Computational Discovery of Novel Secondary Structures from Non-Canonical Amino Acids

Adam P Moyer

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

David Baker, Chair

Christine Luschome

Frank Dimaio

Jesse Zalatan

Program Authorized to Offer Degree:

Engineering

University of Washington

Abstract

Computational Discovery of Novel Secondary Structures from Non-Canonical Amino Acids

Adam P Moyer

Chair of the Supervisory Committee:
David Baker
Biochemistry

Protein secondary structures are a fundamental component of biological macromolecules, which are responsible for the myriad molecular processes of life. However, these biological protein macromolecules are not evolved to be amenable for rational molecular engineering due to the conformational polymorphism of alpha amino acid polymers. Therefore, developing new protein secondary structures that exhibit exceptional stability and consequently programmability could accelerate the efforts to engineer and, more importantly, utilize designed macromolecules. In this thesis, I describe a computational method to identify new secondary structures which are composed of non-canonical amino acids. Further I describe the subsequent experimental validation of these secondary structures to expand the known secondary structures for future molecular engineering efforts. Here I considered a pool of 135 non-canonical amino acid building blocks to create combinations of di-peptide repeat units, because this theoretical space has only been

sparsely investigated previously. In total, the combinations of these building blocks yielded over 15,000 unique sequences which were computationally evaluated for their propensity to form a stable helical structure. Due to the lack of experimental data on the conformational properties of these residues, I developed an exhaustive and adaptive resolution computational search method to efficiently sample the enormous space of potential conformations. This method enabled the computational evaluation of the entire molecular conformational ensemble. Using this method, I identified 10 novel secondary structures which were expected to occupy a single low energy state. Experimental evidence suggests that these molecules are well-folded, engineerable helical polypeptides. Moreover, a select secondary structure was characterized as a polymer to explore the potential for these molecules as new classes of helical polymers for future materials applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank the countless researchers that developed theories, experimental methods, and software that enabled my work over the many years before I began my scientific endeavor. It is cliché, but scientific progress is built on the shoulders of giants.

David Baker and the lab that he has fostered was an incredible environment to develop as a scientist. The number of great ideas that float around the lab is truly special. Sometimes it can be hard to focus on one, and David knows that. He gives you the time to explore and develop your skills until you find the perfect project. For that, I will always be grateful.

My colleagues have been exceptional support, mentors, and inspiration. I have met people from all over the world and learned so much from all of them. I always enjoy going to group meeting because I know that the talk will be world class research.

Some colleagues have even become friends, and I am lucky for that. Florian, Brian C, Derrick, Dmitri, Tim, and Brian W have climbed literal mountains with me. They helped me develop as a scientist and a person.

My partner, Rose, has been a constant inspiration and understanding support. She is an amazing scientist and teacher as well as an exceptionally thoughtful and funny person.

Lastly, I would like to thank my parents, Sheldon and Michelle. They were my first inspiration to be a leader and inventor, and they have always supported me to follow my dreams. For that, I will always be thankful.

Chapter 1. INTRODUCTION

Because this work focuses on re-inventing the fundamental structure of macromolecules, I must first describe the current perspective of fundamental structures. I will go through the history of the original macromolecules, biomolecules, which includes protein, DNA, and RNA. I will focus on protein structure because proteins are most similar to the novel polypeptides that are described in this thesis. After describing biological protein structure, I will discuss alternative macromolecules which are composed of non-canonical building blocks. These molecules range from obscure examples from nature to fully synthetically developed molecules which were created in a laboratory setting. Next, I will describe the previous attempts from literature to computationally discover and design new macromolecules. This will be useful to compare with the novel work described in this thesis. Finally, I will describe the motivation and challenges to discover new secondary structures computationally.

1.1 THE STRUCTURE OF MOLECULES

Molecules have hierarchy of structure (*1-3*). This hierarchy is known as primary through quaternary structure and can explain interactions ranging from covalent bonds up to multi-molecular assemblies. Usually, this hierarchy of structural definitions is applied in the context of proteins, DNA, and RNA. However, this hierarchy is useful to describe all molecules. Because the following work of this thesis is so coupled to the fundamental definition of structure, I feel it is important to define the various levels of structure within the context of general molecules. These levels of molecular organization are defined as followed:

- Primary structure – chemical composition of macromolecules
- Secondary structure – local and repeatable substructure of macromolecules

- Tertiary structure – non-local structure of macromolecules
- Quaternary structure – intermolecular structure of multi-macromolecular complexes

1.2 The Primary Structure of Proteins

The alpha amino acid sequence is the primary structure of proteins. There are twenty amino acids that are regularly incorporated into proteins. These amino acids can be thought of as tri-substituted carbon centers with two backbone substitutions, the carboxylic acid and the amine, and a side chain substitution which is a variable group. The backbone is named as such because the polymeric chain propagates through those atoms. Nearly all of the amino acids share the same chiral center at the alpha carbon. The only exception is glycine which does not have a chiral center because the alpha carbon is only di-substituted.

The conformation of a residue can be described by the dihedral angles of the backbone atoms. These torsions are the major degrees of freedom of a residue. These torsions are known as phi, psi, and omega, and these torsions are important because they are used to define the secondary structures of proteins. Generally, the side chains of amino acids only moderately affect the backbone torsional preferences, but the sidechains are important because they significantly affect the global tertiary structure of a molecule. Also, residues vary in propensity to form different secondary structures which is a result of the various sidechains. For example, beta branched amino acids have a greater propensity to form beta-sheet structures, but this propensity does not exclude them from alternative structures such as the alpha helix. Proline and glycine are exceptions are exceptions worth noting. Proline forms a 5-member cycle through the side chain to the amine of the backbone. In the context of the polypeptide, this eliminates the hydrogen bonding capabilities of the backbone amide and limits the available conformational space, and proline is not able to form the most common secondary structures that require hydrogen bonding. With that said, the

rigid structure of proline enables other unique secondary structures such as the poly-proline helix. In contrast, glycine is able to occupy a larger conformational space compared to all of the other residues due to the lack of side chain substitution. In general, greater flexibility is not conducive for structuring a polymer into a distinct structure, and glycine is generally found in the irregular regions of proteins.

1.3 The Secondary Structure of Proteins

There are several forms of secondary structure possible given native primary structure of proteins which includes the alpha helix, pi helix, 2-7 helix, 3-10 helix, beta sheet, polyproline I (PPI) helix, polyproline II (PPII) helix, collagen, elastin, keratin, gramicidin (beta helix), valinomycin (beta ribbon), and poly(γ-glutamic acid).

The alpha helix, the 310 helix, beta sheet, and PPII helix are ubiquitous within the context of proteins. They make up nearly 50% of residues of native proteins while the majority of the remaining residues are considered to be irregular (*4*-*6*).

Structures such as collagen, elastin, and keratin are utilized by biological systems to fulfill particular material functions (*7*-*9*). These macromolecular material properties are derived by the nature of the secondary structure which form compressible, elastic, or rigid assemblies, respectively.

Structures like the pi helix, 2-7 helix, and PPI helix, while theoretically possible, are rarely observed in nature because the helices are high energy states in the solvated state of proteins. The PPI helix was experimentally found to be the form of polyproline in isopropyl alcohol (*10*). The pi helix is well documented as a single bulge of an alpha helix, but there are no known examples of a stable pi helix with multiple turns (*11*).

Each of the secondary structures described above utilize alpha amino acids with L chirality. However, there are a few notable exceptions to this canonical structure found in nature. First, the antibiotics gramicidin and valinomycin utilize mixed L-D chirality residues to form unique helices that are structured only in the context of the membrane and when bound to an ion (*12*-*14*). Second, poly(γ-glutamic acid) uses a gamma amino acid backbone, rather than an alpha amino acid backbone. The acid group of the sidechain of glutamic acid polymerizes with the canonical backbone amine group to form helices (*15*). These motifs are the first hints from natural products that secondary structures can form without α-L-amino acids.

All of these native secondary structures demonstrate varying levels of complexity which is useful to recognize when designing new secondary structures. This complexity spans from 1) helices like the alpha helix which are composed of a single repeating residue to 2) secondary structures that are only stabilized in the context of more of the same secondary structure such as the beta sheet to 3) secondary structures that utilize multiple residues within the repeating motif like collagen and elastin to 4) secondary structures that utilize non-canonical residues like gramicidin and valinomycin.

1.4 Predicting the Secondary Structure of Proteins

Before the structure of proteins was determined experimentally, it was suspected that the residues of proteins formed into secondary structures. There were several seminal papers that attempted to predict the structure of protein secondary structures. Pauling et al (*16*) are famously credited for accurately predicting the structure of the alpha helix. Interestingly but less known, they also predicted a second isoform of a helical polypeptide which was termed the pi helix in the same work. The alpha helix was found to be the most abundant secondary structure of proteins;

the pi helix is found in nature but rarely. While the predicted torsion values differed from observed torsion values by approximately 15 degrees, the work is considered a success.

Pauling et al (*16*) describe the method to deduce the structure of the polypeptide chain as "complete and accurate determination of the crystal structure of amino acids, peptides, and other simple substances related to proteins…might be obtained that would permit the reliable prediction of reasonable configurations of the polypeptide chain." In other words, they utilized a fragment-based approach which, when combined with human intuition, accurately predicted the macromolecular assembly of polypeptides. The work is a remarkable extrapolation from first principles to predict emergent macromolecular behavior. However, the method lacks the ability to discriminate false positive predictions; this is necessary when exploring outside the chemical space of proteins.

In contrast, the work by Bragg et al (*17*), which was a previous publication that attempted to predict the structure of the polypeptide helix from the initial diffraction data of myoglobin, missed the possibility of the alpha helix and the pi helix. Instead, they predicted the second most common helical form which is known as the 3-10 helix. The predictions by Braggs et al were not comprehensive because they only considered helices of perfect repeat, or in other words, an integer number of residues per turn of the helix. The alpha and pi helix had 3.6 and 5.1 residues per turn, respectively. Also, they only loosely followed the geometric constraints from known molecular structure. Most notably, the torsion angle of the amide bond was not strictly limited to a trans conformation. This anecdote builds a cautionary tale to follow first principle and avoid applying arbitrary constraints.

Over application of constraints was also the downfall of those predicting the geometry of beta sheets. Again, Pauling et al (*18*) published the most successful prediction of the structure of

beta strands in the context of beta sheets. At the time, it was believed that the structure of the polypeptide backbone in a pleated structure would be fully extended. However, this would result in a repeat spacing of 3.6 A structural repeat instead of the observed 3.3 A diffraction. Pauling et al found by scanning the dihedral angles of the polypeptide backbone and calculating atomic coordinates that there was a set of dihedrals that maintained the planarity of the amide bond hydrogen bonds between chains while reducing the repeating distance of the polypeptide in agreement with the diffraction data. However, while the prediction was much higher resolution than any other previously proposed model, they did not consider a higher order twist along the polypeptide chains which was ultimately found to be a predominant feature of beta sheet containing proteins.

Predicting structure of the collagen triple helix proved to be quite challenging. Despite multiple attempts by several well-known scientists of the time such as Pauling and Ramachandran, it was properly predicted over a decade after the original alpha helix and beta sheet predictions of 1951. While previous predictions were correct in recognizing there are three parallel molecules that share a common helical axis, Crick et al (*19*) were able to successfully describe the structure of the collagen helix with two critical insights. First, abiding by the torsional constraints of the pyrroline ring for the proline residues was absolutely necessary. Second, accommodating a register shift between the parallel chains was required to realize the structure of collagen, which is not an apparent degree of freedom a priori.

The structure of gramicidin was also predicted before the structure was elucidated (*20*). Gramicidin was the first prediction that took into account a polypeptide with mixed stereocenters of the side chain functional groups. Therefore, it could be considered the first prediction of a non-canonical secondary structure.

1.5 The Primary Structure of Non-Canonical Amino Acids

Since the first structure of a protein was determined by x-ray crystallography, synthetic chemists have been inspired by the incredible structure of biological proteins. Synthetic molecules which fold into unique structures are known as "foldamers" within the broader community of synthetic chemists (*21*). There have been wide variety of foldamers that are made with synthetic derivatives of amino acids ranging from alpha, beta, gamma, delta, and aromatic amino acids. These residues may include cycles that constrain the conformational degrees of freedom, which has produced foldamers with exceptional propensity to fold. Peptoids, or N-substituted glycine polymers, are a well-represented building block in the literature due to the readily available synthetic derivatives. Molecules of mixed building blocks have also been shown to adopt well folded structures. A spectrum of amino acids that have been utilized in foldamers is described below.

Even before the structures of proteins were known, companies were utilizing the properties of polyamide polymers for various innovative materials. The most known examples, Nylon (*22*) and Kevlar (*23*), are synthetic polyamides developed by the DuPont Company which are composed of aliphatic and aromatic building blocks, respectively. These molecules are excellent examples to demonstrate the utility of the emergent materials before the structure of the polymers was considered. It would be expected that with rational design and a molecular understanding, even better materials could be developed.

There are many other linkages that have been utilized within foldamers besides the amide bond of amino acid building blocks. Most notably, there have been examples of molecules with urea, triazole, sulfonamide, aminooxy, hydrazino, and meta-phenylacetylene linkages between subunits and various combinations thereof (*24-29*). However, these linkages will not be considered

in the context of this work because I am focusing on the well-developed amide coupling chemistry. The methods developed in this work will transfer to these chemistries as well as any chemistry that is developed in the future.

Notably, enzymology labs have begun to focus on developing enzymes that can produce a diverse set of new molecules which could be used as macromolecular building blocks. The Arnold Group is leading that progression. Recently, they have published multiple examples of enzymes which can produce new amino acid building blocks at a commercial scale (*30*).

1.6 The Secondary Structure of Non-Canonical Amino Acids

The literature of reported secondary structures is expansive. There have been several reviews within the last 10 years (*21, 31-40*). Most exploration of the space has utilized human intuition to speculate the potential secondary structure of various synthetic building blocks. There have been theoretical studies which have attempted to rationalize discovered and predict undiscovered secondary structures (*41-43*). Most of the space of aliphatic amino acids with full satisfied hydrogen bond donors has been fully enumerated up to delta amino acids by theoretical studies. The review by Vasudev et al (*44*) from 2011 is an excellent source that describes most of the possible helical conformations of alpha, beta, and gamma amino acids with fully satisfied hydrogen bonding known at the time.

1.6.1 *Beta Amino Acids*

The helices composed of beta amino acids were first described around 1996 by Seebach and Gellman (*45, 46*). They stabilized the C14-helix by using beta-3-substituted-amino acids and trans-2-amino-cyclohexanecarboxylic acid. These helices have been found to be more thermodynamically favorable and protease resistant compared to the alpha amino acid counter

parts. The C12 helix was found to be stabilized by the constrained residue, trans-2-aminocyclopentanecarboxylic acid. Likewise, it was found that the more extended C10 and C8 helices were able to be stabilize via cyclobutane and cyclopropane derivatives, by Fleet et al (*47*) and Abele et al (*48*), respectively.

1.6.2 *Gamma Amino Acids*

Then in 1998, Seebach reported the structures of the first polymers composed of gamma amino acids (*49*). Using gamma-4-substituted-amino acids, they found that the C14 helix is natively adopted. Hoffman predicted that the C9 helix was also possible for gamma amino acids, which was later realized by utilizing residues with large sidechains and backbone cyclized residue. It is suspect that gamma-3-substituted-amino acids and gamma-2-substituted-amino acids do not adopt a well-defined structure because they exhibit limited dispersion by NMR. Interestingly the C7 ribbon-like helix was stabilized by utilizing a constrained residue that formed the intra-residue hydrogen bond. The most extensive review of this work is by Bouillère et al in 2011 (*40*).

1.6.3 *Heterogeneous Backbones*

Alpha amino acids with mixed stereochemistry were the first examples of secondary structures with heterogeneous patterns. The first solid state characterization of a non-canonical secondary structure was from Di Blasio et al in 1991 (*50*). They found that polymers of proline (Pro) and alpha aminoisobutyric acid (Aib) form a beta bend ribbon, previously found only in native proteins due to the torsional strain of the conformation without the di-substituted alpha carbon of alpha amino isobutyric acid. The work of Kulp et al (*51*) in 2011 demonstrated that a single isoform of the beta helix similar to gramicidin could be stabilized in solution by cyclizing the antiparallel dimer structure into a single chain.

