

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 3.1 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.

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 [110]. 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 [110]. As we collected more data on aggregation of amino acids, we found that this value was oligomer size dependent. We also found the monomer and dimer to be the only oligomer sizes that 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.

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

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

Here we probe the relationship between DID and amyloid propensity. For the most part, Figure 4.5 C shows few features in the amyloid region and the peptides are generally on the extended side of DID. The top left shows some signs of compact DID. This is also the same region where the hydrophobic core motif is represented. Like the monomer cross section result, here the hydrophobicity is likely the larger factor governing amyloid formation, as evidenced by the larger diversity of hydrophobicity motifs in the amyloid region. In the non-amyloid region, there exists a region of mixed amyloid and non-amyloid points (middle of the plots), as well as a region of pure non-amyloid points (the right of the plots), reminiscent to the pattern we saw in the distribution of points during the first validation experiment (Fig. 3 A). Within these regions the hydrophobicity motifs have relatively low diversity, being generally hydrophilic, while there is greater diversity in the DID motifs. Critically, as one moves deeper into the non-amyloid region, one observes a rise in the compactness of the residues. Thus, in the same way the model from Fig. 3 A determined the archetype in the pure green region, the CAE has determined a strong relationship between compactness and a failure to grow fibrils – the extrapolated "least amyloidogenic" peptides (those that would appear in the bottom right of Fig. 5 C) are most strongly

characterized by a higher degree of compactness, with less distinguishing features in hydrophobicity representation.

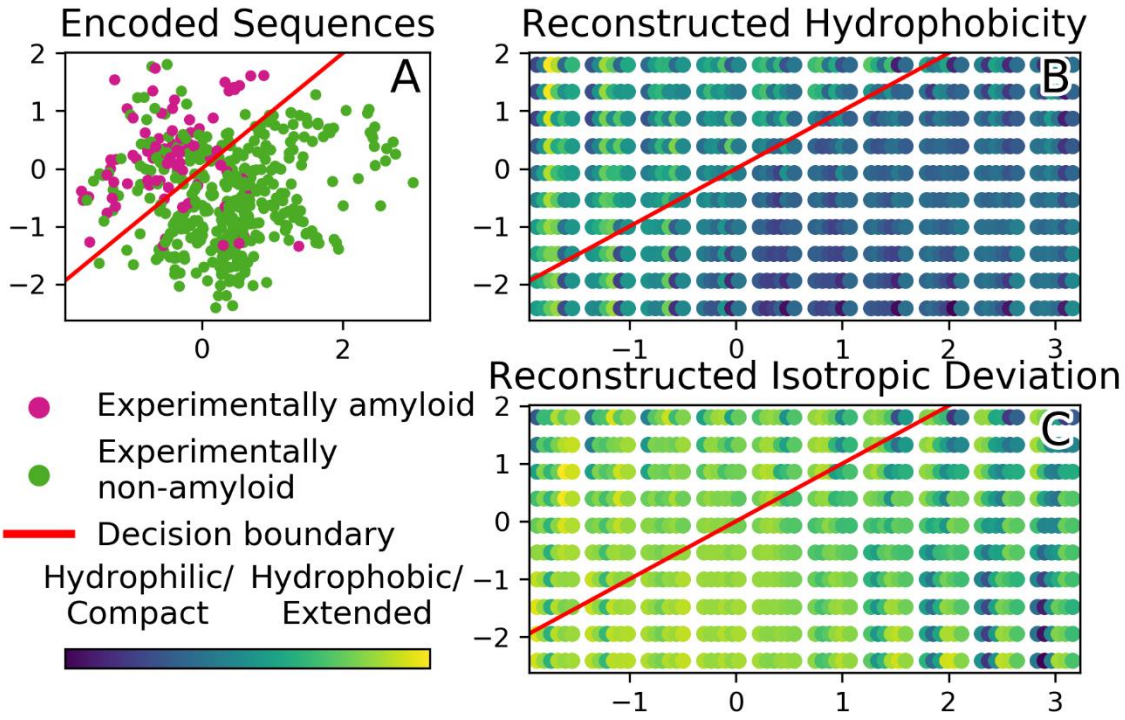

Figure 3.5 Representative model trained using hydrophobicity and DID. This figure is the same representation of a model as Fig. 4except (C) depicts the DID of each point in latent space. Here yellow is extended growth, and blue is compact growth.

These results provide potential insight about how DID relates to amyloid formation. Namely, compact growth of the amino acids could block the amyloid process of the peptide when hydrophobic interactions are not a significant driving force of amyloid formation. While it was not found that DID could be used by itself to attain reliable correlations with amyloid-forming behavior, likely due to the specificity of the interaction observed at the dimer level, the CAE determined that DID could be strongly related to a failure to form fibrils. Further, from this observation we may gain some insight about the differences between amyloid forming hexapeptides, and larger proteins. The residue with the most compact isotropic deviation is

glycine, and indeed the peptides in the non-amyloid forming/compact isotropic deviation region of the latent space are rich in glycine. This is a curious result since amyloids are often associated with glycine rich proteins as they tend to be intrinsically disordered [123,124]. Further, it has also recently been shown that glycine is an essential residue in cylindrin formation[125–127], 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.

4.4 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.

Chapter 5) Characterizing Epistasis of fRNA Fitness Landscapes Using Entropy

5.0 Forward

So far, the characterization of biopolymers has focused on polypeptides, either in the abstract, as in the coarse-grained models of Chapters 2 & 3, or in the immediate, as in the database of Chapter 4. Chapter 5 represents a shift away from polypeptides to focus on polynucleotides, specifically ribonucleic acid polymers, or RNA.

This is not outside the purview of this dissertation so far, as polynucleotides and polypeptides share key isomorphisms that make their modeling extremely similar in key regards. As described in Chapter 1, both categories of polymer can be described as a sequence of monomers drawn from a characteristic "alphabet". In human polypeptides, this alphabet contains the 20 amino acids. In polynucleotides, the alphabet contains the 4 nucleotides associated with either the RNA nucleobases or the DNA nucleobases.

Because of this abstract shared identity, both classes of polymers can be treated similarly at the level of information modeling. The methods of simulation applied to polypeptides in Ch. 1 & Ch. 2 could, and often are, be applied to chains of ribonucleic acids. Similarly, the process of mining a database for sequence motifs associated with a desired binary activity could be applied to polynucleotides, assuming that a broad enough database of active and inactive sequences can be acquired.

Polynucleotides tend to be viewed primarily under the paradigm of storing or transmitting genetic information. While this is a key function, RNA in particular is known to occasionally serve in a manner more similar to the protein machineries – that is, it will possess a chemical function, and can catalyze reactions with itself or other molecules. Such RNA is known as function RNA

(fRNA), and due to theories regarding the earliest biomolecules, investigation in fRNA abilities are implicated in elucidating the origins of terrestrial life.

An essential informatic structure regarding these explorations is the fitness landscape, which relates variations in a sequence of RNA, or genotype, to its ability to perform a specific activity. Advances in experimental methods has enabled the collection of large databases of information in the form of these landscapes, inviting a need to model these structures in terms that easily quantifies the factors most relevant to developing theories of molecular evolution.

As such, Chapter 5 applies a core tool of information theory to extract information from a fitness landscape model and represent it in terms of how sites along an fRNA sequence interact with one another, affecting the activity of interest. It represents the power of stepping away from seeing the biopolymer world as implicitly linked the physics of chemistry and seeing through the lens of pure statistics and information quantities. This is a fitting testament to the paradigm of information first modeling.

5.1 Introduction

The ‘fitness landscape’ is one of the determining factors in molecular evolution. These landscapes encode fundamentally the propensity of a given molecular sequence to possess a given activity. Knowledge of these landscapes is necessary for a fully understanding of these activities then, with collection and synthesis of the information they contain representing an important goal in advancing scientific understanding of functional RNAs. A predominant paradigm in the analysis of fitness landscapes is viewing evolution as a random walk along these landscapes[20], with much research being done on improving understanding of these topographical structures.

Substantial work had been done developing theoretical models of fitness landscapes to assist in their characterization, and have been covered in recent reviews[128,129]. These works tend to view evolution as a serial process in which an organism moves through some favored path toward a genetic optimum. Endemic to this paradigm is the question of how an genotype moves through this landscape, either by crossing low-fitness valleys[130] or exploiting the probabilistic implications of high dimensionality in sequence space to move monotonically toward higher fitness[131,132]. Many models of fitness landscapes exist, ranging from simple additive models to enhanced models capable of integrating the complexities of epistatic interaction[133–136]. Indeed, quantifying epistasis is such an essential element of characterizing the structure of fitness landscapes that many direct or indirect measures of epistasis, such the divergence from additivity, the ratio of roughness over additive fitness, the fraction of sign epistasis, Fourier spectral analysis and multiple other techniques have been explored[137].

With the realization of new methods in the massive collection of data across fitness landscapes[138–141] new means of quantifying their attributes have become accessible. Methods for approximating RNA secondary structure[142–146] from the primary sequence has been extremely useful in this regard however when considering the interpretation of the vast quantities of data available from contemporary methods, simulation techniques are limited by the need for researcher oversight in the examination of each potential sequence. This can easily become prohibitive.

In this paper, we use the paradigm that the fitness – here the activity of a self-aminoacylating piece of fRNA – can be viewed as a random variable whose samples can be drawn from controlled ensembles. We further choose to study the mutation space that is implicitly associated with any sequence space – we view the fitness space in terms of changes to the fRNA activity rather than focusing on the value of the activity for any given sequence. With this combination of viewpoints,

we can apply the tools of information theory to study two empirical fitness landscapes and quantify insights regarding interactions between residues and site-specific relevance of epistatic effects. This bears resemblance to work that considers epistasis in terms of statistical correlations of mutations[147]. This work adopts those mutational ensembles, but rather than focusing on correlations instead approaches the problem from the perspective of the entropy content of these associated distributions. This abstracts the problem to the level of information content, providing a highly general framework of analyzing these spaces in terms of variables of potential interest.