Heterogeneous backbones of alpha, beta, and gamma amino acids have been explored as well. However, the chemical space of combinations of residues starts to become much bigger because there are two dimensions which vary, both residue type and chirality. The most prominent example is the 10/12 helix, which is composed of alternating B2 and B3 amino acids of matching chirality. Patterns of alpha and beta amino acids have been investigated with 1:1 and 1:2 patterns by Choi et al (*52*), but the explored sequence space was very limited. Hoffmann, again, published a paper (*42*) predicting the possible helices of mixed alpha and beta residues in 2006 which is before Choi published those results. Even peptides composed of all three residues, alpha, beta, and gamma, have recently been explored by Shin et al (*53*). Heterochiral beta polymers have been studied to a limited extent by Martinek et al (*54*).

1.6.4 *Peptoids*

Peptoids are polymers of N-substituted glycine. Conformationally, this class of polypeptide is unique because it is not capable of forming backbone hydrogen bonds, phi can adopt both negative and positive values, and omega can adopt both cis and trans conformations. Therefore, it may seem that peptoids have more degrees of freedom compared to other residues, but N-substitutions constrain the preceding residue similar to the effect of proline (*55*). Also, there is an exceptionally large gamut of sidechains available for peptoids because any primary amine can be incorporated into the polymer. Blackwell et al (*56*) used peptoids to stabilize a PPI helix and a ribbon-like structure that alternates cis and trans, both of which utilize a peptoid sidechain that favors the cis conformation. Mannige et al (*57*) have described a sheet-like conformation for peptoids which is stabilized in the context of an assembly.

1.6.5 *Polyaramides*

Polyaramides are the class of polypeptides with aromatic functional groups within the backbone chain of atoms. These residues are interesting because they are highly constrained and predictable. While synthetic chemists have explored many aramid polymers, and the lab of Ivan Huc has led the innovation of aramids. They have demonstrated helices with remarkable stability and functionality. For example, the polymers have shown to be capable of effective charge transfer and separation across defined length which is ideal for engineering organic electronics (*58*). Huc has also shown that heteropolymers of amino acids and aramids are capable of forming structured helices (*59*). They have also found that aramids are so rigid, one can stabilize secondary structures without long range contact order (*60*). Aramids are poorly explored compared to the aliphatic counter parts because anilines are more challenging to synthesize due to poor nucleophilicity.

1.6.6 *Peptide Nanotubes*

Peptide nanotubes are ring-like assemblies of stacking cyclic peptides of residues with alternating chirality in a sheet-like conformation. Ghadiri first described this structure in 1993, and the peptide nanotubes exhibited exceptional stability. Later work by Clark and Granja expanded the cyclic sheet architecture to beta and gamma amino acids. Interestingly, Clark has demonstrated the utility of nanotubes in the context of high-performance materials by creating a single chain polymer via sidechain crosslinking. These polymers possess extensive intermonomer hydrogen bonding which is calculated to yield exceptional elasticity properties.

1.6.7 *Unexplored Space*

Overall, the field of foldamers is incredibly expansive, and while a lot has been done already, there is still far more space to explore. It is interesting to notice that there has been no

investigation of the secondary structures containing combinations of residues with hydrogen bonding and non-hydrogen bonding to my knowledge. Few studies have utilized non-covalent interactions besides hydrogen bonding or pi stacking to direct folding, such as halogen bonding or chalcogen bonding. Utilizing noncovalent sidechain interactions to stabilize a secondary structure has not been explored either.

1.7 The Motivation to Discover New Secondary Structures

The discovery of new secondary structures is fundamental to the development of molecular engineering. While proteins are currently the most used macromolecular system for engineering purposes, fundamentally, proteins were not developed with the intention of rational engineering. Rather, proteins evolved to create and support the functions of life. It is important to note that properties that are important to engineering such as stability and predictability are detrimental features to systems that utilize evolution to develop. Instead, evolution has worked through promiscuity and polymorphism, which allows small random changes to significantly affect the performance of the molecules.

Just as humans do not use the wings of birds to fly and instead developed aviation from robust mechanical parts, molecular machines must be developed similarly. Ideally, engineers would like to develop a set of molecular tools and machines, so that engineers could eventually build the molecular equivalent of an airplane. To do this, there need to be molecular devices that are equivalent to a hammer or a screwdriver that are resilient to modification and environmental conditions.

While the field of protein engineering is attempting to use proteins to fulfill the role of these molecular devices and machines, there is no fundamental reason, besides the convenience of existence and consequential ease of synthesis, to limit molecular engineering to natural proteins.

Superior building blocks which yield superior macromolecules almost certainly exist because the chemical space of possible molecules is incredibly large.

Specifically, new secondary structures could enable new technologies. Most importantly, secondary structures with exceptional stability and predictability would enable the rapid and robust design of globular macromolecules that resemble proteins. With highly predictable design of structure due to the greater performance of the new secondary structures, the *de novo* design of enzymes, therapeutic binders, and material assemblies could become routine, rather than the grand challenge these aims currently are.

The challenge that most interests me and why I decided to focus on secondary structures is the design of new helical polymers for materials and electronics applications. Helical polymers have already been shown to be capable of both of these tasks. In the context of materials, polymers with intramolecular or intermolecular hydrogen bonding demonstrate high elasticity or high strength, respectively. These are the most desirable properties of high-performance materials. If we could design these atomic level interactions rationally, we could begin to optimize and understand these properties from first principles. For electronic applications, highly structured polymers are critical for maximizing the charge transfer efficiency through the polymer chain. Electronic polymers have been used in light emitting devices and light absorbing devices. Improving the efficiency of these devices is critical for conserving and producing energy, the greatest technological challenges of humanity. For both of these applications, the design and organization of hierarchical assembly of helical polymers, while is not included in the context of this work, could be another future direction that would be enabled by these new examples of helical polymers.

1.8 The Challenges to Discover New Secondary Structures

Discovering new secondary structures has both computational and experimental challenges. Herein, I will describe the issues that I recognized while working on this problem. For some of these challenges, I have found adequate solutions which will be described in later chapters. In general, I focused my effort on solving computational challenges.

1.8.1 *Computationally Evaluating the Energy of Non-Canonical Amino Acids*

Computationally modeling non-canonical amino acids is a challenge because there is limited experimental data that documents the conformational preferences of those non-canonical residues. This experimental data is critical for molecular mechanical modelling, the most common modelling technique used to analyze the energy proteins. These molecular mechanical models are heuristic fits to the empirical data of crystal structures, where the frequency of a conformation of a residue in the database of protein crystal structures correlates with the energy of the conformation; specifically, a high frequency correlates with a low energy.

When modelling canonical residues, there are more than 100,000 crystal structures to use. In contrast, for many non-canonical residues, there are less than 10 crystal structures to use.

To make matters worse, most crystal structures of non-canonical residues are not in the context of macromolecules. This is important because the context of a macromolecule with many competing factors that contribute to the energy allow residues to occupy higher energy states than the lowest energy state. In small molecule crystallography, molecules do not need to make accommodations to satisfy a global energy minimum like macromolecules. Therefore, small molecule crystal structures will be redundant and give little information on the potential energy landscape of the non-canonical residue in the context of the secondary structure of a macromolecule.

To address this lack of knowledge, the solution is to apply computational techniques which do not utilize experimental data to calculate the energy of the non-canonical residues. Using *ab initio* quantum mechanics, one can calculate the energy of a molecular conformation with accuracy, but these calculations require a significant amount of computational time which is grossly unacceptable for molecular design. Semi-empirical density functional theory (DFT) represents a compromise between full quantum mechanical calculations and molecular mechanic calculations, by compromising accuracy for speed. Historically, this theory was appropriate for the design of new secondary structures using non-canonical amino acids. However, recent advances in machine learning techniques approximate the results of the most accurate and expensive *ab initio* calculations with a 1 kcal/mol error while reducing the time requirements by 100,000x. This leaves the method at the same speed and accuracy as molecular mechanical methods which are fit for proteins, but these machine learning models are generalizable to a wide range of chemical space. Therefore, machine learning methods are of great use when designing macromolecules with non-canonical amino acids.

1.8.2 *Computationally Sampling the Conformations of Non-Canonical Amino Acids*

Improving the speed of individual calculations while maintaining accuracy of molecular energy calculations is a significant challenge that must be addressed. However, an alternative approach to improving the speed of individual calculations is to improve the efficiency of sampling. Fundamentally, computational molecular design is a sampling issue. One must find molecules that adopt a unique low energy state, and to do this, one must consider many states of a molecule to identify not just the lowest energy state but also that there are no alternative states that are also low in energy. Ideally, the entire conformational ensemble of a molecule would be evaluated. If the entire ensemble is calculated, one could exactly calculate the probability of being

in any one conformation given the principles from statistical mechanics. However, exhaustively sampling the conformations of any moderately complex molecule is not just a challenge but essentially impossible.

Given this problem, each degree of freedom of a molecule (bond lengths, bond angles, and bond torsions) is a dimension which must be sampled to yield the complete ensemble, and each degree of freedom must be perturbed in the context of all other degrees of freedom. Even a simple molecule like ethane (C_2H_6) has 7 bond lengths, 6 bond angles, and 1 bond torsion which should be considered to create a complete ensemble. If one samples each degree of freedom with an arbitrary but small value of 10, that will total $1.0*10^{14}$ unique conformations. Given 1 second per calculation, that is $2.8*10^{10}$ hours to evaluate the full ensemble. Obviously, this is not acceptable for practical applications. Therefore, simplifications must be made. One solution is to consider only the most critical degrees of freedom that affect the energy of a molecule, such as the bond torsions. This is a reasonable simplification because molecular design ignores the highest resolution features such as bond length and bond angle because those degrees of freedom are essentially locked at the resolution of molecular design. The resolution of molecular design is currently limited to predicting the bond torsions and the subsequent structure at ~1A resolution. Modifying bond lengths and bond angles which are significant enough to change the macromolecular structure are simply energetically prohibitive.

Even after identifying and considering only the most critical degrees of freedom of a molecule, many molecules of interested will have at least 2 degrees of freedom. For example, alpha amino acids have 2 significant degrees of freedom, phi and psi, within the backbone alone. Non-canonical residues can have arbitrary amounts of degrees of freedoms. If one assumes that adequate sampling of these torsional degrees of freedom is 10-degree increments, each torsion will

need 36 samples. Below is a table that lists the total number of samples required to fulfill that resolution with respect to the number of degrees of freedom in the molecule:

Table 1.1. Total Number of Samples With 36 Samples Per Dimension

Dimensions (N)	**36^N**
1	36
2	1,296
3	46,656
4	1,679,616
5	60,466,176
6	2,176,782,336
7	78,364,164,096
8	2,821,109,907,456

The number of samples scales exponentially with the number of dimensions. This is known as the "curse of dimensionality" and developing effective methods for sampling this hypercube of space is a hard problem. While the volume of the possible space continues to grow as the number of dimensions grow, the "needle in the haystack" remains the same size. Solutions to improve the efficiency of sampling must be specific to the given problem. For molecules, three critical assumptions can be made: 1) only low energy states are important, 2) dimensions are not equally important, and 3) the interactions between dimensions are close to linear. The consequences of these assumptions are that some dimensions can be sampled less intensely and information from sampling one dimension independently can be useful when sampling dimensions in conjunction. Most importantly, sampling techniques can focus on the low energy regions. Therefore, identifying and preferentially sampling low energy states in this large search space is critical to improving efficiency.

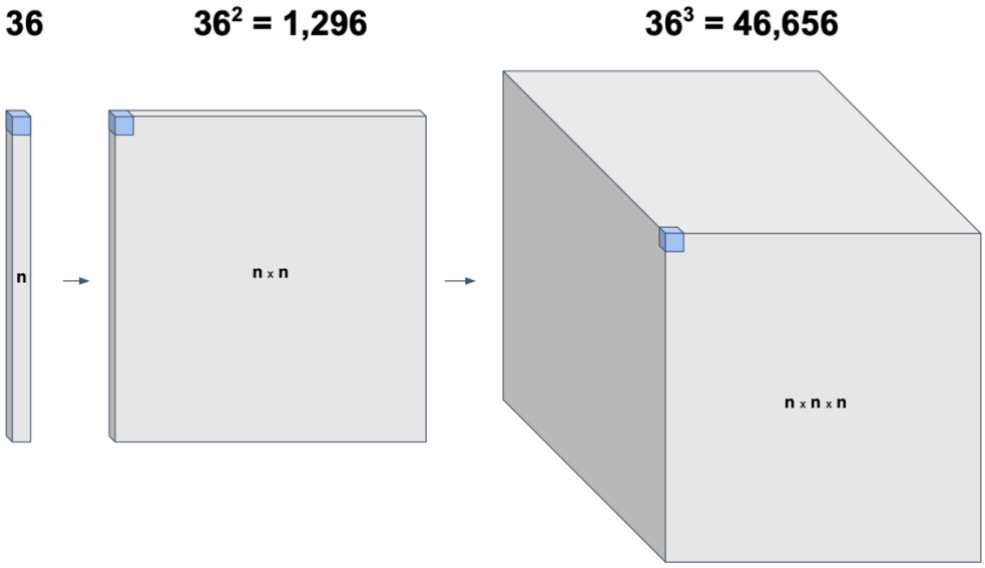


Figure 1.1. Visual representation of the relationship between volume and dimensions. A point of interest (blue cubes) is included as a comparison to the total volume (white cubes). As dimensions increase, the fractional volume of the point of interest and the total volume decreases.

Molecular systems also have symmetry which can be utilized to reduce the amount of sampling required to exhaustively sample a space. Generally, in the context of molecular design, symmetry has been a powerful tool which has been utilized to design large molecular systems. Fortunately, the discovery and design of new secondary structures can also take advantage of symmetry. Because secondary structures are helically symmetric systems, multi-residue helical polymers can be simplified into the degrees of freedom contained within the simplest subunit. For example, if you model a polymer of alpha amino acids to rediscover the alpha helix, the relevant degrees of freedom can be reduced to the phi and psi angle of one residue. All other residues within the helix can be modeled as copies of that single residue. This same concept will apply to non-canonical systems and more complicated subunits such as examples with multiple residues in the simplest subunit. It is important to point out that assuming symmetry is another simplification. In

this study, I am only interested in helically symmetric molecules, and therefore I am limiting the sampling to helical symmetry. While this assumption is necessary for limiting the search space to an achievable dimensionality, this assumption neglects the possibility of helices where the symmetric subunit is more than one of the simplest subunits (i.e., a secondary structure composed of two alpha amino acids of different conformations that alternate in the polymer chain).

1.8.3 *Experimentally Determining the Structure of Non-Canonical Secondary Structures*

Experimentally determining the structure of molecules is a general challenge, and a key point of this work is validating the computational predictions with experimental data. Common techniques that give insights into the structures of molecules include electron microscopy (EM), nuclear magnetic resonance (NMR), x-ray crystallography, circular dichroism (CD) spectrophotometry, and atomic force microscopy (AFM). EM is applicable only when determining the structure of large macromolecules, so that technique will not be discussed here. While CD spectrophotometry and AFM are relatively simple to perform and analyze, they do not provide atomic level detail. NMR and X-ray crystallography are preferred as they can provide atomic resolution. However, the analysis of NMR data is laborious and has its limitations, and X-ray crystallography requires exhaustive experimental conditions. Also, the contacts of the crystal lattice can affect the conformation of the molecules.

NMR characterization of molecules is a well-established method to determine the structure of a molecule in solution. NMR can provide information on the ensemble of a molecule, and it can detect if there are multiple conformations if the conformations are slow to exchange. NMR is suited for both small molecules and larger proteins with tertiary structure, but the larger the molecule the more challenging it is to interpret and assign the data from the NMR spectrum. Solving the secondary structure of molecules with 100 heavy atoms, or 10 residues, is well within

the size range of the technique; however, the inherent repetitive nature of secondary structures emerges as a unique problem. Repetition of structure is a problem for NMR characterization because the environment of the residues within the molecule is identical, other than the first and last residues of the chain. When residues are identical chemically and environmentally, the residues will overlap on the NMR spectrum, making it impossible to unambiguously assign peaks for accurate structure determination.

X-ray crystallography is the gold standard for molecular structure determination because the method can solve structures with sub-angstrom accuracy. However, x-ray crystallography has drawbacks. Preparing a sample for x-ray crystallography is a fickle endeavor experimentally. The molecule of interest must precipitate from a solution as a large, single crystal which is a rare event. Therefore, many trials to yield a crystal must be done. Also, the structure of the molecule is solved in the context of a solid instead of solution. Molecules may take on different, arbitrary conformations as a solid, and typically the solution properties of a molecule are the more important properties for function.

CD is a spectroscopic technique that is performed in solution yielding a spectrum that is dependent on the structure of a macromolecule. Therefore, molecules that are identically chemical but differ structurally due to environmental factors like temperature and pH will produce different spectrums. This makes CD ideal for observing changes in structure.

Atomic force microscopy (AFM) measures the topology of molecules on a surface. AFM is able to analyze the conformation of individual molecules, which is exclusive to this technique compared to the previous methods. However, AFM is not able to resolve atomic level details therefore it cannot replace more involved techniques like x-ray crystallography and NMR.

Chapter 2. COMPUTATIONAL DISCOVERY OF NOVEL NON-CANONICAL SECONDARY STRUCTURES

While the previous chapter goes through the history of molecular structure and more specifically, primary and secondary structure, I will herein describe the process that I developed to computationally discover new secondary structures from non-canonical amino acids. I sought to systematically explore the space of possible helical polypeptides to identify new secondary structures that are composed from non-canonical amino acids because I hypothesized that these new secondary structures could improve the engineerability and diversity of known secondary structures. Specifically, I focused on secondary structures composed di-peptide building blocks. I presumed that a large-scale computational effort would be able to screen far more residue compositions than the previous low throughput synthetic efforts. Also, I recognized that a computational protocol could enable the evaluation of costly or synthetically challenging residues that would typically be excluded from a comparable experimental effort. Lastly, I sought to develop a method that precluded preconceived intuition and heuristics to avoid human bias into the resulting structures.

To identify the di-peptide building blocks that could form into a favorable conformation with helical symmetry, I devised a hierarchical computational sampling protocol. Broadly, the protocol has three parts. First, calculate the potential energy surface of all of the residues as monomeric units. Second, use the aforementioned calculations to bias sampling of di-peptide fragments in the context of helical symmetry. Finally, analyze the ensembles of the helically symmetric polymers to identify di-peptide building blocks that exhibit a single low energy conformation.

2.1 The Gamut of Non-Canonical Building Blocks

In total, I considered a set of 135 non-canonical amino acid building blocks. Figure 2.1 and Figure 2.2 visually describe the residues considered in this study. Most of these residues are readily available from commercial vendors. While canonical amino acids utilize only one pattern of atom types along the backbone chain, this gamut of non-canonical amino acids contains 25 unique patterns of atom types along the backbone chain. The remaining variety of the residues is from the sidechain substitutions and cyclization patterns, which were selected because we expected these residues to significantly affect the conformational preference in the context of a helical structure. Of the 135 residues, 90 of the residues had a chiral center, and 40 residues were proline-like, meaning that the terminal amine of the amino acid was secondary. While many of these residues have been considered as building blocks for novel secondary structures in previous studies, the combinations of these residues as di-peptide building blocks have only been sparsely explored. Therefore, we focused our efforts on finding secondary structures that were composed of combinations of two residues. When taking into account the chirality of the building blocks, the 135 building blocks yielded approximately 15,000 combinations of unique di-peptide building blocks.

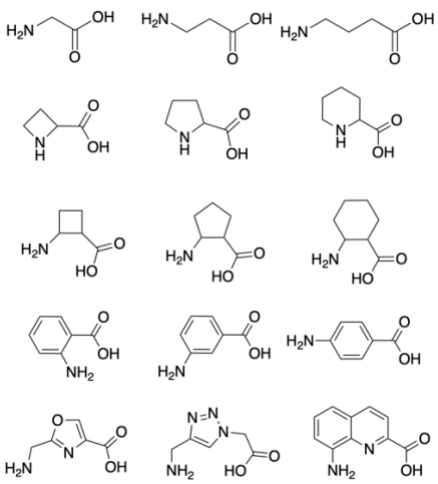


Figure 2.1. An example assortment of residues considered in this study. The residues vary in size, molecular composition, and geometry. All residues are amino acids which means that the residue contains an amine and carboxylic acid group.

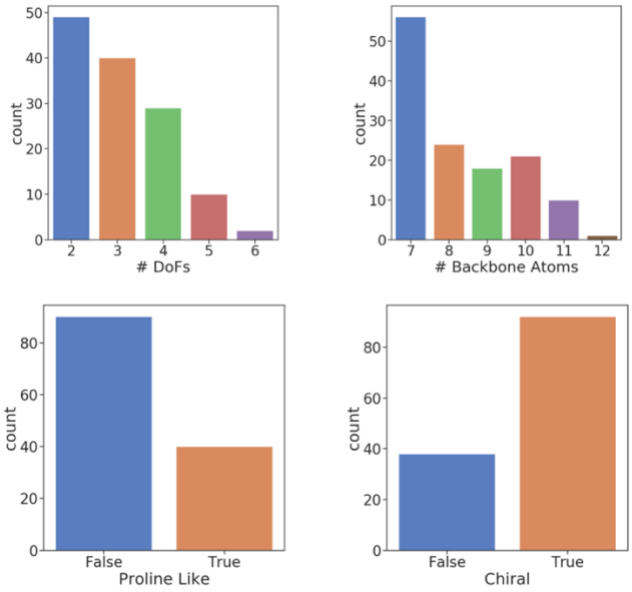


Figure 2.2. Descriptive properties of the gamut of non-canonical amino acids. Top left is the count of residues given the number of major degrees of freedom. Top right is the count of residues given the number of atoms in the backbone. Bottom left is the number of residues that are proline-like, contain a secondary amine. Bottom right is the number of residues that contain a chiral center.

2.2 Parameterizing Non-Canonical Building Blocks

Accurate and exhaustive calculation of the potential energy surfaces of the monomeric residues is a critical challenge because this data biases all downstream sampling steps. This step is a challenge because there is limited experimental data to accurately parameterize the energy landscape and the non-canonical residues have up to 6 degrees of freedom which requires extensive sampling. To address these issues, (1) we evaluate the energy of the non-canonical residues with AIMNet(SMD)-D4 (*61-66*), a neural network trained on millions of density functional theory (DFT) calculations with a dispersion correction and implicit solvation, and (2) a novel sampling technique for molecular systems which we call Boltzmann Adaptive Sampling (BAS). These improvements made it trackable to sample the entire potential energy surface of the non-canonical residues without *a priori* knowledge. The molecular data structures were managed using RDKit (*67*, *68*), ASE (*69*), and NumPy (*70*).

Monomeric residues are parameterized in a context that is as close as possible to the context of a polymeric chain. Therefore, all residues are generated with two variations. The first variation is the normal context with a N-terminal acetylation and a C-terminal methylamidation. These terminal modifications mimic the close-range interactions that would be found in a polypeptide context. All residues are amino acids, so this is a reasonable approximation. The second variation is with a C-terminal di-methylamidation which I refer to "pre-proline." This modification is designed to mimic the context of a residue which precedes a tertiary amide. The name of the variation is attributed to the prominent affect that proline has within native proteins. However, this name applies to all non-canonical residues with a tertiary amide within the context of this study.

Generating the initial structures for each residue is the next critical step. The initial conformation of the residues, both variations, were generated via the Cambridge Structure Database conformer generator (*71*). Molecules with discrete rotamers like ring systems with multiple puckering states such as the pyrrolidine ring of proline are separated at this point, and each discrete rotamer will be sampled independently. Discrete rotamers must be considered independently because the adaptive sampling technique that will be described can only sample continuous degrees of freedom like torsions. At this point the critical degrees of freedom are identified for each residue by analyzing the conformers that have been generated. For example, the phi and psi torsions are the significant degrees of freedom of alpha amino acids. For residues with sidechains, the sidechain torsions are also considered to be significant degrees of freedom. Then for each discrete rotamer, the lowest energy conformer is identified by minimization and that conformer is the starting point for the adaptive sampling protocol.

2.3 BOLTZMANN ADAPTIVE SAMPLING

The protocol for sampling the entire potential energy landscape of the monomeric residues is described here. The motivation to use adaptive sampling was described in the previous chapter, but in short, the conformational space of molecules with even a few degrees of freedom is vast and I would like to find a way to intelligently sample that space by focusing on regions with low energy states. Therefore, I devised a scheme that guides sampling that is tuned for molecular systems and scales to arbitrary dimensions.

Adaptive sampling techniques are utilized to solve sampling and mesh generation challenges in sciences and engineering. The work of Nijolt et al (*72*) inspired me to recognize that adaptive sampling could be useful for my problem. However, I needed to develop a new version

of the technique because their work 1) was limited to 1 and 2 dimensions and 2) considered only local features of the search space.

The critical concept of the adaptive sampling scheme is the utilization of a Delaunay triangulation for maintaining the relationship between all of the points that have been sampled. The implementation of Delaunay triangulation was used from SciPy (*73*) without modification. A Delaunay triangulation is a technique that creates a mesh from an arbitrary set of points and, despite its name, can be generalized to arbitrary dimensions to create a mesh from irregular points. In the 2-dimensional case, a Delaunay triangulation can be visualized as a mesh of triangles where each vertex of a triangle is a point, which in our case is the dihedrals of a molecule that have been sampled. In higher dimensions, the triangles become the respective dimensional counterpart. For example, in 3-dimensions, each partition of the mesh will be a tetrahedron. In general, the partitions are known as simplices and each simplex will have N+1 vertices where N is the dimension of the simplex. This is implied from the definition, but each simplex of a Delaunay triangulation contains no points that have already been sampled. Also, the area of each triangle can be variable because the points do not need to be regular. Each point can also have meta-data associated with it. While that is important for the BAS technique, meta-data does not affect the calculations of the triangulation protocol. Other properties of simplices and Delaunay triangulations are not important for this work. However, I will note that I believe improvements of the protocol could be found by optimizing the Delaunay triangulation protocol.

Given a protocol that creates a mesh that relates all of the neighboring points of a given set of points, I can apply that to molecular sampling. For this protocol, all points of the Delaunay triangulation are the dihedral values of conformations that have been sampled. I store the conformation and the energy of the molecule that is associated with the point.

The sampling protocol is initialized by sampling a sparse regular grid with 7 points per dimension. This was found to sample the most common dihedral angles (-180°, -120°, -60°, 0°, 60°, 120°, 180°). Note that the first and last dihedral angle are identical molecularly, but Delaunay triangulation implementations that work in higher dimensions cannot account for periodicity.

Once a grid of conformations is sampled, I create the Delaunay triangulation that will guide further sampling. Every simplex of the Delaunay triangulation is evaluated with a loss function (see below). The simplex or simplices with the highest loss are then chosen. Once a simplex is chosen, I select the center of the simplex with small random perturbations to avoid artifactual sampling. That point corresponds to the next torsions that will be sampled. After evaluating that point by minimizing the conformation with the new dihedral angles, that point is added back into the Delaunay triangulation. This results in the simplex that was previously identified to break into multiple new simplices. The number of new simplices is dependent on the dimension. This process of identifying simplices, evaluating the centers of the simplices, and updating the Delaunay triangulation can be done one at a time or in batches. While performing this process one simplex at a time is optimal for sampling the simplex with the highest loss, I perform the process in small batches (<32) because the time to update the Delaunay triangulation is significant. I developed a dynamic controller that adjusted the batch size to ensure that 99% of the runtime was spent on evaluating the molecular energy. This controller was especially necessary in high dimensions because the Delaunay triangulation calculation scales poorly with number of dimensions and number of samples. If a controller is not used to adjust the batch size, the protocol will spend most of the time selecting which points to sample instead of sampling points.

The loss function that determines which simplices to sample next is the most critical part of this protocol. For this loss function, I considered what is important to molecular systems. Of

course, the most important factor to consider for molecular systems is the energy of the molecule. Therefore, the core of the loss function is the unnormalized Boltzmann probability. The Boltzmann probability is a useful metric because it is derived from first principles of statistical mechanics. Conformations with a high Boltzmann probability are the most probable and, consequently, the most important states of a molecule. Naïvely, taking the average Boltzmann probability of each vertex for each simplex would yield a loss function that always selected the simplex with the lowest energy. That would result in sampling low energy states but would continue to select subsections of the same simplex which results in narrow sampling. However, if you adjust the average Boltzmann probabilities of a simplex by the volume of the simplex, small simplices become penalized, which relieves the issue of exclusive sampling on the lowest energy states. Instead, a balance is formed that weights large regions of unexplored space and low energy regions together. I will refer to this approach as the volume adjusted Boltzmann probability.

Another critical step in the process is the evaluation of the energy of the conformation given a set of dihedral angles. The goal of this step is to accurately measure the energy of a molecule given a set of torsions. Broadly speaking, this process has two steps: 1) set the torsions of the major degrees of freedom to the values that are selected to be sampled, and 2) minimize the conformation of the molecule with those torsions constrained. I found that minimizing the conformation with constraints on the major degrees of freedom and without constraints on the minor degrees of freedom produced smooth potential energy surfaces that qualitatively recapitulated known potential energy surfaces like alanine. It was essential to perform a full minimization on the molecule because minute changes in bond lengths and bond angles can significantly affect the energy of a conformation, and when the torsions of the major degrees of freedom are modified, the ideal bond lengths and angles of the molecule can change These effects

are most pronounced with polar interactions like hydrogen bonding, which are prevalent in the amino acids that are included in this study. Without minimization, the potential energy surfaces of the molecules become sharper and can even become discontinuous. Lastly, I found that using the conformation of the nearest neighbor as the starting point before setting the major degrees of freedom and applying constraints improved the run time and convergence of the minimization trajectories. Presumably this is because the nearest neighbor point has similar perturbations to bond length and bond angle.

The entire protocol is represented visually as a process flow diagram in Figure 2.3. The iterative process of building the Delaunay triangulation, selecting points, and evaluating the energy of the molecule is repeating until the potential energy surface of the molecule has converged. An example of a converged potential energy surface for alanine is shown in Figure 2.4. The adaptive resolution of sampling is apparent when comparing the size of the simplices in the low and high energy regions of the potential energy surface. Regions of low energy have very dense sampling and small simplices while regions of high energy have very sparse sampling and large simplices. This phenomenon emerges directly from the loss function that balances the weight of energy and volume. Because this process does not use any heuristics that are specific to any particular molecule, this process can be applied to any molecule while maintaining the exceptional sampling performance.

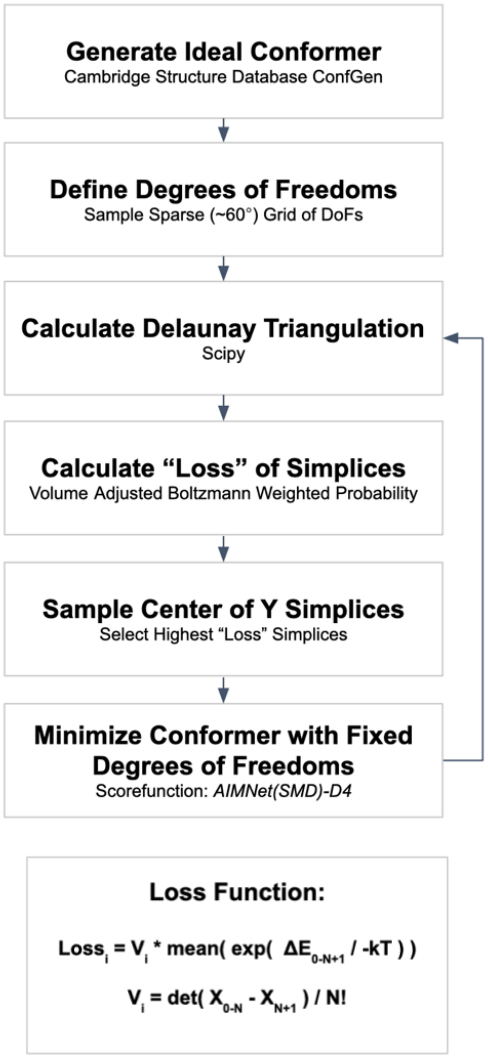


Figure 2.3. Process flow diagram for Boltzmann Adaptive Sampling protocol for monomeric residues. The process iterates indefinitely until sufficient sampling is achieved which is checked periodically. Heuristically I found that $10*10^N$ samples is sufficient where N is the total number of degrees of freedom. The number of centers sampled per iteration (Y) should be dynamically controlled to balance the time calculating the Delaunay triangulation and evaluating the energy of molecules. The loss function is the Volume Adjust Boltzmann Weighted Probability which is calculated by multiplying the volume of a simplex by the mean of the unnormalized Boltzmann probability for each vertex of the simplex.