By adopting the information theoretic perspective of big data, this method will be amicable to the ever-growing capability to produce empirically realized fitness landscapes and rapidly characterize site-specific information from that data. This may enable the improvement of existing theoretical models, refinement of thinking on evolutionary processes, or else enable abstract comparison of many fitness landscapes in search of trends. Ultimately, it allows epistasis to be characterized at the level of the entire fitness landscape or to the degree of nucleotide epistatic interactions with other locations within the sequence.

To this end, we use relative entropy, a quantity associated with the training of machine learning algorithms such as decision trees. Concisely, the relative entropy quantifies the amount of uncertainty about a random variable when one can observe a variable from another potentially related variable. It is related to the Kullback-Leibler divergence[21,148], which has found a wide variety of applications over numerous fields[149–153], acting a critical metric in the development of artificial intelligence and refining machine learned models.

This work uses data gathered by a method introduced in a previous publication[154]. The SCAPE (sequencing to measure catalytic activity paired with in vitro evolution) combined method for sampling large swathes of a fitness landscape. The object of study is a RNA sequence 71

nucleotides in length, of the form 5'-CON1-VAR-CON2-3', where CON1 is a 26 nucleotide length sequence of constant nucleotides, VAR is a 21 length sequence of variable nucleotides and CON2 is a 24 length constant sequence of constant nucleotides. The prior work performed identified 5 families of ribozyme within this sequence space. The five families were individually capable of performing the key activity of covalently attaching aminoacyl groups. This reaction has particular significance due to it potential involvement in the formation of compounds that could have potentially functioned as a prebiotic chemically activated amino acid, capable of forming the earliest peptides[155]. The data is collected using a process called k-seq, which provides the activity of the ribozyme catalysts in terms of kA, where k is a rate constant (min^{-1} M^{-1}) and A is the maximum amplitude of reaction from the experiment, a unitless quantity. For the families of enzymes considered the parameters cannot be estimated separately, however their product provides viable insight into the catalytic rate of the polymer.

This work focuses on the analysis of a family of aminoacylating ribozymes referred to in previous work as S-1B.1-a (see prior work for nomenclature). The database of sequences mined for information represents an exhaustive coverage of the sequence space within two Hamming distance of a core seed sequence, which we call S1B.

In the following sections, we calculate the conditional entropy associated with knowing the identity of a target nucleotide when considering the effect of a mutation on a specific site of the variable region of the RNA polymer. We connect the values of this conditional entropy to significance in assessing the activity of the ribozyme, prescribing a quantitative metric for locating sites of interest along the biopolymer chain. This is, to the author's best knowledge, the first instance of applying conditional entropy to quantify structure-activity relationships and elucidate epistatic connections between activity and site-specific couplings using a systematic process of

considering ensembles of mutations within a fitness landscape. The hope is that the generalizability and deep connections to information content will yield a satisfying and robust approach to characterizing these sorts of landscapes, in a way that does not necessarily depend on any assumed properties of the distributions in question.

5.2 Methods

As stated in Chapter 1, the relative entropy of a random variable **X** when considered with a covarying conditioning random variable **A** is defined as follows:

$$RE_{XA} = D_{KL}(P_X(x,a)||P_X(a)) = \sum_{x \,\epsilon\, \boldsymbol{X}, a \,\epsilon\, \boldsymbol{A}} p(x,a) log_b \left(\frac{p(x,a)}{p(a)p(x)} \right)$$

Where p(x,a) is the joint probability distribution of variable **X** on conditioning variable **A,** and x, a are the values from the sampling ensemble. The logarithmic root b allows for a choice of units. In this work we set this value to 2, so that all information is in units of bits. The sum is taken over the support set of the probability mass functions associated with **X** and **A**. We note that in the case of unrelated variables that p(x,a) = p(x) p(a) and $RE_{XA} = 0$, reflecting that fact that knowledge of one does not inform the other.

We compute the relative entropy for a given site on the ribozyme as follows. The sequences are associated with values kA, which is treated as a variable **kA** that exists as samples from the 'random' space of the variable region's permutations. The probability mass function associated with this variable can be approximated by the generation of a histogram. Conditioning variables can then be chosen at will from the genotypes, and are broadly referred to in this work using the variable **M**. For example, if we are interested in the relative entropy of the first site along sequence then we can create a probability mass function for whether the residue is A (class one), U (class

two), C (class three), etc. This yields a quantity that relates the knowledge of the activity class to the genotype variable in terms of the amount of information one gains about the phenotype when one has knowledge of the genotype variable.

A pdf for kA is approximated by producing a histogram over the range of kA values for the ensemble, similar to the discretization process used to compute the free energy landscapes of Chapter 2 with the WHAM[43] method. The choice of bins in the histogram is treat as an arbitrary decision based on the observation it influences results significant only at extremes. For the purposes of this work, the histogram of kA was chosen to have 21 bins, with a maximum and minimum value defined by the complete mutation ensemble.

With the pdfs for Ka and M, the relative entropy for a given set of sites is calculated as follows:

$$RE_{\{\mathrm{n_i}\},\{\mathrm{M_i}\}} = \sum_{\mathrm{kA,M}} p(\mathrm{kA,M}) log_2 \left(\frac{p(\mathrm{kA,M})}{p(\mathrm{M}) p(kA)} \right)$$

See prior work for experimental details on how the fitness landscapes were collected. Secondary structures were calculated using mfold[156].

ε, Difference from Additivity

We compare relative entropy to the quantity ε as a means of building intuition and verifying the relationship to epistasis. ε quantifies the epistatic contribution as follows:

$$\varepsilon = \Delta kA_{qNp,lMk} - (\Delta kA_{qNp} + \Delta kA_{lMk})$$

Where q, p ,l and k are nucleotide identities; N and M are the sites where the mutation occurs; and Δ implies the change in the quantity associated with the subscripted mutation. The mutants are computed with respect to the seed sequence as the wildtype.

5.3 Results & Discussion

Figure 5.1 shows a similarity between the standard deviation of epsilon and the single-site entropy computed when considering the distribution of **kA** jointly with the nucleotide identity of site N along the variable region.

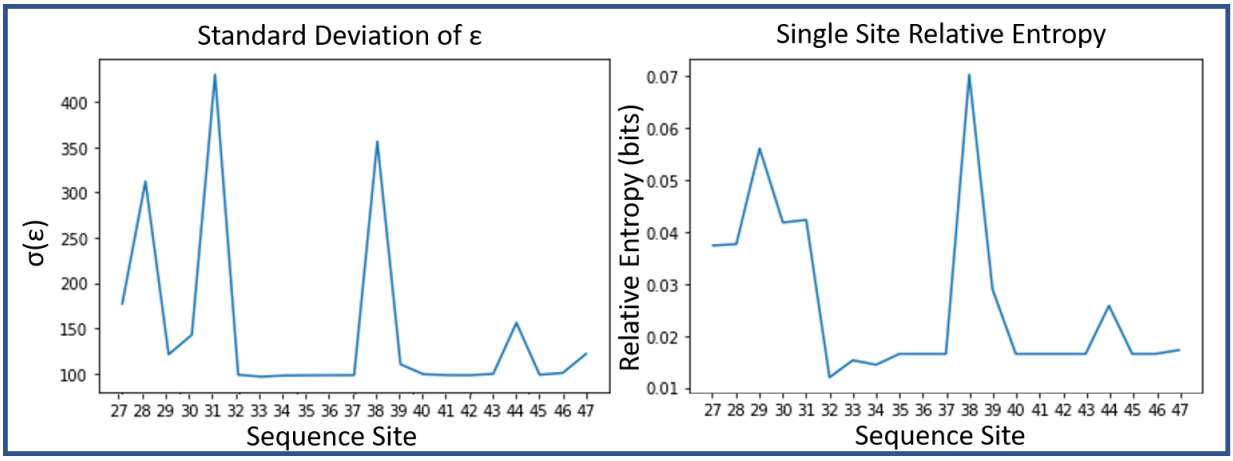

Figure 5.1 Two plots showing similarity between the standard deviation of ε, which quantifies epistasis as the difference between the actual change of a phenotype by a double mutant from the sum of the constituent single mutations, and the single site relative entropy. Both quantify the complexity of the mutations (STD(ε) as the spread of the differences and relative entropy as the information gain between knowing the identity of a site and the activity distribution)

We build intuition around the relative entropy quantity by observing the similarities. The standard deviation of ε captures the spread of the difference between the double mutation and additivity. We can consider Figure 5.2 to build insight into these quantities, in which we plot a representation of the mutations affecting sites from the data cluster. These plots show the effects of mutation of single residues against a variety of backgrounds, reflecting the changes due to these residues and their interactions with the rest of the system.