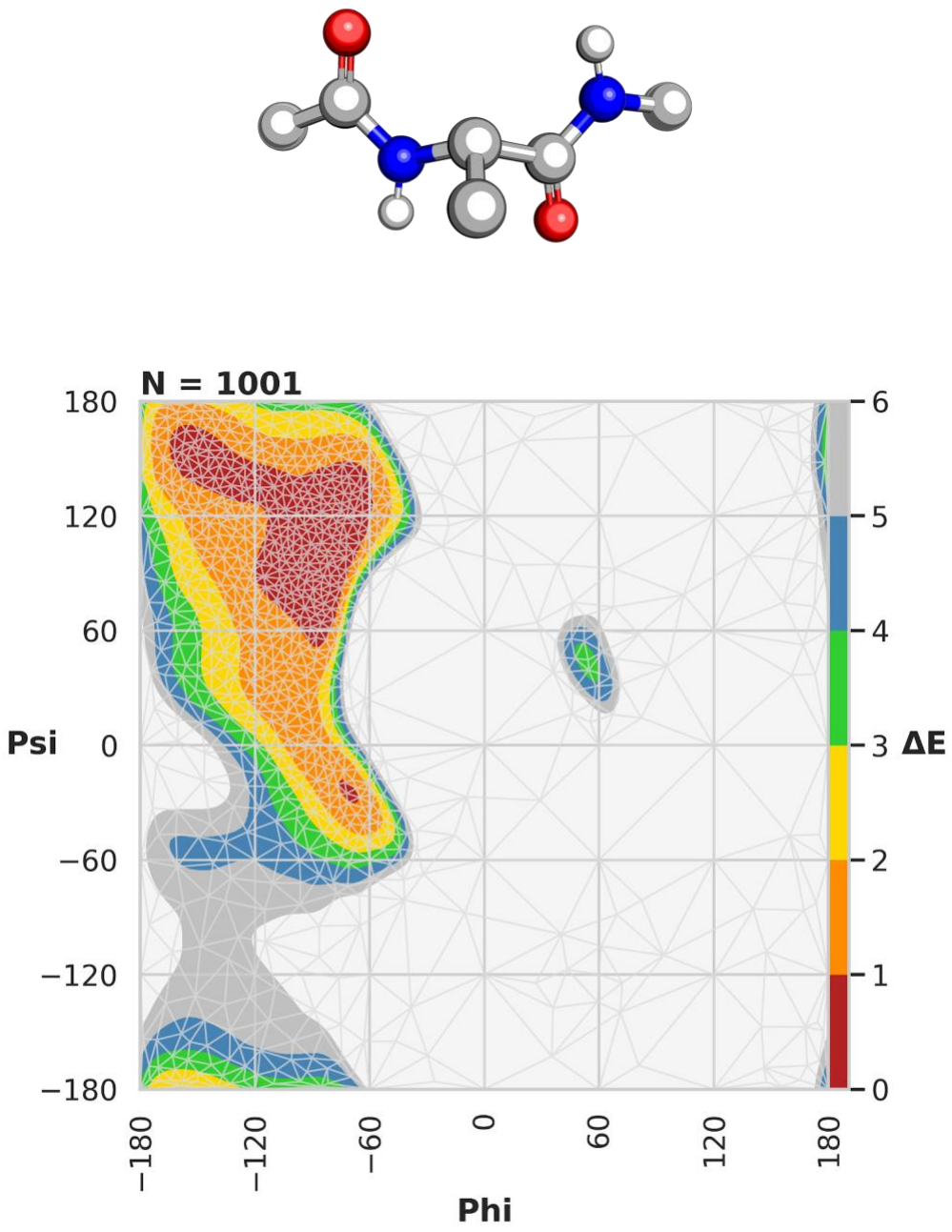


Figure 2.4. Ramachandran plot of alanine with N-terminal acetylation and C-terminal methyl amidation generated with Boltzmann Adaptive Sampling. Each vertex of the Delaunay Triangulation is a sampled point. The surface contour is generated with a linear interpolation of the resulting Delaunay Triangulation. The delta energy compared to the lowest energy conformation and it is in kcal/mol, Phi and Psi refer to the backbone dihedral angle of alanine, and the value of the backbone dihedrals is degrees.

The key metric for performance of the sampling is the rate of convergence of the potential energy surface. The potential energy surface can be generated by linearly interpolating the Delaunay triangulation. As more points are sampled in the Delaunay triangulation, a linear interpolation of the points will become more and more accurate at predicting new interpolated points. Once enough points have been sampled, the potential energy surface for a given molecule will converge and adding more points will not affect the potential energy surface. Once the potential energy surface is converged, the energy of any conformation can be accurately predicted.

Creating a mathematical model that quantifies this convergence was important. The Kullback-Libler (KL) divergences is a metric that compares the difference of probability distributions, and it is the perfect metric for this problem. The KL divergence explicitly measures the difference between two probability distributions. Therefore, large values of KL divergence mean that distributions are different and the value approaches 0 when distributions are identical. The KL divergence works with both discrete and continuous probability distributions. In my case, I generate a discrete probability distribution by sampling the potential energy surface with a uniform grid with 10-degree resolution. At each grid point I record the interpolated energy and translate that energy into a delta energy by subtracting from the lowest energy point. Because the interpolation is linear, the lowest energy point will be a point that I explicitly sampled, so finding the lowest energy point on the surface is trivial. Once I have the delta energy for the grid of interpolated points, I convert all of those values into normalized Boltzmann probabilities. Once I perform that protocol on one potential energy surface, I can perform it again on another potential energy surface. Then I can perform the KL Divergence calculation to directly compare those two distributions. In my case, I compared the potential energy surfaces of the same molecule as more points were sampled. With that information, I could identify when distributions were converged

because the KL divergence approached 0. Visual inspection of the potential energy surfaces confirmed that small values of the KL divergence (<0.05) corresponded to near identical distributions.

I performed a benchmark on this approach on various series of molecules to explore the effect that adaptive sampling has on improving the rate of convergence compared to uniform sampling. This experiment can be found in Figure 2.5. The experiment used a series of molecules which have two degrees of freedom, and which differ significantly in the shape of the potential energy surface. Specifically, I analyzed glycine, alanine, and alpha-aminoisobutyric acid. All of these residues are near identical, differing only in the number of methyl substitutions on the alpha carbon of an alpha amino acid. Residues with more alpha substitutions are more constrained, and therefore, the potential energy surface of those molecules will have a smaller region of low energy space. I hypothesized that the adaptive sampling would converge faster for molecules with more constraints because the resulting focused sampling will be able to generate a more accurate interpolation in the regions that matter most to the probability distribution, the low energy regions. Because high energy regions will have very low probabilities, those regions do not have a significant effect on the KL divergence. Conversely, any errors in the low energy regions have a significant effect on the KL divergence. Figure 2.5 demonstrates this phenomenon. All probability distributions, as sample counts increase, are compared to the final exhaustive probability distribution. I performed the protocol 10 times for each molecule and all replicates are plotted. First it is important to note that all trajectories that were sampled with adaptive resolution converge to a KL divergence value of 0 as the sample counts increase to the exhaustive 1000 samples. However, the result of uniform sampling varies among the different molecules. From left to right, the molecules increase in how restrictive the alpha methyl substitutions become, and uniform

sampling performs worse following the same trend. Glycine performs nearly identically with adaptive and uniform sampling. Both methods converge to the identical potential energy surface for all 10 trajectories by 1000 samples. This trend make sense if one considers the potential energy surface of glycine. It is a broad potential energy surface with a lot of low energy points. Uniform sampling applies well to this situation because there is only a limited amount of wasted sampling. However, this is not true for alanine which is the center pane. Alanine only populates approximately half of the region that glycine can populate. We see that adaptive sampling, which focuses on the regions of high interest, performs better than uniform sampling, which does not utilize that information. While the uniform sampling of alanine is able to converge at 1000 samples for all 10 trajectories, adaptive sampling quickly converges in 200-300 samples. That is even faster than the 400-500 samples needed for glycine, which makes sense intuitively because there is less space that needs to be sampled for alanine. The most interesting case is the final case with alpha-aminoisobutyric acid. This residue is the most constrained, and there is the most divergence in performance between the two methods. From the potential energy surface, it is apparent that alpha-aminoisobutyric acid only populates a very small fraction of the potential energy space compared to the other residues. This means that it is very critical to accurately map the energy of this region because all of the highest probability conformations will be localized there. Incredibly the uniform sampling never actually converges to the exhaustive potential energy surface because the resolution of sampling that can be done with uniform sampling and 1000 points is approximately ~10-degree bins, and that bin size is not small enough to accurately describe the narrow region of low energy space.

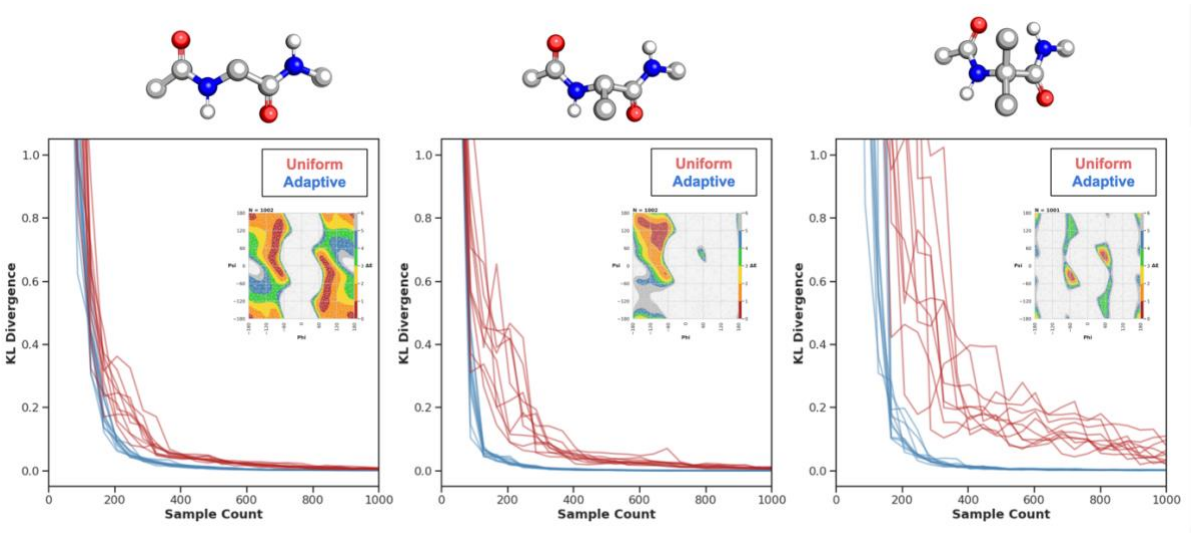


Figure 2.5. A plot of Kullback-Libler (KL) Divergence versus sample count for 10 trajectories of conformer ensemble sampling with (blue) and without (red) adaptive sample for glycine (left), alanine (center), and alpha-aminoisobutyric acid (right). The divergence of the ensemble is calculated between checkpoints during sampling and the final exhaustive ensemble. The potential energy surfaces for each residue are shown as an insert within each respective plot.

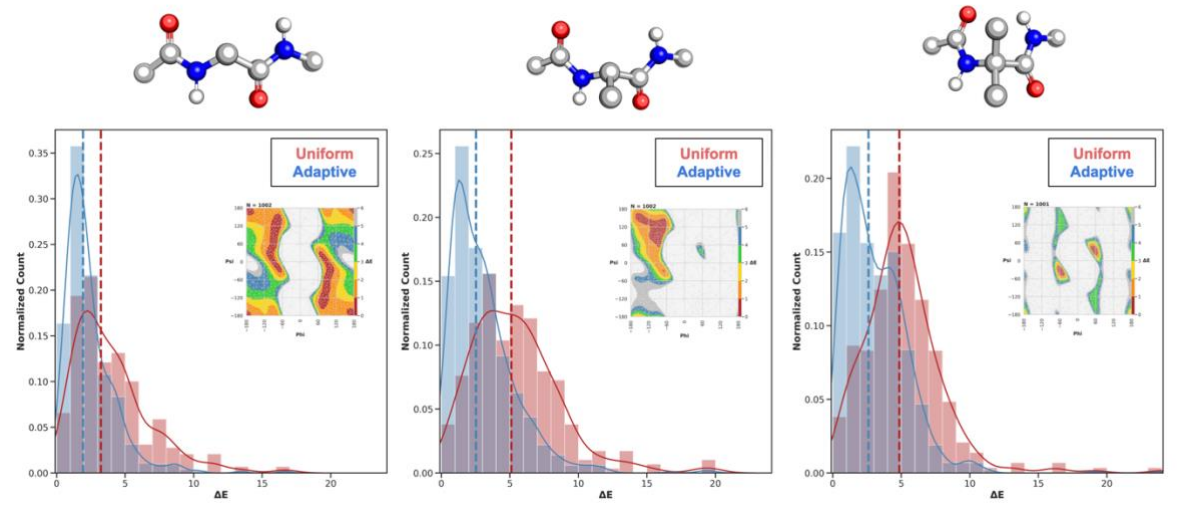


Figure 2.6. A plot of the normalized distribution of points with respect to energy in kcal/mol after conformer ensemble sampling with (blue) and without (red) adaptive sampling for glycine (left), alanine (center), and alpha-aminoisobutyric acid (right). The vertical line depicts the median value of the energy from the distribution. The potential energy surfaces for each residue are shown as an insert within each respective plot.

Another aspect of the same data is depicted in Figure 2.6. Here I have plotted the energy distribution of the sampled point in the final ensemble produced by the two different methods. It would be expected that the adaptive sampling would focus on regions with lower energy, and that trend can be identified from these graphs. The vertical dashed lines represent the median energy of all points. For all molecules, the median energy is shifted 1-2 kcal/mol for the adaptive sampling compared to the uniform sampling.

Another interesting experiment to consider is the effect of number of dimensions on the performance of adaptive and uniform sampling. This is studied in Figure 2.7. Because higher dimensions increase the total volume of the space, the regions of low energy space will become a lower fraction of a total space. This is similar to the effect that was recognized in 2-dimensions, and in that case, adaptive sampling was able to outperform uniform sampling significantly. Again, in this series of examples, the molecule with a potential energy surface that is more sparse shows significant improvements in performance with adaptive sampling over uniform sampling. In this case, performing the benchmark becomes much more computationally intensive in higher dimensions, so I reduced the number of replicates in the higher dimensions. The total number of samples required to reach convergence even with adaptive sampling increased 10-fold between dimensions. If sampling required approximately 10-degree binning to reach convergence, the expected scaling would be 36-fold for each subsequent dimension. Therefore, it is impressive that the adaptive sampling only requires a 10-fold increase in sampling. That corresponds to very crude sampling at 36-degree binning. At a uniform level of sampling, the potential energy surface is not close to converging.

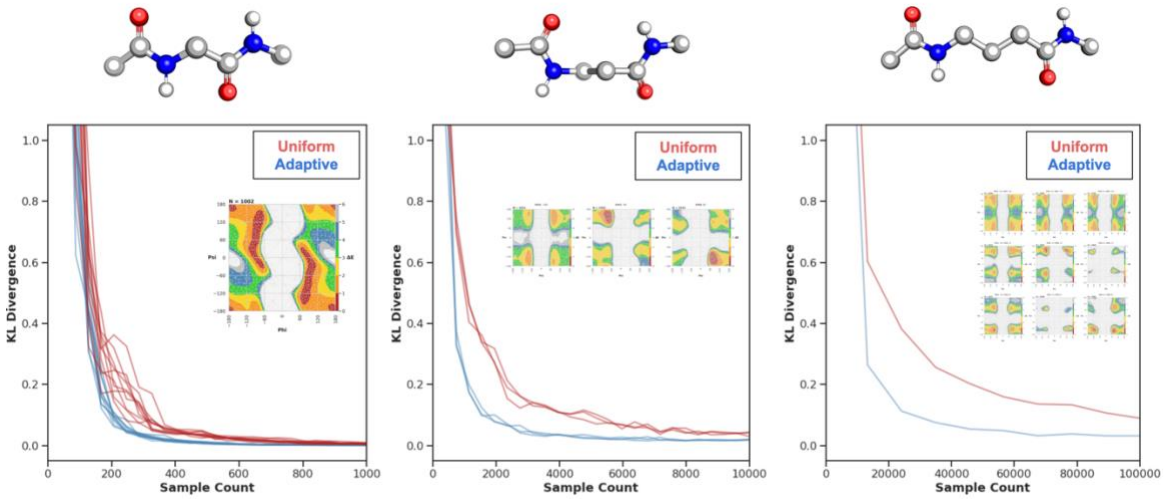


Figure 2.7. A plot of Kullback-Libler (KL) Divergence versus sample count for 10, 3, and 1 trajectories, respectively, of conformer ensemble sampling with (blue) and without (red) adaptive sample for glycine (left), beta-glycine (center), and gamma-glycine (right). The divergence of the ensemble is calculated between checkpoints during sampling and the final exhaustive ensemble. The potential energy surfaces for each residue are shown as an insert within each respective plot.

Overall, these results are very promising to demonstrate the significant performance upgrade that adaptive sampling can provide when performing conformational ensemble calculations. The ability for the protocol to adaptively focus on the most important regions of a molecular ensemble without *a priori* knowledge is impressive. This consistent and scalable performance is necessary for applying this protocol to a large gamut of non-canonical residues. This protocol could be applied to other molecular sampling protocols in which low energy states are important.

The major limitation of this approach is the scaling to high dimensions. While this method was sufficient for the molecules that I studied, any molecules with more than 6 degrees of freedom will become prohibitive due to the time to calculate the Delaunay triangulation. This effect is exacerbated by the fact that higher dimensions require more samples in general. These two effects together cripple the protocol. Figure 2.8 shows how significant this effect is. When a table in 6

dimensions has 1,000,000 points, it requires about 10 minutes to update the Delaunay triangulation with 100 new points. This is approximately equal to the time required to computationally evaluate the energy of 100 molecules. Therefore, half of the time will be spent managing the Delaunay triangulation instead of evaluating molecules. Improved methods for Delaunay triangulation could resolve this issue, but that is not within the scope of this work.

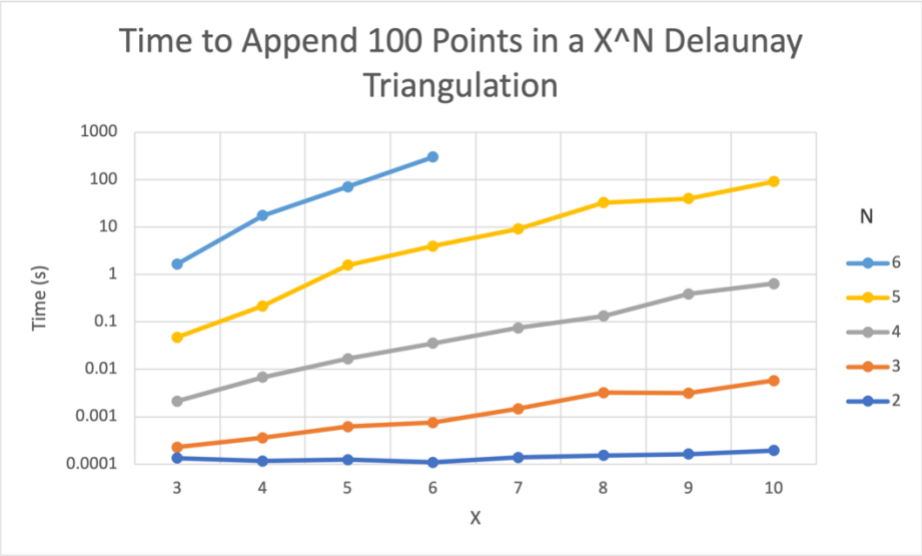


Figure 2.8. The time required to append 100 points into a Delaunay triangulation while varying the number of points in the set and the dimension of the points. X represents the size of the table and N represents the dimension of the table. For 6 dimensions, the calculations were stopped when the size of the table reached 1,000,000 points. Time is in seconds, and the axis is plotted with logarithmic scaling.

2.4 Sampling Secondary Structure Ensembles

Now that potential energy surfaces have been created for each monomeric residue, I can utilize that information to bias my sampling of secondary structures which is the ultimate goal of this process. Again, the secondary structures that I designed are built from di-peptide subunits with helical symmetry. Therefore, the process can be broken down into 4 simple steps which are performed on all possible combinations of di-peptides: 1) select a conformation for both residues biased by the Boltzmann probability of the conformation, 2) fuse those two residues together to create a di-peptide fragment, 3) propagate that di-peptide fragment with helical symmetry, and 4) evaluate and record the energy and conformation of the helically symmetric polymer. The first 3 steps of this process are visualized in Figure 2.9. This process is indefinitely iterated, and there is no way to analyze the convergence of the sampling. Because there are between 1,000 and 100,000 conformations per residue, exhaustively sampling the combinations of conformations is prohibitive. For the simplest di-peptide units, there are 1,000,000 unique pairs of conformations. Also, the time to evaluate the energy for the larger helically symmetric systems is significantly longer. With this in mind, I added an additional step that helped to explore the polymeric space more efficiently. After initially sampling secondary structures biased by the energy of the monomeric units, I instead used the energy of the polymer ensemble to bias the selection of new residue fragments. I can identify what torsion values are producing low energy structures in the context of a helically symmetric polymer. Then I can find which monomeric conformations have torsion values that are close in value and proceed to evaluate those monomer conformations in the context of a polymer as well. To avoid over sampling the low energy states, this selection process is a random selection that is biased by the Boltzmann probability. The total number of samples for each ensemble was about 50,000 conformations.

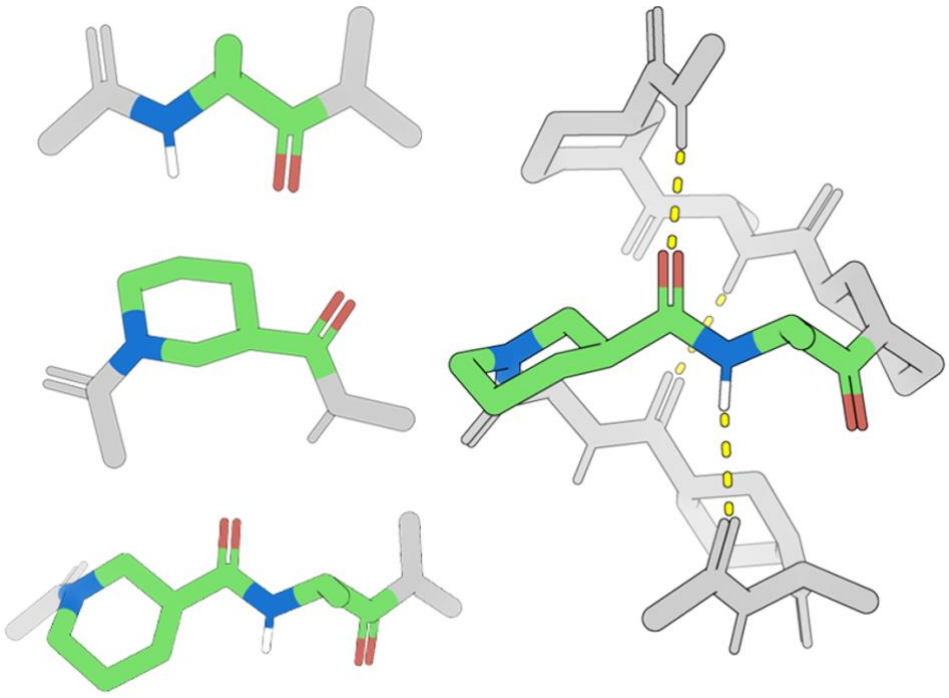


Figure 2.9. Computational assembly of monomers into di-peptide fragments and subsequent polymerization into a helically symmetric secondary structure.

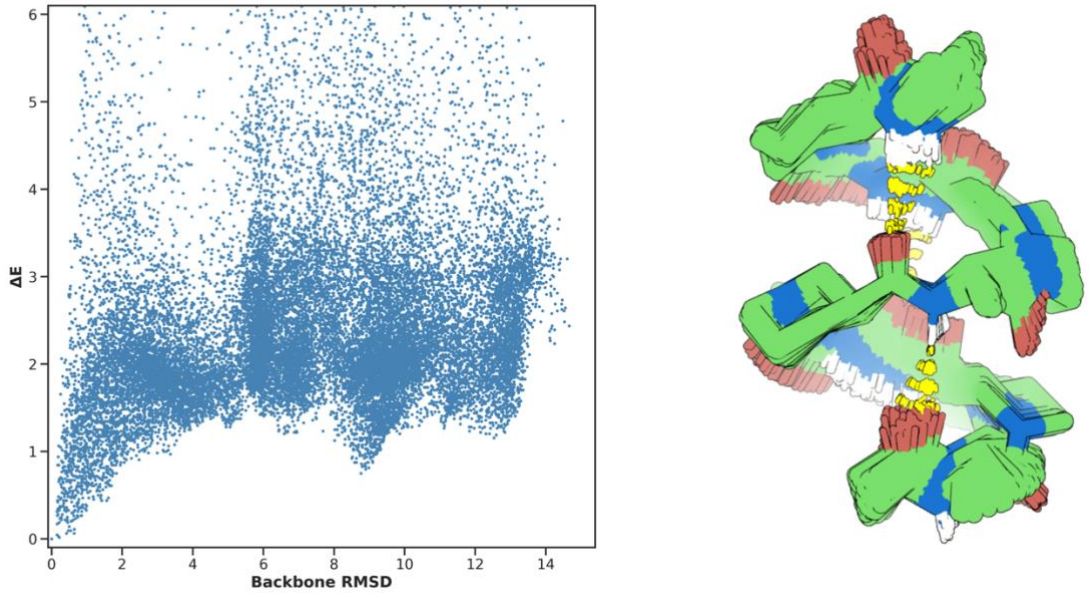


Figure 2.10. An example ensemble of helically symmetric conformations that is plotted to compare delta energy to the lowest energy state versus the root mean squared deviation (RMSD) of the backbone atoms to the lowest energy state. Energy is in kcal/mol/residue, and backbone RMSD is in angstroms. The model is the lowest energy structures aligned and overlayed.

Once the ensembles of helically symmetric polymers have been sampled, the goal is to evaluate the ensembles that fold into a unique low energy state. An example of an ensemble that folds into a unique state can be seen in Figure 2.10. The sampling of the protocol is quite exhaustive as is apparent in the ensemble of molecules that are overlayed. The aligned ensemble of conformations is all conformations that are below 0.5 kcal/mol/residue, though these conformations differ in backbone RMSD up to 1 Angstrom. Visually, it is apparent that these conformations are identical. Therefore, I created a filter to identify ensembles that do not have any conformations below 0.5 kcal/mol/residue that have an RMSD of greater than 2 Angstroms. I applied this filter to all 15,000 ensembles.

As a sanity check, this filter was able to identify the structure of known secondary structures which is shown in Figure 2.11. This secondary structure was an inspiration for this work and being able to accurately rediscover this structure is a promising step toward having confidence that the other predictions from this protocol are accurate enough to correctly identify previously undiscovered secondary structures.

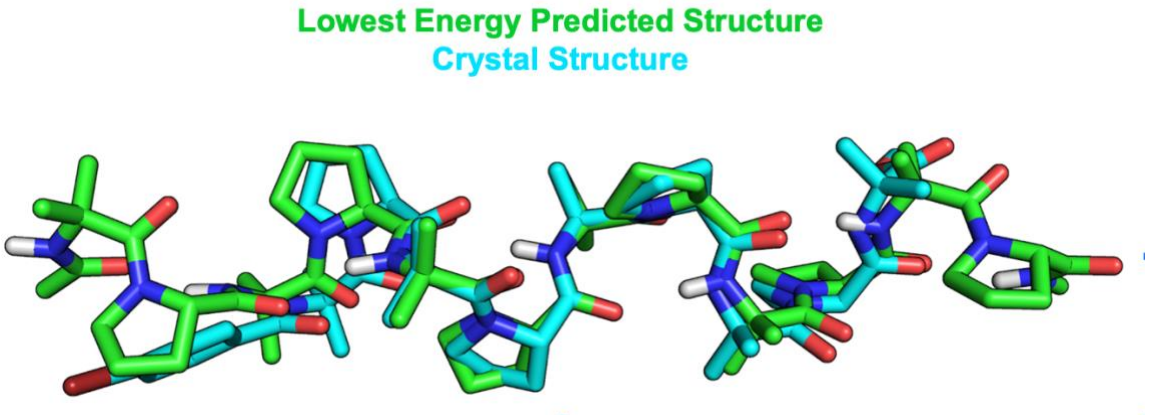
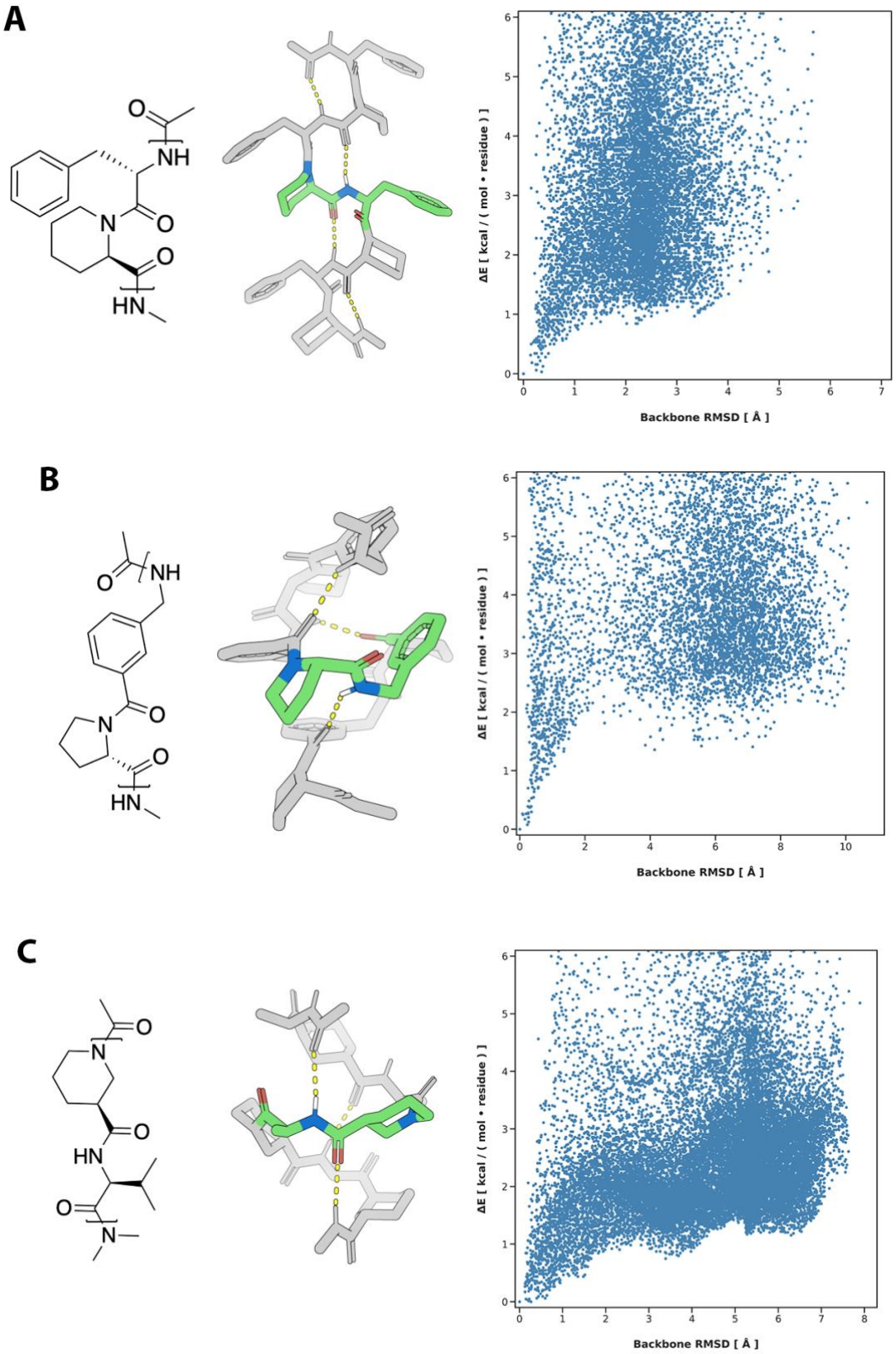
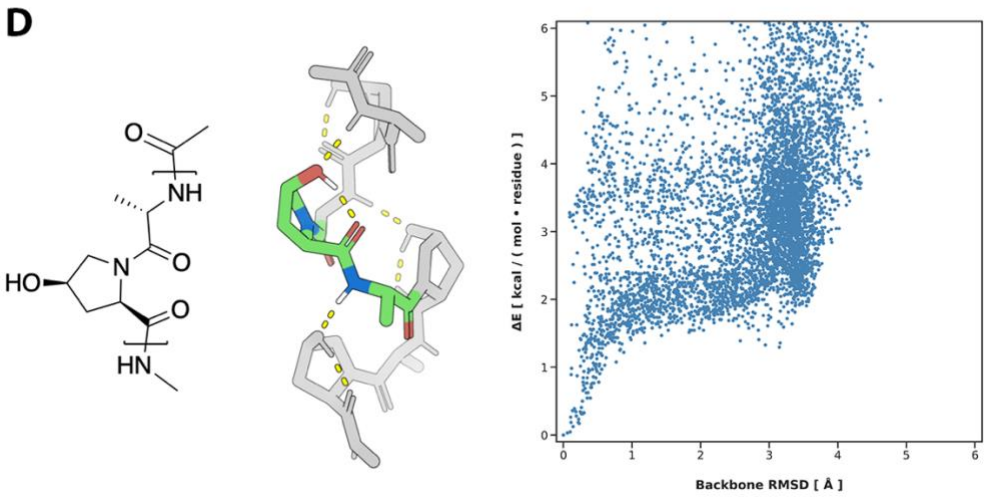