The mutations are color coded according to the wildtype nucleotide at the indicated position. For high-entropy sites like 29 and 38 the high relative entropy is characterized by a greater

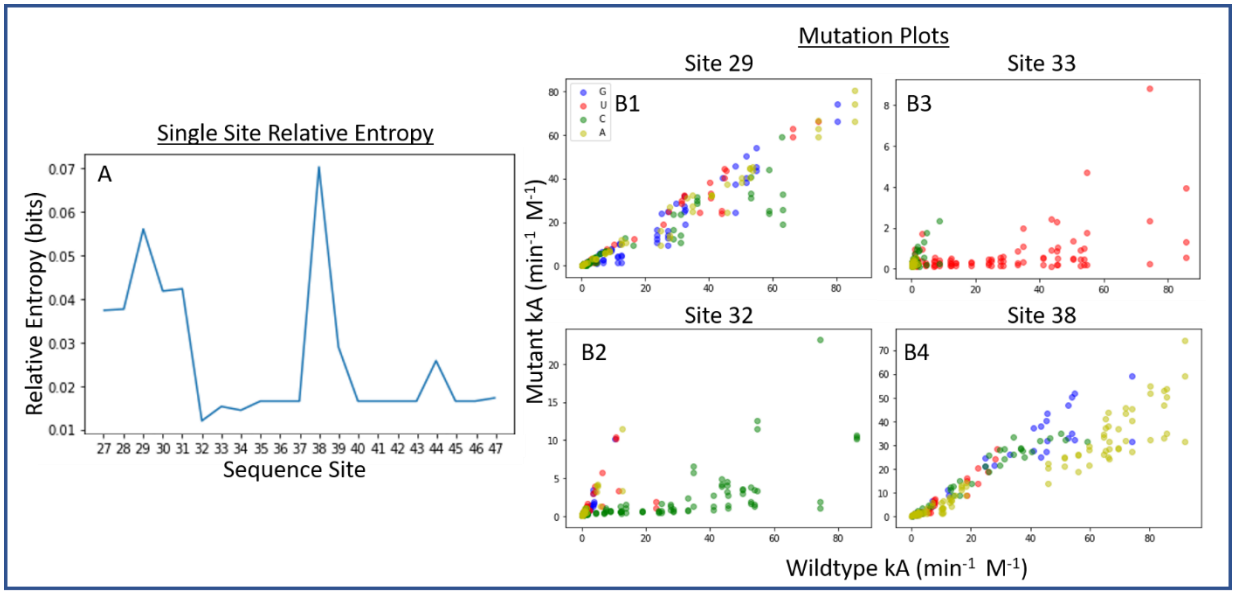

Figure 5.2 Figures showing the single-site relative entropies calculated between the distribution of activities and the genotype of specific residues. Color of a point represents the wildtype's nucleotide at that location, indicating the breakdown of genotype populations. Sites with higher relative entropy show more hierarchy, while sites with low relative entropy have distributions that contain relatively little information regarding the distribution of activities.

diversity of mutations. The information gains by knowing the identity of these residues better informs our understanding of the **kA** distribution due to the wide set of values the mutation stretches over. Conversely, locations with low relative entropy (32 and 33 in Figure 5.2) tend to have their mutations clustered in smaller areas, speaking to the fact the identity of these nucleotides do not have complex interactions with the background and relatively low epistatic character. These residues might be considered 'critical' sites. In the cases of 32 and 33 (Figure 5.2) we can observe that the wildtype only has appreciable activity in cases where a certain nucleotide occupies the position. Conversely, we see more complex coloration patterns for the high-RE sites. We can note that this property was not well-captured by the ε-derived quantity, which characterized site 29 as being more similar to site 32 than site 38. In terms of the effects of mutation on this residue when

considering the complexities of interaction with the genetic background, this characterization is manifestly untrue for the examined data.

The relative entropy provides reason to further investigate the importance of certain sites. Site 38 is shown to have the highest single site RE, suggesting its identity provides the most information regarding the kA distribution. Secondary structure predictions by mfold[156] helps interpret the significance of these residues. In Figure 5.3 we can observe that the predicted nucleophile in the acylation, G65, exists in flexible base pairing with C35 within the variable region.

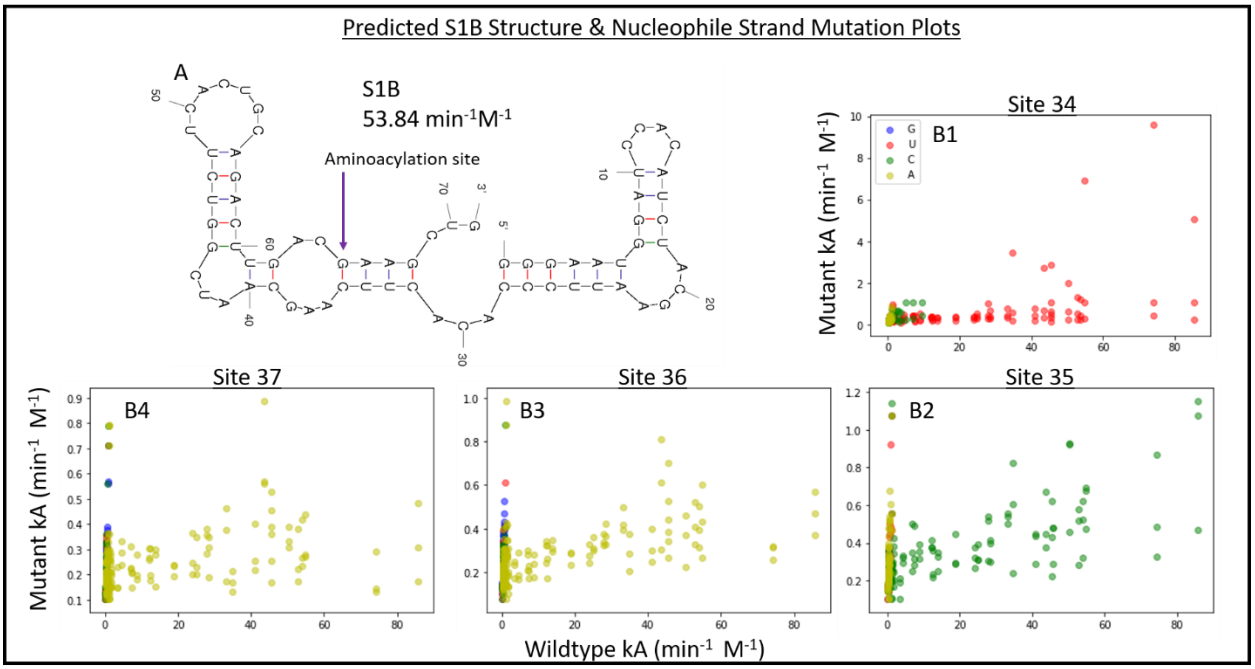

Figure 5.3 The predicted structure of seed S1B (kA = 53.84 min^{-1}M^{-1}) and several select mutation plots showing distributions characteristic of sites whose nucleotide identity is critical to ribozyme function. C35 (B2) is base paired to the nucleophile G65. Nearby residues have similarly strict identities, implying that the specific arrangement of C35 is important to the catalytic mechanism. Because this sites only confer information regarding whether or not a ribozyme is functional they possess low relative entropy compared to sites with more flexible identities, but whose interactions with the genetic background have sophisticated relation to the catalytic process.

Figures 5.3B show mutations of sites including and nearby C35. Unsurprisingly, mutation of C35 itself disrupts the function of the ribozyme entirely, suggesting that the base pairing itself is of critical importance. Sites 36 & 37 exhibit similar intolerance of variable residues. This may suggest that closing the loop that opens after C35 is similarly destructive. Interestingly, site 34

shows signs of tolerance toward mutation – function is severely impacted, but there exist sequences for which function unambiguously exists for C34.

The importance of the 36-38 loop is qualified by the observation that the highest functionality sequences are intimately tied to having A38 (Figure 5.2B4). However, the relationship of this location to the ribozyme function is the highest in the variable region as evidenced by the high relative entropy to the **kA** distribution. We can observe that all possible nucleotides are can possibly yield functional ribozymes. Interestingly, and predicted by the high relative entropy, these values are hierarchical within the distribution – adenine is preferred, then guanine, then cytosine and finally uracil. This existence of clear hierarchy, in addition to the wide span of potential values, is at the heart of the high relative entropy value. For the entire set, knowing the value of site 38 helps 'narrow down' what the activity will be. Knowing the values of critical sites like 35 or 36 will only provide information as to whether the ribozyme is functional or not and provides little information regarding the complex distribution within the subset of active ribozymes, thus despite the importance to function in a binary sense, the critical locales show little relative entropy and might best be detected using complementary analysis like conservation studies.

Knowing Site 38 prefers adenine for functionality strengthens the argument that the catalytic mechanism relies on the open loop 36-38. Figure 5.4 shows the predicted secondary structures for S1B and the three single mutants of site 38. Uracil results in the poorest functional ribozymes, likely due to the partial loop closure that is predicted to occur from the U38-A63 base pairing. C38's activity is comparable to S1B, possibly due to interactions with G62 creating a more stable loop and interfering with the ability of G65 to depart its base pairing to C34. A38 is unambiguously the functional ribozyme, likely because the adenine has no possible base pairings that would not substantially strain the rest of the structure and thus the loop is fixed in the apparently ideal state.

The region with high relative entropy at the beginning of the variable region is involved with Site 29 and is implied, based on relative entropy, to contain a high amount of information regarding

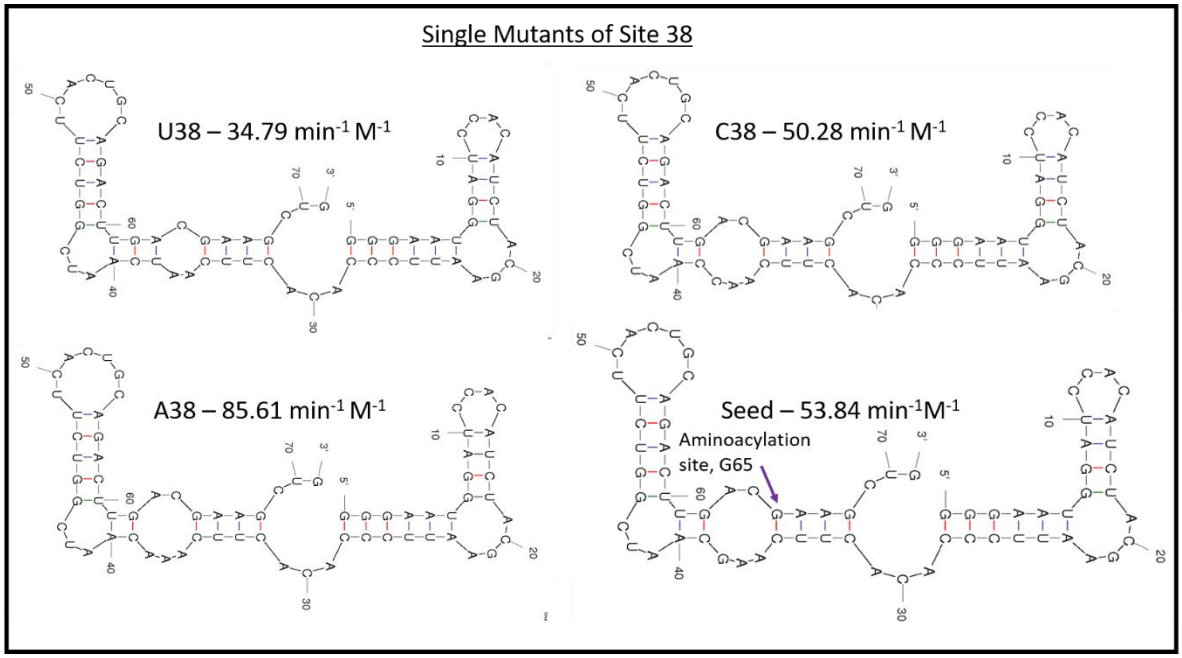

Figure 5.4: Predicted secondary structures for mutants of site 38, showing a drop in function when the loop after the active site is closed. A38, for which the loop must remain open due to a lack of base pairing opportunities for adenine, shows the highest activity. U38, which is predicted to exist in a stable base pair with A63, closes the loop and the sequence has reduced function.

the activity of the ribozyme. Secondary structure calculations show the value of the nucleotide affects the stability of the functional structure, suggesting the reason for its importance is the complex interactions this site and the surrounding region have with genetic background in

stabilizing the functional fold. This is how the area can impact the functionality of the ribozyme despite being distant from the catalytic site. These calculations are summarized in Table 5.1.

Table 5.1		
Ribozyme	Proper Fold Energy	Activity
S1B	-14.20 kcal/mol	53.84 $mol^{-1}M^{-1}$
G29	-14.20 kcal/mol	54.76 $mol^{-1}M^{-1}$
C29	-13.30 kcal/mol	43.66 $mol^{-1}M^{-1}$
U29	-13.70 kcal/mol	45.44 $mol^{-1}M^{-1}$

Table 5.1 Showing the predicted energy of the functional fold for site 29 mutants. The identity of the nucleotide interacts with the sequence to stabilize or destabilizing the active fold and affects the overall catalytic capability.

5.4 Conclusions

We show that the relative entropy between an approximated probability density function of activities for a given database of sequences and the identity of a nucleotide at a given location in the sequence can be used to quickly characterize the complexity of a nucleotide's interactions with the functional process. This is a means of quantifying epistasis, the phenomenon in which genes interact with potentially distant regions in a primary sequence to have non-additive effects.

The relative entropy was related to ε, the difference between an additive mutation and the true mutation. Analysis was performed to demonstrate the relative entropy's ability to characterize a system's interaction and provide guidance in deducing a possible mechanism.

Using relative entropy in combination with plots of mutations extracted from the network graph of the sequence space, we were able to establish the relative entropy's deep connections to clarifying which locations in the primary sequence provide the most information regarding the phenotype distribution. Critical sites – sites essential for ribozyme function – contained low

information. While superficially counter-intuitive, these regions only contain information as to whether or not a sequence is function and are thus less informative than knowledge about sites with more complex interactions with functionality. Sites with high relative entropy were implicated through consideration of mutations and secondary structure prediction with more sophisticated effects on the catalytic capability, either through direct interaction with the catalytic site or by globally affecting the stability of the functional fold.

Thus, relative entropy in combination with this sort of data can help detect locations in the primary sequence that mathematically provide the most guidance for forming a hierarchical understanding of subset of functional RNA. Paired with complementary analysis methods, this can be used to build insight into the mechanistic process.

Because relative entropy is a foundational quantity in information theory, it is intimately related with machine learning algorithms, such as the formation of decisions trees. Thus, this research acts to help perform fundamental analysis using the quantity in hopes that these insights can be used to develop better machine intelligence algorithms for characterizing ribozyme behavior using the sorts of large datasets now available with high throughput experimentation.

Broadly, relative entropy can be used to interrogate the information content of genetic variables with respect to their impact on activities. Because it is so fundamental, it is an extremely general and powerful approach that can be applied to any situation in which ample data is available relating primary sequence and an observable of interest. Combined with existing techniques such as conservation analysis and secondary structure prediction, the relative entropy can indicate regions in the gene that are mathematically most pertinent to informing our understanding of the distribution of activity within the sequence space.

Chapter 6) Conclusions

The techniques present in this dissertation are promising and novel means of elucidating biological heteropolymers using the sorts of data that have become possible to generate in modern times. Molecular dynamics and information-based machine intelligence approaches provide a complementary framework for investigating the behavior and characterizing the dynamics of polymer systems.

In Chapter 2, work was performed in which we established the utility of coarse grained molecular dynamics in assessing the general behavior of heteropolymer models with the intent of validating an assumption that has been used to make more efficient calculations regarding the behavior of proteins in solvated systems. In broader scheme of this dissertation it represented work showing the importance of validation when making approximations of complex behavior. We concluded a simpler scheme that the analyte approach was in certain cases stronger, and the analyte model of interaction was vulnerable to significant error. This provided physical insights into the behavior of bulk solutions of molecules versus simplified classical models in which the behavior of such systems are embedded into solutes at the level of pairwise interactions in their Hamiltonians. To capture the proper physics, pairwise terms displayed poor additive effects and consequently affected the predictions of the simulations.

In Chapter 3, a variational autoencoder was used to reduce the representation of a large volume of data describing the assembly of peptide fibril structures using a coarse-grained model. This process allowed for the unsupervised discovery of an ‘order parameter’ for the system, qualitatively capturing the process by which a disordered system approached its favored ordered state. Such a method is helpful as larger volumes of data are generated using molecular dynamics,

as the automatic discovery and representation of broad phases or states is important for the proper analysis of these trajectories. The work in Chapter 3 showed a common form of neural network was compatible with these sorts of data sets, and that the ANNs could be used to extract insight from highly dimensional molecular dynamics trajectories on a large scale.

Chapter 4 modified the neural network architecture used in Chapter 3 by coupling a variational autoencoder to a neural network classifier and forming a latent representation of amyloidogenic peptides from an experimental database of hexamers. This process allowed for the rapid visualization of characteristic motifs that can contribute to amyloid formation and enabled the investigation of novel features for their potential pertinence to amyloid activity. These motifs included the alternating pattern of hydrophobic and hydrophilic residues and hydrophobic core peptides that have been associated with fibril formation. Additionally, an experimental quantity called the dimeric isotropic deviation was analyzed using the classifying autoencoder for its potential relationship to amyloidogenic properties. It was found to be implicated in the blockage of fibril formation when monomers in the peptide are associated with spherical growth in their homogenous dimers. The classifying autoencoder presented an intriguing and useful way of interrogating the relationships between primary sequence patterns and amyloid forming behavior in solution.

Finally, Chapter 5 focused on the application of essential information theory to quantify and characterize the epistatic character of site-specific nucleotides in the behavior of a functional ribozyme. Predictions backed by experimental evidence and secondary structure calculations verified the connection between the quantity calculated, relative entropy, and the complexity of a site's interactions with the catalytic process of a self-acylating fRNA. The relative entropy used here is a quantity involved in the formation of decisions trees, and it is the hope that these

investigations will aid in the development of more advanced machine learning techniques for characterizing the behavior of fRNA fitness landscapes. At a more immediate level, the relative entropy was used to identify loci in the fRNA primary sequence that were associated with complex interactions in the catalytic mechanism, such as the significance of a loop motif proximate to the catalytic site and the importance of a distant stem region seemingly implicated in the stability of the active fold. These sorts of datasets paired with entropy analysis can produce high priority targets in the primary sequence for direct experimental interrogation and could enable more rapid characterization of the nuanced interactions regions with higher epistatic behavior have with their relevant phenotypes.

We conclude by remarking on the importance of creative and innovative application of novel modeling strategies as we move forward. While all models may indeed be wrong, contemporary methods in data collection and analysis have enable us to make them as right as possible. With ever-increasing experimental and computational capabilities, using diverse and complementary methods from simulation through information theory will be necessary to fully unlock the potential of understanding. Critical to this is modeling, either through first-principles approaches such as molecular dynamics or through data-down perspectives like those in neural network approaches. The availability of large datasets will increasingly improve our ability to develop and validate our models, while the ambition of gaining insight into larger and more complicated systems requires clever approximation and thorough validation. It is our hope that this work serves as a useful body for the purposes of improving the use of some of the explored methods in the field of biological chemistry, and that with continued effort biochemical models and analytic techniques will continue to grow and improve.

References

(1) Box, G. E. P. Science and Statistics. *J. Am. Stat. Assoc.* **1976**, *71* (356), 791–799. https://doi.org/10.2307/2286841.