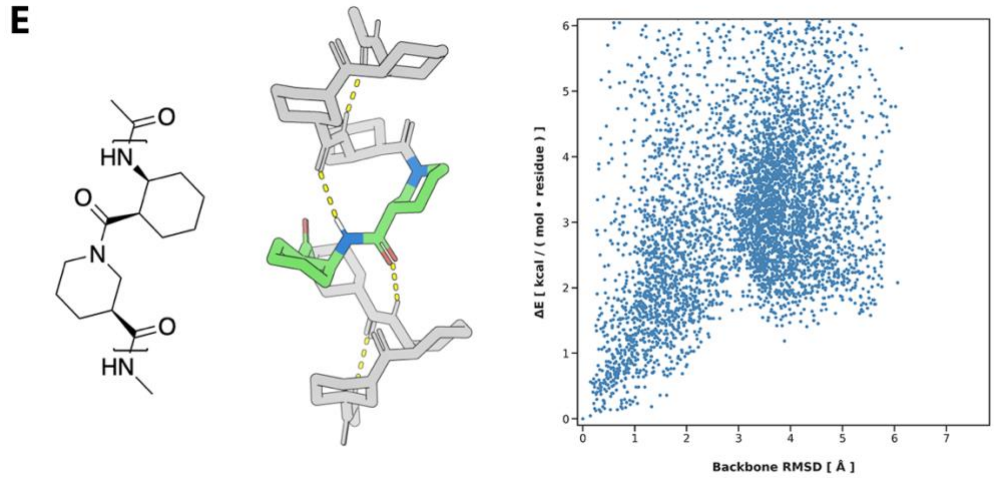
Figure 2.11. An example prediction of the lowest energy structure for a known non-canonical secondary structure. The sequence is a di-peptide unit of alpha-aminoisobutyric acid and proline. The crystal structure is from the CCDC (Deposition number: 1227627)

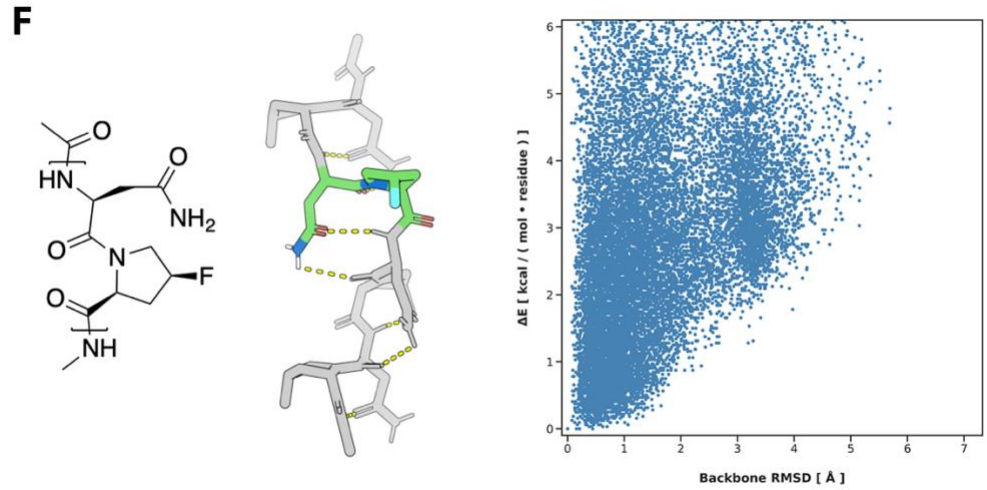
Using this protocol and filtering, I was able to identify 10 novel secondary structures which are previously undescribed. The full list of secondary structures and their information can be found in Figure 2.12. These structures vary in structure in interesting and unexpected ways. The pattern of hydrogen bonding, chemical functional groups, and shape differ between all of the structures, which highlights the generalizability of this protocol. Some molecules utilize sidechains to stabilize the secondary structures. This aspect of secondary structure design has not been explored previously because it would be very hard to predict these effects with human intuition alone. Also, exotic effects like halogen interactions showed up as being important for stabilizing conformations. Even molecules without hydrogen bonding could be detected as stable, and these types of molecules have been specifically excluded from comparable studies. With exhaustive sampling and avoiding human bias, very interesting molecules can emerge from familiar and unfamiliar building blocks. It is important to note that evaluating 15,000 combinations of di-peptides with all of these exotic residues would have been incredibly expensive. Instead, I can perform calculations which guide downstream experimental steps to avoid the costly experimental synthesis and characterization. Applying this protocol to discover new secondary structures from new residues is a promising new direction.



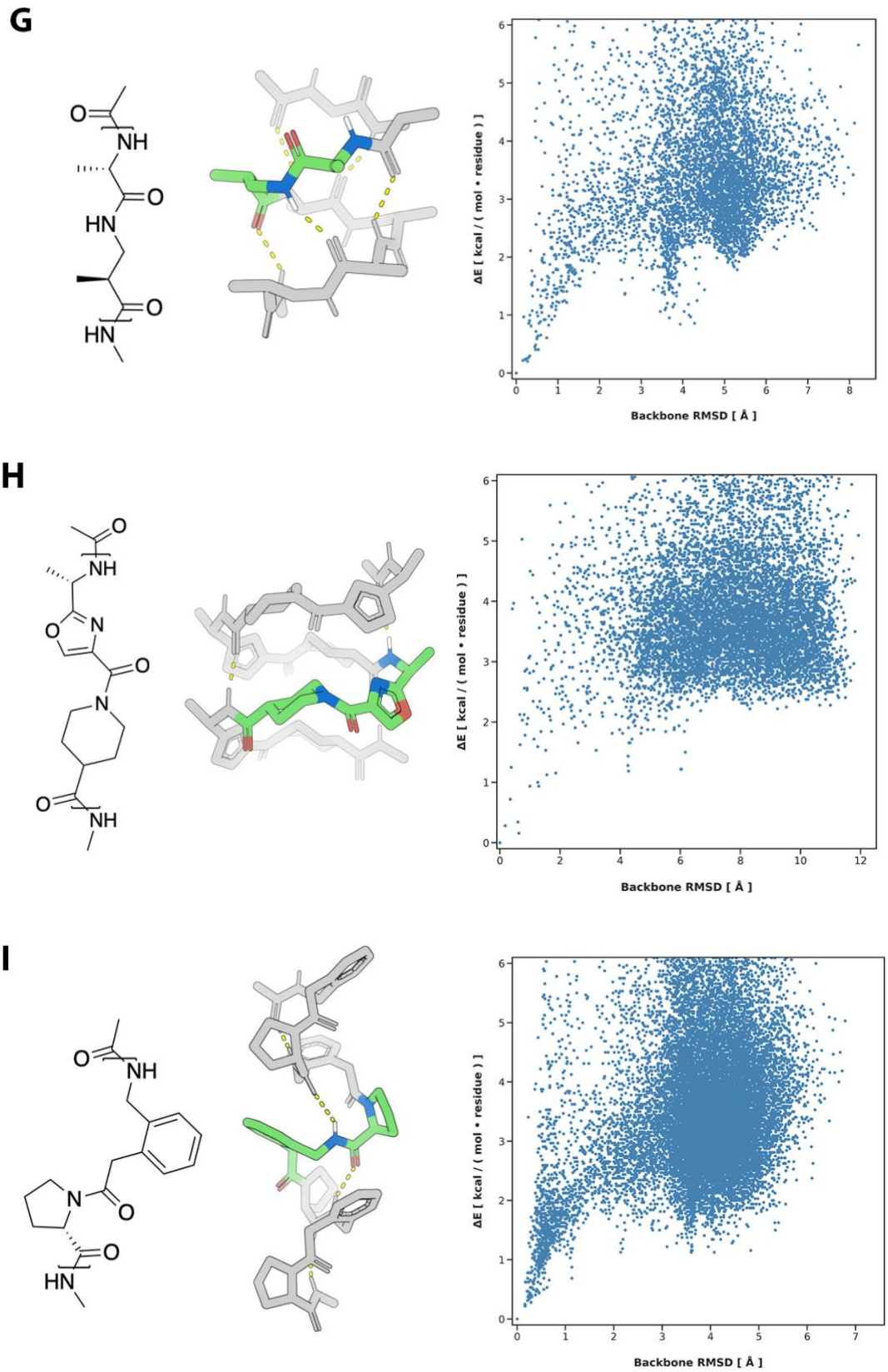
Caption on following page.







Caption on following page.



Caption on following page.

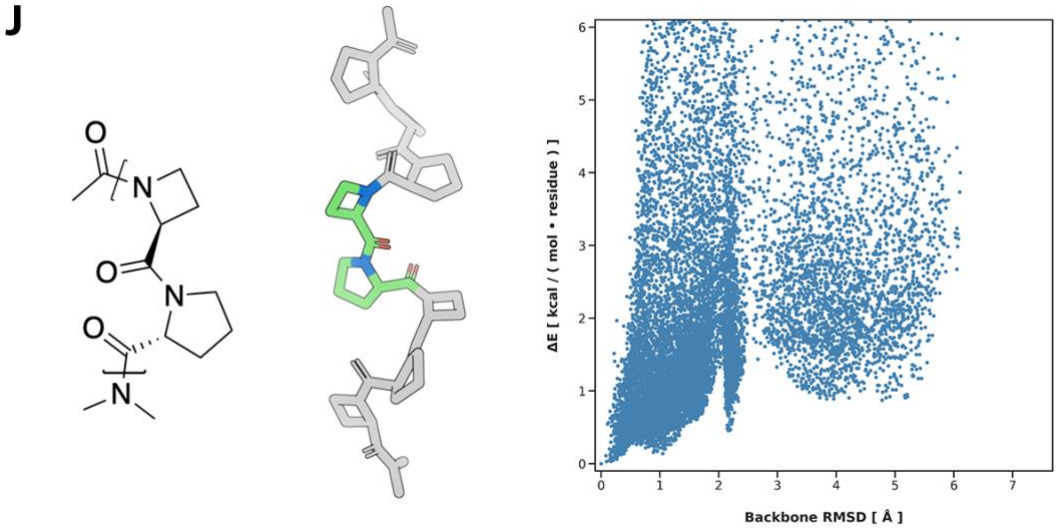


Figure 2.12. The complete set of novel secondary structures (A-I) that were identified by this computational protocol. In total, 10 secondary structures were identified. All secondary structures are polymers of di-peptide subunits. These subunits are draw as chemical diagrams, and the polymer subunit is delineated by the brackets. The lowest energy conformation of the molecules is also shown. The di-peptide subunit is shown in green, and the rest of the molecule is shown in gray to improve visualization. Hydrogen bonds are highlighted with yellow dashed lines. The conformational ensemble is plotted as delta energy per residue in kcal/mol versus the backbone atom RMSD to the low energy state.

Chapter 3. EXPERIMENTAL CHARACTERIZATION OF NOVEL SECONDARY STRUCTURES

Experimentally characterizing the structure of these novel secondary structures is a challenging but critical portion of this work. In this thesis chapter, I will focus on the characterization of the secondary structure A (ssA) from Figure 2.12 because this peptide has been nearly completely characterized. The characterization of all of the novel secondary structures is ongoing work.

Broadly, the characterization of these secondary structures includes x-ray crystallography, nuclear magnetic resonance (NMR), circular dichroism (CD) spectrophotometry, and atomic force microscopy (AFM). Because the secondary structures are designed with infinite symmetry, there is no clear size to design the molecules. However, I decided to synthesize the smallest possible version of the peptides that maintains all of the key hydrogen bonding contacts that promote structure because smaller peptides are easier to crystallize and evaluate on NMR. Also, many of these secondary structures are non-polar and consequently have poor solubility in water. While water solubility is not an ultimate criterion, I decided that improving the water solubility of the secondary structures would allow me to treat all peptides identically.

These decisions have drawbacks though. Small peptides might not form secondary structures until the length of the secondary structure reaches a critical size. Also, focusing on water solubility may have been a mistake as well. Water is a polar medium which competes with intramolecular hydrogen bonding. Many of these novel secondary structures have intramolecular hydrogen bonding which is critical for structure formation. Other solvents could improve the strength of these hydrogen bonds and consequently improve the folding of the secondary structures. Also, for future applications of these secondary structures like materials applications, using water as a

solvent could actually be detrimental. Exploration in both of these spaces are reasonable considerations for future characterization.

The exact chemical structure of secondary structure A (ssA) that was synthesized and characterized can be found in Figure 3.1 Notice that this peptide differs from the computational model in 3 ways. First, the N-terminus has a glycine with a free amine. This was designed to mimic the N-terminal acetylation in the computational model and include a charged amine functional group to improve water solubility. Second, the phenylalanine residues have been modified to tyrosine, which allows spectroscopic analysis at A280 and improves solubility. Third, the C-terminus is amidated which is the easiest chemical modification that can be applied to the C-terminus to mimic a primary amide bond. Methyl amidation of the C-terminus would have been preferred, but that modification is not readily available from commercial vendors.

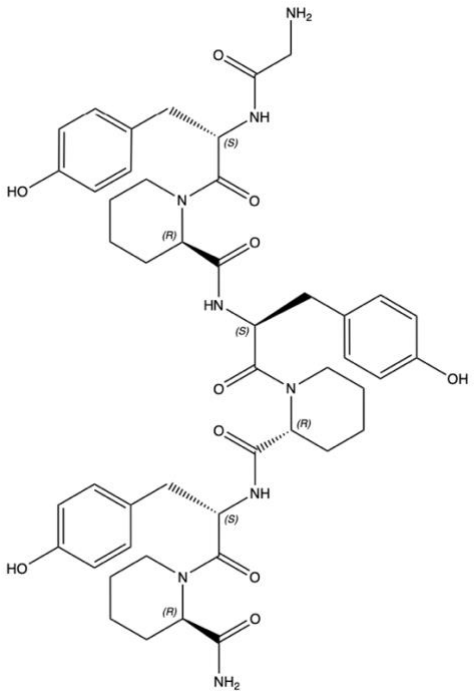


Figure 3.1. Chemical structure of experimentally characterized novel secondary structure A (ssA). The peptide is 3 repeats of the di-peptide motif l-Tyrosine and d-Pipercolic acid. The C-terminus is amidated and the N-terminus is glycinated.

3.1 CRYSTALLOGRAPHIC CHARACTERIZATION

Acquiring a crystal structure of these novel secondary structures is the ultimate goal for structural characterization. Fortunately, I was successful at acquiring a crystal of the novel secondary structure A (ssA). A photograph of the crystal and the crystal structure aligned to the design model can be found in Figure 3.2. The crystals were acquired by slow evaporation of a concentrated solution of the peptide in an equal mixture of water and acetonitrile. While small crystals formed quickly, these crystals could not be collected and diffracted effectively. Therefore, I let the solution evaporate for several weeks and exceptionally large crystals formed. These exceptionally large crystals were abundant, and many crystals could be collected for x-ray diffraction experiments. The crystal with the best diffraction data was used to solve the structure of the peptide. The crystal lattice of the peptide was found to be P 1 2_1 1.

Impressively, the structure of the molecule matched the design model nearly identically. The RMSD of the backbone CA atoms was 0.46 Angstroms which is exceptionally high for computationally designed molecules. Qualitatively, there is no deviation from the design model. Even the side chain rotamers which were hypothesized to be critical for the stability of the structure formed nearly identical contacts to the design model. The crystal packing of the molecules within the crystal lattice have limited contact. There is a terminus-to-terminus hydrogen bonding contact between subunits of the crystal lattice. There is limited sidechain to sidechain interactions, but there is a cluster of tyrosines that seem to form a hydrophobic pocket which could contribute to the stability of the crystal. There is a significant amount of water molecules in the crystal. Many of the interactions between molecules of the crystal structure are bridged by water molecules.

Further crystallographic characterization of the other peptides is on-going. Currently, there have not been any other peptides that have diffracted for structure determination.

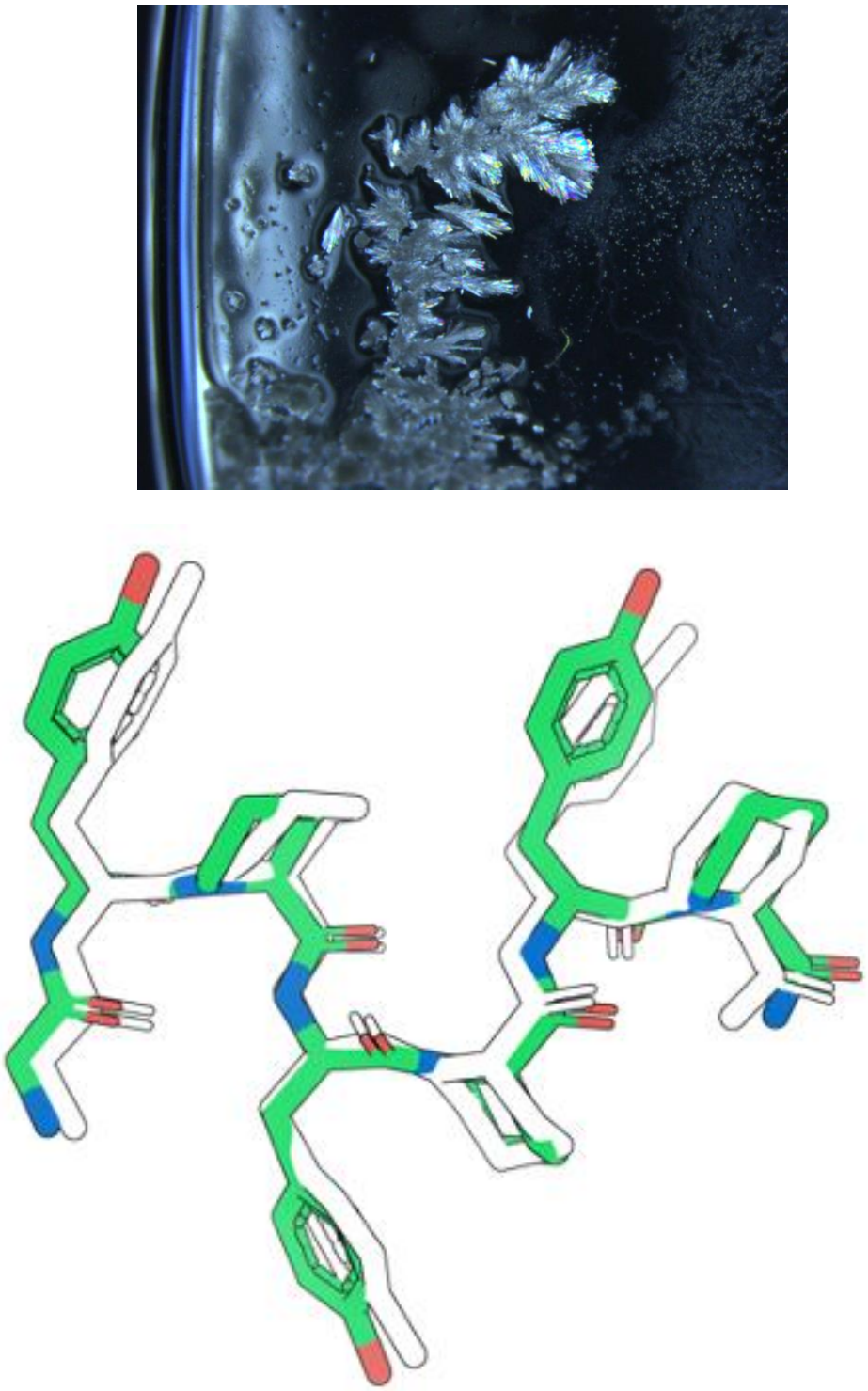


Figure 3.2. Top, a photograph of the peptide crystals acquired from an evaporating solution of water and acetonitrile. Bottom, crystal structure (gray) and design model (green) of secondary structure A (ssA) aligned by backbone CA atoms.

3.2 NUCLEAR MAGNETIC RESONANCE CHARACTERIZATION

The NMR characterization of the peptides has not yielded any structures yet because I do not have sufficient training in the collection and interpretation of NMR data. However, I have developed collaborations to facilitate this effort. My collaborators are Theresa Ramelot and Gaetano Montelione at Rensselaer Polytechnic Institute. The interpretation of all data shown is provided from conversations with them.

To date, the collaboration has been able to collect the spectrum of the secondary structure A (ssA) with high quality data that was collected at 800 mHz in water at 278K. Figure 3.3 shows the key NH region in 1D and 2D. This data suggests that the peptide has two or three conformations that are in slow exchange. The major structural species is approximately 80% of the population which is encouraging. Also, key nuclear Overhauser effects (NOEs) between the tyrosine residues and the pipecolic acids can be identified. We are hopeful that the full structure of this peptide can be solved. The overlap of spectrum due to the repetitive nature of the secondary structures does not seem to be an issue for this peptide.

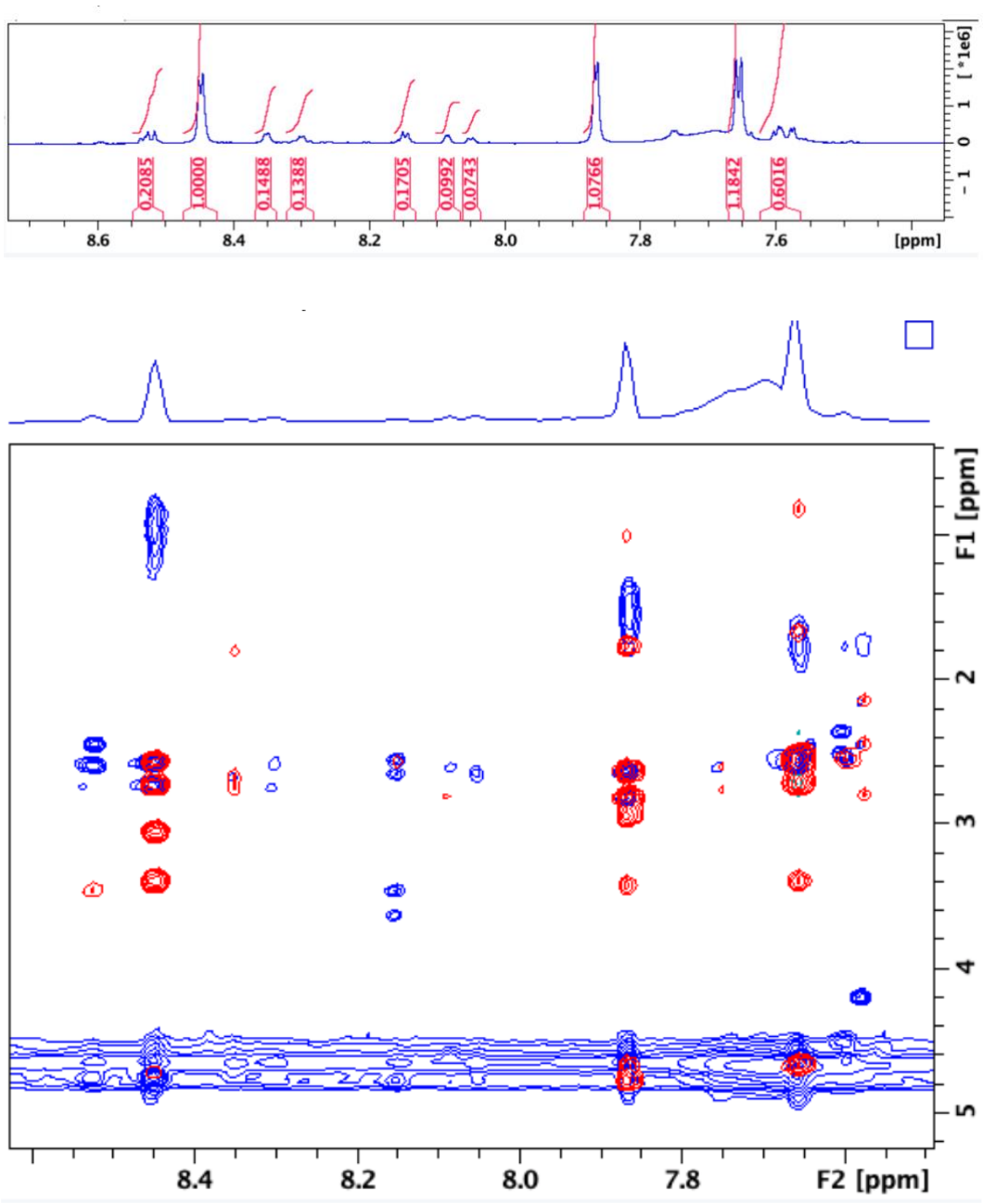


Figure 3.3. Top, the integrated 1D spectrum for secondary structure A (ssA) zoomed to the NH ppm region. Bottom, the 2D TOCSY (blue) and NOESY (red) spectrum of secondary structure A (ssA) zoomed to the NH ppm region. Both spectra are collected in water at 278K with an 800 mHz probe.

3.3 Circular Dichroism Characterization

Compared to NMR, CD is an easier experiment to conduct, and the technique can be useful for comparing modifications between similar peptides. While the technique cannot tell us about the exact structure of a molecule, it is useful for telling us how a structure changes with different variables.

I was interested in the effect of length of peptide has on the structure of the peptide. Therefore, for secondary structure A (ssA), I synthesized the peptide at multiple different lengths. Specifically, I synthesized it with 3, 6, and 9 repeats. The original peptide that I synthesized and crystallized had 3 repeats, so all of the new peptides are longer than the original one.

The results from CD are interesting. I collected the data with all of the peptide concentrations normalized by weight because I wanted all of the peptides to yield the same absorbance. Therefore, the molar concentration of the smaller peptides is higher than the largest peptide. However, when normalized by the number of amide bonds and aromatic groups in the peptides, which is what contributes to CD signal, all of the peptides have equal concentration. I used sodium bicarbonate solution because the longer peptides were not soluble in a pure water solution. All of the peptides had the same terminal modifications.

I found multiple interesting characteristics in this study which can be found in Figure 3.4. First, all of the peptides have qualitatively similar spectrum with a minimum at 210, maximum at 230, a shoulder at 250, and a minimum at 290. A strong signal in the near UV region of 250+ is interesting because it indicates that the tyrosine phenol group is contributing to the CD signal. This indicates that the tyrosine is structured in solution which fits the model. The other interesting thing to note is that the 6 repeat peptide has the strongest signal which is hard to explain.

In Figure 3.5, I wanted to analyze the effect on temperature on the structure of the peptide. I had hoped to see an unfolding of the peptide to determine the stability of the peptide. However, the effect of temperature did not have a significant effect on the structure of the peptide. I monitored the near UV signal for the thermal studies because I suspected that the sidechain residues would be the first part of the peptide to unfold. However, I did not detect cooperative unfolding. Instead, the loss in signal with temperature is linear and minimal. Even at 95C, the peptide maintained a similar near UV signal. The results from this study are inconclusive because this evidence could mean that the peptide is not folded, or this evidence could mean that the peptide is very well folded even at elevated temperatures.

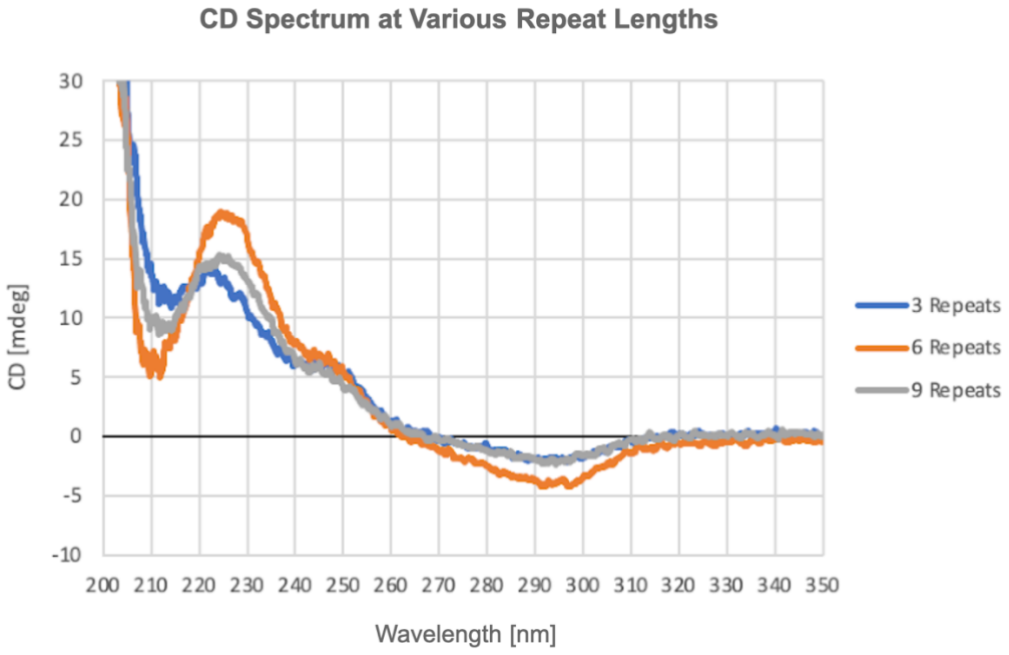


Figure 3.4. Full CD Spectrum of secondary structure A (ssA) in 10 mM NaHCO3 solution at 278K at various repeat lengths. The concentrations of the peptides were normalized by weight. All peptides were analyzed at 0.2 mg/ml which was calculated by the A280 of tyrosine.

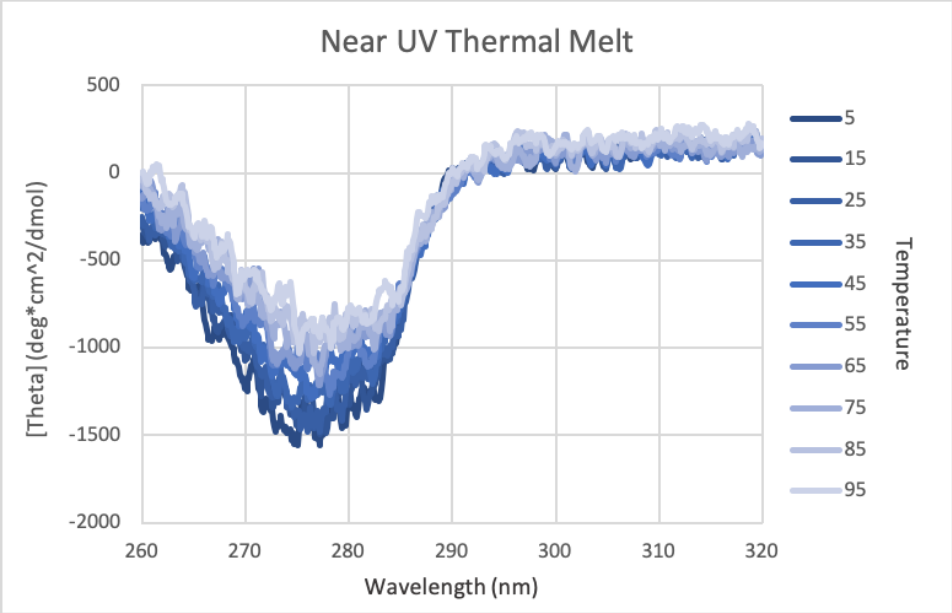


Figure 3.5. Thermal melt CD spectrum of secondary structure A (ssA) in DI water with 3 repeats. The concentration of the peptide was 0.2 mg/ml calculated by the A280 of tyrosine. Temperatures are in Celcius.

3.4 Atomic Force Microscopy Characterization

One of the long-term goals of this work is to develop new materials. Therefore, I was particularly interested in studying these secondary structures as helical polymers. Because the secondary structure can be broken down into a simple di-peptide repeat unit, I can synthesize this di-peptide unit and chemically ligate this unit into a full polymer. This method of synthesis does not control the length of the polymer, but the polymers are much larger than the typical fixed length peptides. The full protocol for synthesizing these polymers is as follows: 1) dissolve 10 mg of di-peptide subunit with free amine and acid termini in 10 ml of DCM and mix with 5eq of PyAOP and 200 microliters of DIEA followed by mixing the solution at 20C for 72 hours, 2) spin down the reactions and remove the supernatant, 3) 3x wash the solid polymer product with DCM, 4) 3x wash the solid polymer product with 50:50 water:acetonitrile, and 5) lyophilize. This protocol is similar to the reaction for the cyclization of peptides, but the reaction is run at a much higher

concentration to promote intermolecular reactions. The purification of the polymer has not been optimized beyond the washing of the insoluble product. If the peptide was soluble, a chromatography purification would be necessary. A visualization of this process can be found in Figure 3.6.