(2) Boyle, J. Lehninger Principles of Biochemistry (4th Ed.): Nelson, D., and Cox, M. *Biochem. Mol. Biol. Educ.* **2005**, *33* (1), 74–75. https://doi.org/10.1002/bmb.2005.494033010419.

(3) Sipski, M. L.; Wagner, T. E. Probing DNA Quaternary Ordering with Circular Dichroism Spectroscopy: Studies of Equine Sperm Chromosomal Fibers. *Biopolymers* **1977**, *16* (3), 573–582. https://doi.org/10.1002/bip.1977.360160308.

(4) Noller, H. F. Structure of Ribosomal Rna. *Annu. Rev. Biochem.* **1984**, *53* (1), 119–162. https://doi.org/10.1146/annurev.bi.53.070184.001003.

(5) Wu, Y.; Wang, F.; Shen, C.; Peng, W.; Li, D.; Zhao, C.; Li, Z.; Li, S.; Bi, Y.; Yang, Y.; Gong, Y.; Xiao, H.; Fan, Z.; Tan, S.; Wu, G.; Tan, W.; Lu, X.; Fan, C.; Wang, Q.; Liu, Y.; Zhang, C.; Qi, J.; Gao, G. F.; Gao, F.; Liu, L. A Noncompeting Pair of Human Neutralizing Antibodies Block COVID-19 Virus Binding to Its Receptor ACE2. *Science* **2020**, *368* (6496), 1274–1278. https://doi.org/10.1126/science.abc2241.

(6) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(7) Schultz, E. P.; Vasquez, E. E.; Scott, W. G. Structural and Catalytic Effects of an Invariant Purine Substitution in the Hammerhead Ribozyme: Implications for the Mechanism of Acid–Base Catalysis. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70* (9), 2256–2263. https://doi.org/10.1107/S1399004714010608.

(8) Design Patterns: Elements of Reusable Object-Oriented Software [Book] https://www.oreilly.com/library/view/design-patterns-elements/0201633612/ (accessed Jun 18, 2020).

(9) Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodríguez-Ropero, F.; van der Vegt, N. F. A. Systematic Coarse-Graining Methods for Soft Matter Simulations – a Review. *Soft Matter* **2013**, *9* (7), 2108–2119. https://doi.org/10.1039/C2SM27201F.

(10) Giuseppe Milano*, † and; Müller-Plathe‡, F. Mapping Atomistic Simulations to Mesoscopic Models: A Systematic Coarse-Graining Procedure for Vinyl Polymer Chains https://pubs.acs.org/doi/pdf/10.1021/jp0523571 (accessed Jun 15, 2020). https://doi.org/10.1021/jp0523571.

(11) Fritz, D.; Harmandaris, V. A.; Kremer, K.; van der Vegt, N. F. A. Coarse-Grained Polymer Melts Based on Isolated Atomistic Chains: Simulation of Polystyrene of Different Tacticities. *Macromolecules* **2009**, *42* (19), 7579–7588. https://doi.org/10.1021/ma901242h.

(12) Molinero, V.; Moore, E. B. Water Modeled As an Intermediate Element between Carbon and Silicon. *J. Phys. Chem. B* **2009**, *113* (13), 4008–4016. https://doi.org/10.1021/jp805227c.

(13) Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? *Curr. Opin. Struct. Biol.* **2008**, *18* (1), 10–15. https://doi.org/10.1016/j.sbi.2007.10.005.

(14) Schmidt, M.; Pfetzer, N.; Schwab, M.; Strauss, I.; Kämmerer, U. Effects of a Ketogenic Diet on the Quality of Life in 16 Patients with Advanced Cancer: A Pilot Trial. *Nutr. Metab.* **2011**, *8*, 54. https://doi.org/10.1186/1743-7075-8-54.

(15) Cooke, I. R.; Kremer, K.; Deserno, M. Tunable Generic Model for Fluid Bilayer Membranes. *Phys. Rev. E* **2005**, *72* (1), 011506. https://doi.org/10.1103/PhysRevE.72.011506.

(16) Drouffe, J. M.; Maggs, A. C.; Leibler, S. Computer Simulations of Self-Assembled Membranes. *Science* **1991**, *254* (5036), 1353–1356. https://doi.org/10.1126/science.1962193.
(17) Wu, C.; Shea, J.-E. Coarse-Grained Models for Protein Aggregation. *Curr. Opin. Struct. Biol.* **2011**, *21* (2), 209–220. https://doi.org/10.1016/j.sbi.2011.02.002.
(18) Chan, H. S.; Dill, K. A. Origins of Structure in Globular Proteins. *Proc. Natl. Acad. Sci.* **1990**, *87* (16), 6388–6392. https://doi.org/10.1073/pnas.87.16.6388.
(19) Bellesia, G.; Shea, J.-E. Diversity of Kinetic Pathways in Amyloid Fibril Formation. *J. Chem. Phys.* **2009**, *131* (11), 111102. https://doi.org/10.1063/1.3216103.
(20) Blanco, C.; Janzen, E.; Pressman, A.; Saha, R.; Chen, I. A. Molecular Fitness Landscapes from High-Coverage Sequence Profiling. *Annu. Rev. Biophys.* **2019**, *48*, 1–18. https://doi.org/10.1146/annurev-biophys-052118-115333.
(21) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22* (1), 79–86. https://doi.org/10.1214/aoms/1177729694.
(22) Hornik, K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Netw.* **1991**, *4* (2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T.
(23) Song, B.; Charest, N.; Morriss-Andrews, H. A.; Molinero, V.; Shea, J.-E. Systematic Derivation of Implicit Solvent Models for the Study of Polymer Collapse. *J. Comput. Chem.* **2017**, *38* (16), 1353–1361. https://doi.org/10.1002/jcc.24754.
(24) Larini, L.; Lu, L.; Voth, G. A. The Multiscale Coarse-Graining Method. VI. Implementation of Three-Body Coarse-Grained Potentials. *J. Chem. Phys.* **2010**, *132* (16), 164107. https://doi.org/10.1063/1.3394863.
(25) Chaimovich, A.; Shell, M. S. Anomalous Waterlike Behavior in Spherically-Symmetric Water Models Optimized with the Relative Entropy. *Phys. Chem. Chem. Phys.* **2009**, *11* (12), 1901–1915. https://doi.org/10.1039/B818512C.
(26) Sanyal, T.; Shell, M. S. Coarse-Grained Models Using Local-Density Potentials Optimized with the Relative Entropy: Application to Implicit Solvation. *J. Chem. Phys.* **2016**, *145* (3), 034109. https://doi.org/10.1063/1.4958629.
(27) Bizjak, A.; Urbi, T.; Vlachy, V.; Dill, K. The Three-Dimensional "Mercedes Benz: Model of Water. *Acta Chim Slov* **2007**, No. 54, 532–537.
(28) Wu, Z.; Cui, Q.; Yethiraj, A. A New Coarse-Grained Model for Water: The Importance of Electrostatic Interactions. *J. Phys. Chem. B* **2010**, *114* (32), 10524–10529. https://doi.org/10.1021/jp1019763.
(29) Song, B.; Molinero, V. Thermodynamic and Structural Signatures of Water-Driven Methane-Methane Attraction in Coarse-Grained MW Water. *J. Chem. Phys.* **2013**, *139* (5), 054511. https://doi.org/10.1063/1.4816005.
(30) Nguyen, A. H.; Molinero, V. Cross-Nucleation between Clathrate Hydrate Polymorphs: Assessing the Role of Stability, Growth Rate, and Structure Matching. *J. Chem. Phys.* **2014**, *140* (8), 084506. https://doi.org/10.1063/1.4866143.
(31) Ferguson, A. L.; Debenedetti, P. G.; Panagiotopoulos, A. Z. Solubility and Molecular Conformations of N-Alkane Chains in Water. *J. Phys. Chem. B* **2009**, *113* (18), 6405–6414. https://doi.org/10.1021/jp811229q.
(32) Athawale, M. V.; Goel, G.; Ghosh, T.; Truskett, T. M.; Garde, S. Effects of Lengthscales and Attractions on the Collapse of Hydrophobic Polymers in Water. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (3), 733–738. https://doi.org/10.1073/pnas.0605139104.
(33) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102* (14), 2569–2577. https://doi.org/10.1021/jp972543+.

(34) Jacobson, L. C.; Molinero, V. A Methane−Water Model for Coarse-Grained Simulations of Solutions and Clathrate Hydrates. *J. Phys. Chem. B* **2010**, *114* (21), 7302–7311. https://doi.org/10.1021/jp1013576.
(35) Nguyen, A. H.; Molinero, V. Identification of Clathrate Hydrates, Hexagonal Ice, Cubic Ice, and Liquid Water in Simulations: The CHILL+ Algorithm. *J. Phys. Chem. B* **2015**, *119* (29), 9369–9376. https://doi.org/10.1021/jp510289t.
(36) Song, B.; Nguyen, A. H.; Molinero, V. Can Guest Occupancy in Binary Clathrate Hydrates Be Tuned through Control of the Growth Temperature? *J. Phys. Chem. C* **2014**, *118* (40), 23022–23031. https://doi.org/10.1021/jp504852k.
(37) Le, L.; Molinero, V. Nanophase Segregation in Supercooled Aqueous Solutions and Their Glasses Driven by the Polyamorphism of Water. *J. Phys. Chem. A* **2011**, *115* (23), 5900–5907. https://doi.org/10.1021/jp1102065.
(38) Bullock, G.; Molinero, V. Low-Density Liquid Water Is the Mother of Ice: On the Relation between Mesostructure, Thermodynamics and Ice Crystallization in Solutions. *Faraday Discuss.* **2014**, *167* (0), 371–388. https://doi.org/10.1039/C3FD00085K.
(39) Hudait, A.; Molinero, V. Ice Crystallization in Ultrafine Water–Salt Aerosols: Nucleation, Ice-Solution Equilibrium, and Internal Structure. *J. Am. Chem. Soc.* **2014**, *136* (22), 8081–8093. https://doi.org/10.1021/ja503311r.
(40) Perez Sirkin, Y. A.; Factorovich, M. H.; Molinero, V.; Scherlis, D. A. Vapor Pressure of Aqueous Solutions of Electrolytes Reproduced with Coarse-Grained Models without Electrostatics. *J. Chem. Theory Comput.* **2016**, *12* (6), 2942–2949. https://doi.org/10.1021/acs.jctc.6b00291.
(41) Xu, L.; Molinero, V. Liquid−Vapor Oscillations of Water Nanoconfined between Hydrophobic Disks: Thermodynamics and Kinetics. *J. Phys. Chem. B* **2010**, *114* (21), 7320–7328. https://doi.org/10.1021/jp102443m.
(42) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.
(43) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021. https://doi.org/10.1002/jcc.540130812.
(44) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1–19. https://doi.org/10.1006/jcph.1995.1039.
(45) Hummer, G.; Garde, S.; Garcia, A. E.; Pohorille, A.; Pratt, L. R. An Information Theory Model of Hydrophobic Interactions. *Proc. Natl. Acad. Sci.* **1996**, *93* (17), 8951–8955. https://doi.org/10.1073/pnas.93.17.8951.
(46) Liu, Z.; Chan, H. S. Solvation and Desolvation Effects in Protein Folding: Native Flexibility, Kinetic Cooperativity and Enthalpic Barriers under Isostability Conditions. *Phys. Biol.* **2005**, *2* (4), S75-85. https://doi.org/10.1088/1478-3975/2/4/S01.
(47) Graziano, G. On the Solubility of Long N-Alkanes in Water at Room Temperature. *Chem. Phys. Lett. - CHEM PHYS LETT* **2011**, *511*, 262–265. https://doi.org/10.1016/j.cplett.2011.06.034.
(48) Underwood, R.; Tomlinson-Phillips, J.; Ben-Amotz, D. Are Long-Chain Alkanes Hydrophilic? *J. Phys. Chem. B* **2010**, *114* (26), 8646–8651. https://doi.org/10.1021/jp912089q.
(49) Cheung, M. S.; García, A. E.; Onuchic, J. N. Protein Folding Mediated by Solvation: Water Expulsion and Formation of the Hydrophobic Core Occur after the Structural Collapse. *Proc. Natl. Acad. Sci.* **2002**, *99* (2), 685–690. https://doi.org/10.1073/pnas.022387699.
(50) Matysiak, S.; Das, P. Effects of Sequence and Solvation on the Temperature-Pressure Conformational Landscape of Proteinlike Heteropolymers. *Phys. Rev. Lett.* **2013**, *111* (5), 058103. https://doi.org/10.1103/PhysRevLett.111.058103.

(51) Chen, T.; Chan, H. S. Native Contact Density and Nonnative Hydrophobic Effects in the Folding of Bacterial Immunity Proteins. *PLOS Comput. Biol.* **2015**, *11* (5), e1004260. https://doi.org/10.1371/journal.pcbi.1004260.

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

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

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

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

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

(57) Fitzpatrick, A. W. P.; Debelouchina, G. T.; Bayro, M. J.; Clare, D. K.; Caporini, M. A.; Bajaj, V. S.; Jaroniec, C. P.; Wang, L.; Ladizhansky, V.; Müller, S. A.; MacPhee, C. E.; Waudby, C. A.; Mott, H. R.; Simone, A. D.; Knowles, T. P. J.; Saibil, H. R.; Vendruscolo, M.; Orlova, E. V.; Griffin, R. G.; Dobson, C. M. Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril. *Proc. Natl. Acad. Sci.* **2013**, *110* (14), 5468–5473. https://doi.org/10.1073/pnas.1219476110.

(58) Geddes, A. J.; Parker, K. D.; Atkins, E. D. T.; Beighton, E. "Cross-β" Conformation in Proteins. *J. Mol. Biol.* **1968**, *32* (2), 343–358. https://doi.org/10.1016/0022-2836(68)90014-4.

(59) Guenther, E. L.; Ge, P.; Trinh, H.; Sawaya, M. R.; Cascio, D.; Boyer, D. R.; Gonen, T.; Zhou, Z. H.; Eisenberg, D. S. Atomic-Level Evidence for Packing and Positional Amyloid Polymorphism by Segment from TDP-43 RRM2. *Nat. Struct. Mol. Biol.* **2018**, *25* (4), 311–319. https://doi.org/10.1038/s41594-018-0045-5.

(60) Niu, L.; Liu, L.; Xi, W.; Han, Q.; Li, Q.; Yu, Y.; Huang, Q.; Qu, F.; Xu, M.; Li, Y.; Du, H.; Yang, R.; Cramer, J.; Gothelf, K. V.; Dong, M.; Besenbacher, F.; Zeng, Q.; Wang, C.; Wei, G.; Yang, Y. Synergistic Inhibitory Effect of Peptide–Organic Coassemblies on Amyloid Aggregation. *ACS Nano* **2016**, *10* (4), 4143–4153. https://doi.org/10.1021/acsnano.5b07396.

(61) Zou, Y.; Qian, Z.; Chen, Y.; Qian, H.; Wei, G.; Zhang, Q. Norepinephrine Inhibits Alzheimer's Amyloid-β Peptide Aggregation and Destabilizes Amyloid-β Protofibrils: A Molecular Dynamics Simulation Study. *ACS Chem. Neurosci.* **2019**, *10* (3), 1585–1594. https://doi.org/10.1021/acschemneuro.8b00537.

(62) Jin, Y.; Sun, Y.; Lei, J.; Wei, G. Dihydrochalcone Molecules Destabilize Alzheimer's Amyloid-β Protofibrils through Binding to the Protofibril Cavity. *Phys. Chem. Chem. Phys.* **2018**, *20* (25), 17208–17217. https://doi.org/10.1039/C8CP01631C.

(63) Straub, J. E.; Thirumalai, D. Membrane–Protein Interactions Are Key to Understanding Amyloid Formation. *J. Phys. Chem. Lett.* **2014**, *5* (3), 633–635. https://doi.org/10.1021/jz500054d.

(64) Martinez, A. V.; Małolepsza, E.; Rivera, E.; Lu, Q.; Straub, J. E. Exploring the Role of Hydration and Confinement in the Aggregation of Amyloidogenic Peptides Aβ16−22 and Sup357−13 in AOT Reverse Micelles. *J. Chem. Phys.* **2014**, *141* (22), 22D530. https://doi.org/10.1063/1.4902550.

(65) Zhuravlev, P. I.; Reddy, G.; Straub, J. E.; Thirumalai, D. Propensity to Form Amyloid Fibrils Is Encoded as Excitations in the Free Energy Landscape of Monomeric Proteins. *J. Mol. Biol.* **2014**, *426* (14), 2653–2666. https://doi.org/10.1016/j.jmb.2014.05.007.
(66) Reddy, G.; Straub, J. E.; Thirumalai, D. Dynamics of Locking of Peptides onto Growing Amyloid Fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (29), 11948–11953. https://doi.org/10.1073/pnas.0902473106.
(67) Thirumalai, D.; Reddy, G.; Straub, J. E. Role of Water in Protein Aggregation and Amyloid Polymorphism. *Acc. Chem. Res.* **2012**, *45* (1), 83–92. https://doi.org/10.1021/ar2000869.
(68) Carballo-Pacheco, M.; Strodel, B. Advances in the Simulation of Protein Aggregation at the Atomistic Scale. *J. Phys. Chem. B* **2016**, *120* (12), 2991–2999. https://doi.org/10.1021/acs.jpcb.6b00059.
(69) Nguyen, H. L.; Krupa, P.; Hai, N. M.; Linh, H. Q.; Li, M. S. Structure and Physicochemical Properties of the Aβ42 Tetramer: Multiscale Molecular Dynamics Simulations. *J. Phys. Chem. B* **2019**, *123* (34), 7253–7269. https://doi.org/10.1021/acs.jpcb.9b04208.
(70) Ilie, I. M.; Caflisch, A. Simulation Studies of Amyloidogenic Polypeptides and Their Aggregates. *Chem. Rev.* **2019**, *119* (12), 6956–6993. https://doi.org/10.1021/acs.chemrev.8b00731.
(71) Chiricotto, M.; Melchionna, S.; Derreumaux, P.; Sterpone, F. Multiscale Aggregation of the Amyloid Aβ 16–22 Peptide: From Disordered Coagulation and Lateral Branching to Amorphous Prefibrils. *J. Phys. Chem. Lett.* **2019**, *10* (7), 1594–1599. https://doi.org/10.1021/acs.jpclett.9b00423.
(72) Chen, M.; Schafer, N. P.; Wolynes, P. G. Surveying the Energy Landscapes of Aβ Fibril Polymorphism. *J. Phys. Chem. B* **2018**, *122* (49), 11414–11430. https://doi.org/10.1021/acs.jpcb.8b07364.
(73) Rojas, A. V.; Maisuradze, G. G.; Scheraga, H. A. Dependence of the Formation of Tau and Aβ Peptide Mixed Aggregates on the Secondary Structure of the N-Terminal Region of Aβ. *J. Phys. Chem. B* **2018**, *122* (28), 7049–7056. https://doi.org/10.1021/acs.jpcb.8b04647.
(74) Morriss-Andrews, A.; Shea, J.-E. Simulations of Protein Aggregation: Insights from Atomistic and Coarse-Grained Models. *J. Phys. Chem. Lett.* **2014**, *5* (11), 1899–1908. https://doi.org/10.1021/jz5006847.
(75) Baul, U.; Chakraborty, D.; Mugnai, M. L.; Straub, J. E.; Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2019**, *123* (16), 3462–3474. https://doi.org/10.1021/acs.jpcb.9b02575.
(76) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *arXiv:1312.6114*; 2013.
(77) Wetzel, S. J. Unsupervised Learning of Phase Transitions: From Principal Component Analysis to Variational Autoencoders. *Phys. Rev. E* **2017**, *96* (2), 022140. https://doi.org/10.1103/PhysRevE.96.022140.
(78) Wang, Y.; Lamim Ribeiro, J. M.; Tiwary, P. Machine Learning Approaches for Analyzing and Enhancing Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139–145. https://doi.org/10.1016/j.sbi.2019.12.016.
(79) Ribeiro, J. M. L.; Collado, P. B.; Wang, Y.; Tiwary, P. Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE). *ArXiv180203420 Cond-Mat Physicsphysics* **2018**.
(80) Ding, X.; Zou, Z.; Brooks III, C. L. Deciphering Protein Evolution and Fitness Landscapes with Latent Space Models. *Nat. Commun.* **2019**, *10* (1), 5644. https://doi.org/10.1038/s41467-019-13633-0.
(81) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J. Chem. Phys.* **2018**, *148* (24), 241703. https://doi.org/10.1063/1.5011399.

(82) Morriss-Andrews, A.; Bellesia, G.; Shea, J.-E. β-Sheet Propensity Controls the Kinetic Pathways and Morphologies of Seeded Peptide Aggregation. *J. Chem. Phys.* **2012**, *137* (14), 145104. https://doi.org/10.1063/1.4755748.
(83) Bellesia, G.; Shea, J.-E. Diversity of Kinetic Pathways in Amyloid Fibril Formation. *J. Chem. Phys.* **2009**, *131* (11), 111102. https://doi.org/10.1063/1.3216103.
(84) Stephen, M. J.; Straley, J. P. Physics of Liquid Crystals. *Rev. Mod. Phys.* **1974**, *46* (4), 617–704. https://doi.org/10.1103/RevModPhys.46.617.
(85) Saupe, A. Recent Results in the Field of Liquid Crystals. *Angew. Chem. Int. Ed. Engl.* **1968**, *7* (2), 97–112. https://doi.org/10.1002/anie.196800971.
(86) Eppenga, R.; Frenkel, D. Monte Carlo Study of the Isotropic and Nematic Phases of Infinitely Thin Hard Platelets. *Mol. Phys.* **1984**, *52* (6), 1303–1334. https://doi.org/10.1080/00268978400101951.
(87) Ray, S.; Holden, S.; Martin, L. L.; Panwar, A. S. Mechanistic Insight into the Early Stages of Amyloid Formation Using an Anuran Peptide. *Pept. Sci.* **2019**, *111* (5), e24120. https://doi.org/10.1002/pep2.24120.
(88) Xiong, H.; Buckwalter, B. L.; Shieh, H. M.; Hecht, M. H. Periodicity of Polar and Nonpolar Amino Acids Is the Major Determinant of Secondary Structure in Self-Assembling Oligomeric Peptides. *Proc. Natl. Acad. Sci.* **1995**, *92* (14), 6349–6353. https://doi.org/10.1073/pnas.92.14.6349.
(89) West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R.; Hecht, M. H. De Novo Amyloid Proteins from Designed Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (20), 11211–11216.
(90) Tro, M. J.; Charest, N.; Taitz, Z.; Shea, J.-E.; Bowers, M. T. The Classifying Autoencoder: Gaining Insight into Amyloid Assembly of Peptides and Proteins. *J. Phys. Chem. B* **2019**, *123* (25), 5256–5264. https://doi.org/10.1021/acs.jpcb.9b03415.
(91) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802. https://doi.org/10.1002/jcc.20289.
(92) Chollet, F.; Others. *Keras*; 2015.
(93) Martín Abadi; Ashish Agarwal; Paul Barham; Eugene Brevdo; Zhifeng Chen; Craig Citro; Greg S. Corrado; Andy Davis; Jeffrey Dean; Matthieu Devin; Sanjay Ghemawat; Ian Goodfellow; Andrew Harp; Geoffrey Irving; Michael Isard; Rafal Jozefowicz; Yangqing Jia; Lukasz Kaiser; Manjunath Kudlur; Josh Levenberg; Dan Mané; Mike Schuster; Rajat Monga; Sherry Moore; Derek Murray; Chris Olah; Jonathon Shlens; Benoit Steiner; Ilya Sutskever; Kunal Talwar; Paul Tucker; Vincent Vanhoucke; Vijay Vasudevan; Fernanda Viégas; Oriol Vinyals; Pete Warden; Martin Wattenberg; Martin Wicke; Yuan Yu; Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.
(94) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(95) Ray, S.; Singh, N.; Pandey, S.; Kumar, R.; Gadhe, L.; Datta, D.; Patel, K.; Mahato, J.; Navalkar, A.; Panigrahi, R.; Chatterjee, D.; Maiti, S.; Bhatia, S.; Mehra, S.; Singh, A.; Gerez, J.; Chowdhury, A.; Kumar, A.; Padinhateeri, R.; Riek, R.; Krishnamoorthy, G.; Maji, S. K. Liquid-Liquid Phase Separation and Liquid-to-Solid Transition Mediate α-Synuclein Amyloid Fibril Containing Hydrogel Formation. *bioRxiv* **2019**, 619858. https://doi.org/10.1101/619858.
(96) Hughes, M. P.; Sawaya, M. R.; Boyer, D. R.; Goldschmidt, L.; Rodriguez, J. A.; Cascio, D.; Chong, L.; Gonen, T.; Eisenberg, D. S. Atomic Structures of Low-Complexity Protein Segments Reveal Kinked β Sheets That Assemble Networks. *Science* **2018**, *359* (6376), 698–701. https://doi.org/10.1126/science.aan6398.

(97) Molliex, A.; Temirov, J.; Lee, J.; Coughlin, M.; Kanagaraj, A. P.; Kim, H. J.; Mittag, T.; Taylor, J. P. Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell* **2015**, *163* (1), 123–133. https://doi.org/10.1016/j.cell.2015.09.015.

(98) Maharana, S.; Wang, J.; Papadopoulos, D. K.; Richter, D.; Pozniakovsky, A.; Poser, I.; Bickle, M.; Rizk, S.; Guillén-Boixet, J.; Franzmann, T. M.; Jahnel, M.; Marrone, L.; Chang, Y.-T.; Sterneckert, J.; Tomancak, P.; Hyman, A. A.; Alberti, S. RNA Buffers the Phase Separation Behavior of Prion-like RNA Binding Proteins. *Science* **2018**, *360* (6391), 918–921. https://doi.org/10.1126/science.aar7366.

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

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

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

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

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

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

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

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

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

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

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

(110) 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.

(111) Louros, N.; Konstantoulea, K.; De Vleeschouwer, M.; Ramakers, M.; Schymkowitz, J.; Rousseau, F. WALTZ-DB 2.0: An Updated Database Containing Structural Information of Experimentally Determined Amyloid-Forming Peptides. *Nucleic Acids Res.* **2020**, *48* (D1), D389–D393. https://doi.org/10.1093/nar/gkz758.

(112) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2008**, *36* (Database issue), D202–D205. https://doi.org/10.1093/nar/gkm998.
(113) Kawashima, S.; Ogata, H.; Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **1999**, *27* (1), 368–369. https://doi.org/10.1093/nar/27.1.368.
(114) Tomii, K.; Kanehisa, M. Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins. *Protein Eng.* **1996**, *9* (1), 27–36. https://doi.org/10.1093/protein/9.1.27.
(115) Fauchere, J.; Pliska, V. Hydrophobic Parameters II of Amino Acid Side-Chains from the Partitioning of N-Acetyl-Amino Acid Amides. *Eur J Med Chem* **1983**, *18*.
(116) Kemper, P. R.; Dupuis, N. F.; Bowers, M. T. A New, Higher Resolution, Ion Mobility Mass Spectrometer. *Int. J. Mass Spectrom.* **2009**, *287* (1), 46–57. https://doi.org/10.1016/j.ijms.2009.01.012.
(117) Gidden, J.; Ferzoco, A.; Baker, E. S.; Bowers, M. T. Duplex Formation and the Onset of Helicity in Poly d(CG)n Oligonucleotides in a Solvent-Free Environment. *J. Am. Chem. Soc.* **2004**, *126* (46), 15132–15140. https://doi.org/10.1021/ja046433+.
(118) Mason, E. A.; McDaniel, E. W. *Transport Properties of Ions in Gases*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 1988. https://doi.org/10.1002/3527602852.
(119) 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.
(120) Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. A Database of 660 Peptide Ion Cross Sections: Use of Intrinsic Size Parameters for Bona Fide Predictions of Cross Sections. *J. Am. Soc. Mass Spectrom.* **1999**, *10* (11), 1188–1211. https://doi.org/10.1016/S1044-0305(99)00079-3.
(121) Dilger, J. M.; Glover, M. S.; Clemmer, D. E. A Database of Transition-Metal-Coordinated Peptide Cross-Sections: Selective Interaction with Specific Amino Acid Residues. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (7), 1293–1303. https://doi.org/10.1007/s13361-016-1592-9.
(122) Counterman, A. E.; Clemmer, D. E. Volumes of Individual Amino Acid Residues in Gas-Phase Peptide Ions. *J. Am. Chem. Soc.* **1999**, *121* (16), 4031–4039. https://doi.org/10.1021/ja984344p.
(123) Fink, A. L. Natively Unfolded Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15* (1), 35–41. https://doi.org/10.1016/j.sbi.2005.01.002.
(124) Uversky, V. N. Targeting Intrinsically Disordered Proteins in Neurodegenerative and Protein Dysfunction Diseases: Another Illustration of the D2 Concept. *Expert Rev. Proteomics* **2010**, *7* (4), 543–564. https://doi.org/10.1586/epr.10.36.
(125) Laganowsky, A.; Liu, C.; Sawaya, M. R.; Whitelegge, J. P.; Park, J.; Zhao, M.; Pensalfini, A.; Soriaga, A.; Landau, M.; Teng, P. K.; Cascio, D.; Glabe, C.; Eisenberg, D. Atomic View of a Toxic Amyloid Small Oligomer. *Science* **2012**, *335* (6073), 1228–1231. https://doi.org/10.1126/science.1213151.
(126) Do, T. D.; LaPointe, N. E.; Nelson, R.; Krotee, P.; Hayden, E. Y.; Ulrich, B.; Quan, S.; Feinstein, S. C.; Teplow, D. B.; Eisenberg, D.; Shea, J.-E.; Bowers, M. T. Amyloid β-Protein C-Terminal Fragments: Formation of Cylindrins and β-Barrels. *J. Am. Chem. Soc.* **2016**, *138* (2), 549–557. https://doi.org/10.1021/jacs.5b09536.
(127) Sangwan, S.; Zhao, A.; Adams, K. L.; Jayson, C. K.; Sawaya, M. R.; Guenther, E. L.; Pan, A. C.; Ngo, J.; Moore, D. M.; Soriaga, A. B.; Do, T. D.; Goldschmidt, L.; Nelson, R.; Bowers, M. T.; Koehler, C. M.; Shaw, D. E.; Novitch, B. G.; Eisenberg, D. S. Atomic Structure of a Toxic, Oligomeric Segment of SOD1 Linked to Amyotrophic Lateral Sclerosis (ALS). *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (33), 8770–8775. https://doi.org/10.1073/pnas.1705091114.

(128) Obolski, U.; Ram, Y.; Hadany, L. Key Issues Review: Evolution on Rugged Adaptive Landscapes. *Rep. Prog. Phys.* **2017**, *81* (1), 012602. https://doi.org/10.1088/1361-6633/aa94d4.
(129) de Visser, J. A. G. M.; Krug, J. Empirical Fitness Landscapes and the Predictability of Evolution. *Nat. Rev. Genet.* **2014**, *15* (7), 480–490. https://doi.org/10.1038/nrg3744.
(130) Wright, S. Evolution in Mendelian Populations. *Genetics* **1931**, *16* (2), 97–159.
(131) Sewall Wright and Evolutionary Biology, Provine https://press.uchicago.edu/ucp/books/book/chicago/S/bo5963711.html (accessed Jun 10, 2020).
(132) Fisher, R. A.; Bennett, J. H. *Natural Selection, Heredity, and Eugenics : Including Selected Correspondence of R.A. Fisher with Leonard Darwin and Others, Edited by J.H. Bennett*; Oxford : Clarendon Press, 1983.
(133) Neidhart, J.; Szendro, I. G.; Krug, J. Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model. *Genetics* **2014**, *198* (2), 699–721. https://doi.org/10.1534/genetics.114.167668.
(134) Aita, T.; Uchiyama, H.; Inaoka, T.; Nakajima, M.; Kokubo, T.; Husimi, Y. Analysis of a Local Fitness Landscape with a Model of the Rough Mt. Fuji-Type Landscape: Application to Prolyl Endopeptidase and Thermolysin. *Biopolymers* **2000**, *54* (1), 64–79. https://doi.org/10.1002/(SICI)1097-0282(200007)54:1<64::AID-BIP70>3.0.CO;2-R.
(135) Estimating some features of NK fitness landscapes | Department of Statistics https://statistics.berkeley.edu/tech-reports/590 (accessed Jun 10, 2020).
(136) Durrett, R.; Limic, V. Rigorous Results for the N K Model. *Ann. Probab.* **2003**, *31* (4), 1713–1753. https://doi.org/10.1214/aop/1068646364.
(137) Szendro, I. G.; Schenk, M. F.; Franke, J.; Krug, J.; Visser, J. A. G. M. de. Quantitative Analyses of Empirical Fitness Landscapes. *J. Stat. Mech. Theory Exp.* **2013**, *2013* (1), P01005. https://doi.org/10.1088/1742-5468/2013/01/P01005.
(138) Andreasson, J. O. L.; Savinov, A.; Block, S. M.; Greenleaf, W. J. Comprehensive Sequence-to-Function Mapping of Cofactor-Dependent RNA Catalysis in the GlmS Ribozyme. *Nat. Commun.* **2020**, *11* (1), 1663. https://doi.org/10.1038/s41467-020-15540-1.
(139) Puchta, O.; Cseke, B.; Czaja, H.; Tollervey, D.; Sanguinetti, G.; Kudla, G. Network of Epistatic Interactions within a Yeast SnoRNA. *Science* **2016**, *352* (6287), 840–844. https://doi.org/10.1126/science.aaf0965.
(140) Payea, M. J.; Sloma, M. F.; Kon, Y.; Young, D. L.; Guy, M. P.; Zhang, X.; Zoysa, T. D.; Fields, S.; Mathews, D. H.; Phizicky, E. M. Widespread Temperature Sensitivity and TRNA Decay Due to Mutations in a Yeast TRNA. *RNA* **2018**, *24* (3), 410–422. https://doi.org/10.1261/rna.064642.117.
(141) High-Throughput Mutational Analysis of a Twister Ribozyme - Kobori - 2016 - Angewandte Chemie International Edition - Wiley Online Library https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201605470 (accessed May 28, 2020).
(142) Becquey, L.; Angel, E.; Tahi, F. BiORSEO: A Bi-Objective Method to Predict RNA Secondary Structures with Pseudoknots Using RNA 3D Modules. *Bioinformatics* **2020**, *36* (8), 2451–2457. https://doi.org/10.1093/bioinformatics/btz962.
(143) Lorenz, R.; Bernhart, S. H.; Höner Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **2011**, *6*, 26. https://doi.org/10.1186/1748-7188-6-26.
(144) Zuker, M.; Stiegler, P. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic Acids Res.* **1981**, *9* (1), 133–148. https://doi.org/10.1093/nar/9.1.133.

(145) Parisien, M.; Major, F. The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data. *Nature* **2008**, *452* (7183), 51–55. https://doi.org/10.1038/nature06684.
(146) Theis, C.; Zirbel, C. L.; Zu Siederdissen, C. H.; Anthon, C.; Hofacker, I. L.; Nielsen, H.; Gorodkin, J. RNA 3D Modules in Genome-Wide Predictions of RNA 2D Structure. *PloS One* **2015**, *10* (10), e0139900. https://doi.org/10.1371/journal.pone.0139900.
(147) Ferretti, L.; Schmiegelt, B.; Weinreich, D.; Yamauchi, A.; Kobayashi, Y.; Tajima, F.; Achaz, G. Measuring Epistasis in Fitness Landscapes: The Correlation of Fitness Effects of Mutations. *J. Theor. Biol.* **2016**, *396*, 132–143. https://doi.org/10.1016/j.jtbi.2016.01.037.
(148) Letters to the Editor: The American Statistician: Vol 41, No 4 https://www.tandfonline.com/doi/abs/10.1080/00031305.1987.10475510 (accessed May 28, 2020).
(149) Satsangi, Y.; Lim, S.; Whiteson, S.; Oliehoek, F.; White, M. Maximizing Information Gain in Partially Observable Environments via Prediction Reward. *ArXiv200504912 Cs* **2020**.
(150) Wang, Y.; Ramezani, M.; Fallon, M. Actively Mapping Industrial Structures with Information Gain-Based Planning on a Quadruped Robot. *ArXiv200209710 Cs* **2020**.
(151) Ben Jaafar Aymen; Bargaoui Zoubeida. Generalized Split-Sample Test Interpretation Using Rainfall Runoff Information Gain. *J. Hydrol. Eng.* **2020**, *25* (1), 04019057. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001868.
(152) Zhang, Q.; Jia, J.; Ding, J.; Kuang, H.; Chen, Z. Maximum Information Gain Relief Feature Weighting. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*; AICS 2019; Association for Computing Machinery: Wuhan, Hubei, China, 2019; pp 630–635. https://doi.org/10.1145/3349341.3349481.
(153) Zhang, Y.; Ren, X.; Zhang, J. Intrusion Detection Method Based on Information Gain and ReliefF Feature Selection. In *2019 International Joint Conference on Neural Networks (IJCNN)*; 2019; pp 1–5. https://doi.org/10.1109/IJCNN.2019.8851756.
(154) Pressman, A. D.; Liu, Z.; Janzen, E.; Blanco, C.; Müller, U. F.; Joyce, G. F.; Pascal, R.; Chen, I. A. Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. *J. Am. Chem. Soc.* **2019**, *141* (15), 6213–6223. https://doi.org/10.1021/jacs.8b13298.
(155) Danger, G.; Plasson, R.; Pascal, R. Pathways for the Formation and Evolution of Peptides in Prebiotic Environments. *Chem. Soc. Rev.* **2012**, *41* (16), 5416–5429. https://doi.org/10.1039/C2CS35064E.
(156) Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, *31* (13), 3406–3415. https://doi.org/10.1093/nar/gkg595.