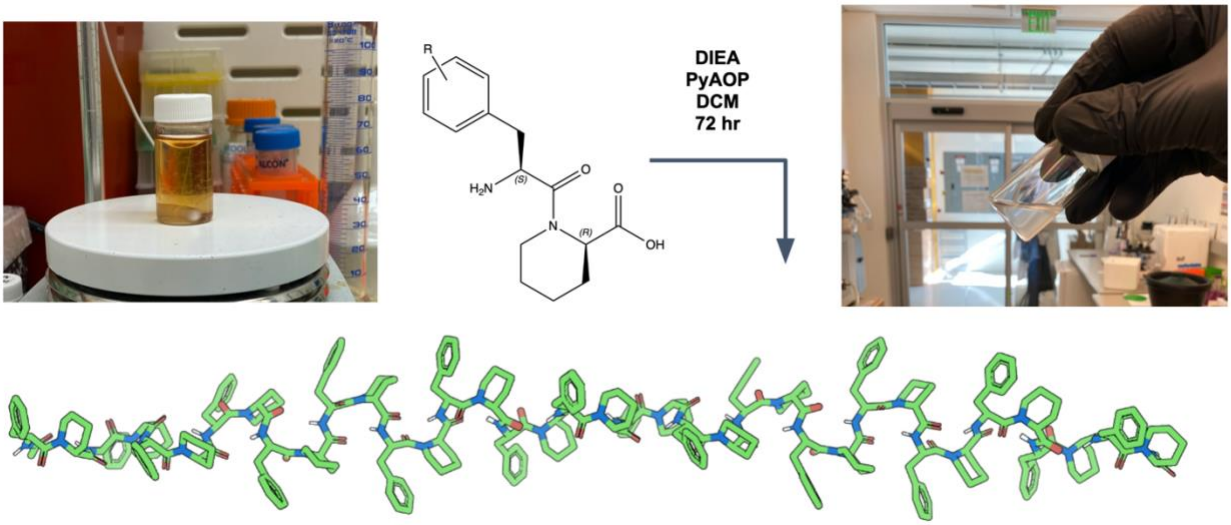


Figure 3.6. Example reaction to form polymers from di-peptide fragments, specifically secondary structure A (ssA) with a l-phenylalanine/d-pipecolic acid subunit. The reaction yields variable length polymers. The final product appeared to be an opaque suspension in 50:50 water:acetonitrile solution.

The polymer version of these secondary structures has an interesting advantage for characterization because it is possible to use AFM to image single molecules to see the behavior of the polymer without the convolution of bulk averaging. Also, another property that can be calculated from AFM imaging is the persistence length of the polymer. Rigid polymers will extend in a rod-like structure on surfaces. The rigidity of the polymer defines how much distance the rod-like structure will persist. I hypothesized that these, presumably, very stable secondary structures would exhibit exceptional persistence lengths. Therefore, AFM was a perfect experiment to test this hypothesis.

To visualize these polymers on AFM, I initiated a collaboration with Susrut Akkineni and James De Yoreo from Pacific Northwest Nation Laboratory. Susrut was able to successfully adhere the polymer onto the surface of highly ordered pyrolytic graphite (HOPG) from a 0.01mg/ml solution of 50:50 water:acetonitrile. Images of the polymer can be found in Figure 3.7 and Figure 3.8. Most notably, the polymers seem to assemble into bundles. The grooves of the assemblies match the dimensions of the expected polymer geometry, but it is apparent that multiple polymers laterally assemble. Most remarkably, the polymers appear to remain rigid over hundreds of nanometers, which is very rare for what is generally considered a flexible polymer like alkyl polyamides. Another interesting property to note is the epitaxial alignment of the polymer assemblies on the surface. Epitaxial alignment on a surface suggests that the polymer is structured, and the polymer is making a favorable contact with the surface that drives the epitaxial alignment. This is strong evidence that the polymer is not just rod-like but also has consistent structure at the atomic level that give rise to this ordered surface interaction.

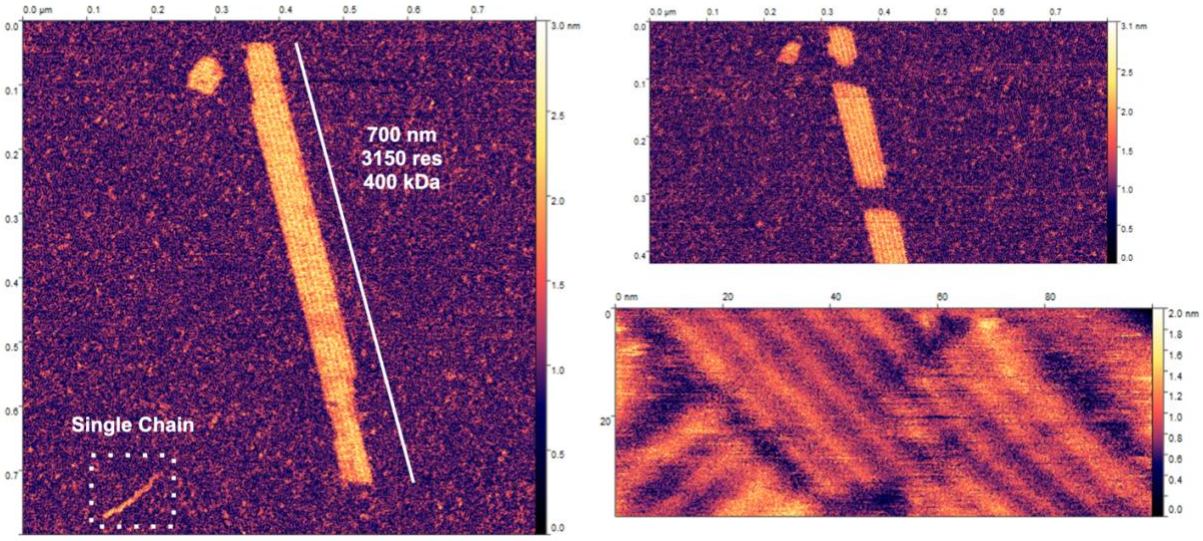


Figure 3.7. Various views of the polymeric version of secondary structure A (ssA) on drop cast at 0.01 mg/ml in a 50:50 water:acetonitrile solution on a HOPG surface. The height is in nanometers and the lateral axes are variable. The peptide appears to bundle into raft-like assemblies. The white line is a measurement of the polymer bundle. What is suspected to be a single chain is outlined in a dashed white box.

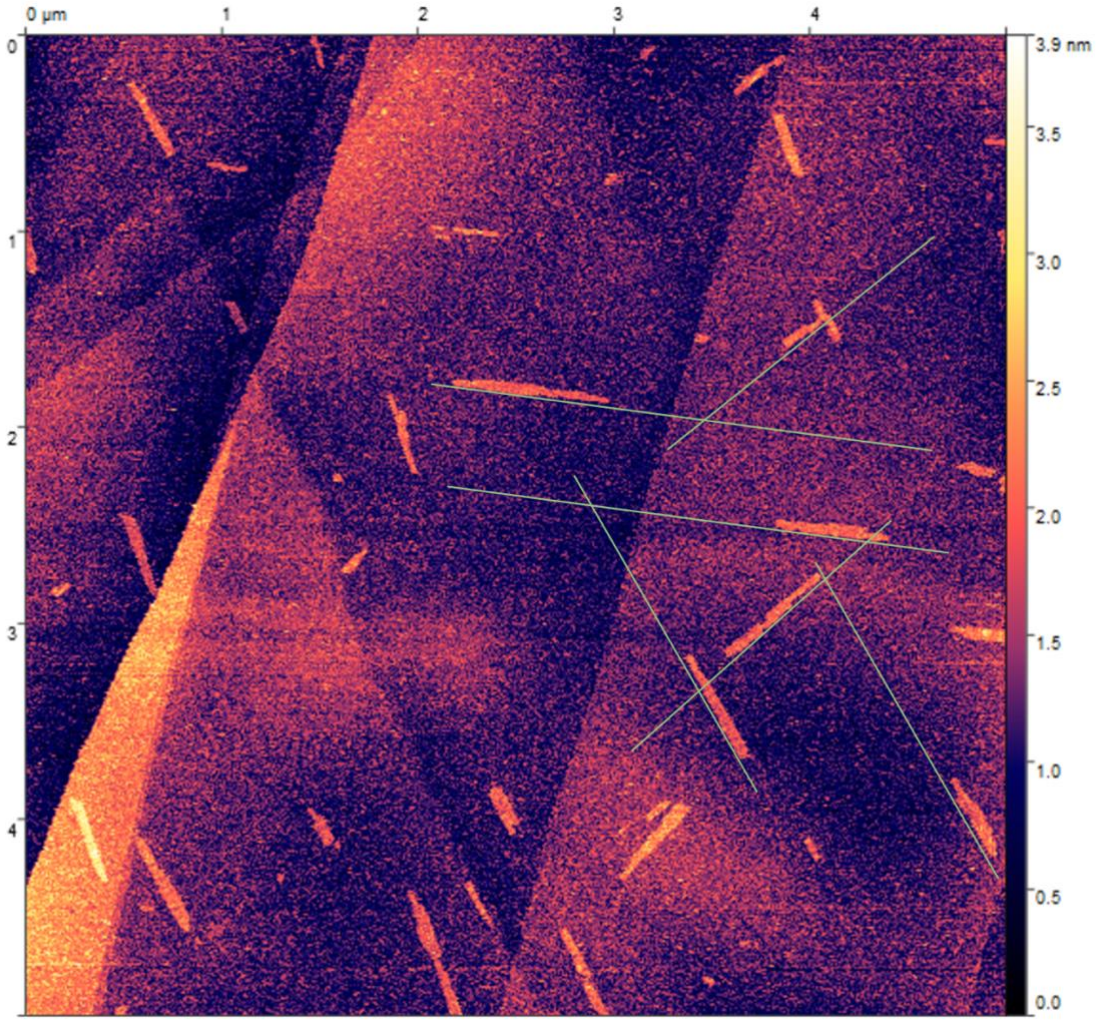


Figure 3.8. Wide view of the polymeric version of secondary structure A (ssA) on drop cast at 0.01 mg/ml in a 50:50 water:acetonitrile solution on a HOPG surface. The height is in nanometers and the lateral axes are in micrometers. The green lines are added to highlight the epitaxial alignment of the polymer bundles.

Chapter 4. CONCLUSION AND FUTURE DIRECTION

I hope that the work herein described is an inspiration to the future of molecular engineering efforts. Viewing molecular design beyond the constraints of protein engineering is a fundamental vision of this work, and I hope to have shared that vision with the reader. There is a large space of molecules that remain unexplored, and this study only begins to scratch the surface of what is possible. While I have begun to address some of the fundamental challenges that stand before the design of molecules from arbitrary building blocks, there are still challenges left unaddressed and directions that remain unexplored. Improving the performance of the adaptive sampling technique in higher dimensions would enable the design of larger residues with greater confidence. Also, expanding the set of residues beyond commercially available amino acids building blocks could be an interesting direction because the space of chemically feasible building blocks is much larger than what is readily available. Lastly, designing the assemble of higher order helices is also an incredibly interesting direction. New methods would need to be developed to detect the interchain contacts of higher order helices, but the fundamental sampling of individual helices is largely solved by this work.

I am excited by the potential future directions of this work. Although the work is not complete because I have yet to characterize many novel secondary structures that I computationally identified, I believe that the first secondary structure that has been experimentally characterized shows incredible promise given the data collected so far. Continuing to characterize these secondary structures as peptides and polymers is the most pressing future direction. However, I am beginning to look into other future directions that I will describe here.

Rigid polymers with precise atomic structure are potentially incredibly useful, specifically when considering my area of interest and future directions: conductance. The field of electrically

conductive polymers is interesting because it is broadly applicable to many challenges of humanity such as light harvesting devices, light emitting devices, and information processing devices. The properties of conducting polymers arise when chemical functional groups with delocalized electrons such as aromatic rings come in close contact with each other along a polymer chain. Using computational design to promote these interactions, similar to what I have done in this work, can yield conductive polymers. Another important consideration is efficiency. While polymers that form a delocalized electron system along the polymer chain are conductive, the performance of the conductivity is structurally dependent. If the delocalized electron system is unstable, the system will occasionally break which leads to loss in current and increases resistance. Therefore, I expect that designed polymers which attempt to improve these properties via computational design could significantly improve the critical performance metrics.

Other future directions for this work include the design of macromolecules with tertiary structure that are designed from the novel secondary structures described in this thesis. However, I do not have a plan to continue in this direction because the applications of these molecules are not particularly evident. While macromolecules built from novel and stable secondary structures would be an important step for improving the reliability of molecular design in the future, the current challenges of macromolecular design like enzyme design and small molecule binder design can be approached with conventional protein design. Therefore, the aspect of novelty is diminished. However, I still suspect that using hyper stable non-canonical secondary structures for these macromolecular design challenges would improve the success rate.

4.1 Conducting Polymer Design

The design of a secondary structure with a particular property rather than an arbitrary, stable secondary structure is a distinctly different challenge. While it is technically

possible to perform similar calculations as those described in chapter 2 to find secondary structures with specific properties, there are ways to improve the efficiency of the algorithm.

In the case of conducting polymers, an aromatic motif that stacks with helical symmetry is desired. Therefore, I can begin the process with that motif and work backwards to find a backbone that will support that molecular arrangement. I decided to design with pyrene for this work because pyrene is a classic molecule for conducting polymer design, and there are commercially available non-canonical residues that include the functional group. Specifically, there is an alpha amino acid with a pyrene sidechain and there is a peptoid residue with a pyrene sidechain. A residue with pyrene in the backbone is not available, but that would be an interesting variation to also consider if custom synthesis is an option. Examples of pyrene stacking with helical symmetry can be found in Figure 4.1.

There are limited degrees of freedom that must be considered when generating this molecular arrangement. If the vertical dimension along the pyrene column is considered the Z dimension, the pyrenes can be displaced in the X and Y dimensions which produces an offset between each layer. Also, the pyrene molecules have another degree of freedom which is the relative rotation between layers. The Z dimension is fixed because the ideal distance for conductance of pyrene molecules is known. However, the relative orientation between pyrene molecules that leads to conductance is not known, and there are many different orientations that can emerge from this sampling technique. There are orientations when the displacement in the X and Y direction is large that eliminate the contacts between the subsequent layers of pyrene, and those configurations must be filtered away because they are irrelevant for conducting polymer design.

Once I have generated many examples of assemblies of stacking pyrene, I can use these as starting points to find polymer backbones that can bridge the layers. I perform this action in two steps. First, I build an ensemble of backbone atoms from the pyrene molecules by aligning the sidechain of the non-canonical residues that contain pyrene onto each pyrene in the array of pyrenes. I generate the ensemble of conformations using the same techniques described in chapter 2. Second, I use the database of conformations that I calculated for the 135 non-canonical amino acids from chapter 2 to find residues that bridge the gap between layers of the pyrene residues. This is performed via a hash-based loop closure method. The fragments that I considered to bridge the gap are single and double residue fragments. Examples of closures using this method can be seen in Figure 4.2.

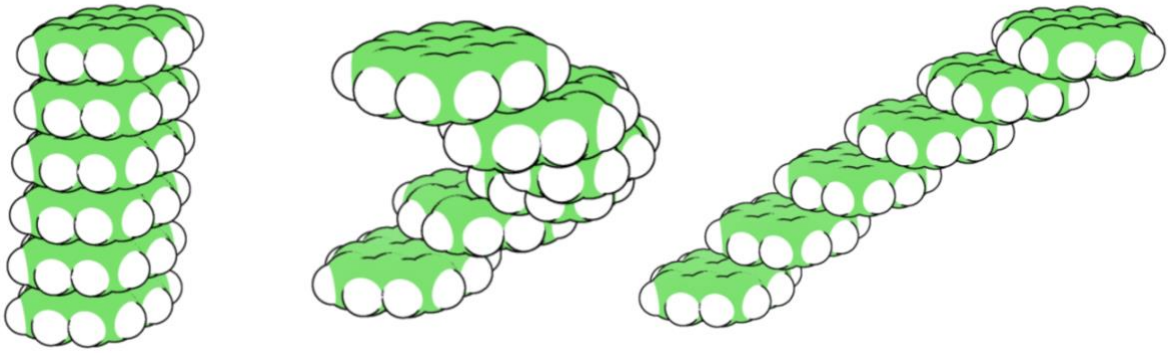


Figure 4.1. Examples of helically symmetric orientations of pyrene stacking generated via computational sampling by varying the offset and relative orientation between each layer. Carbons are shown as green spheres, and hydrogens are shown as white spheres.

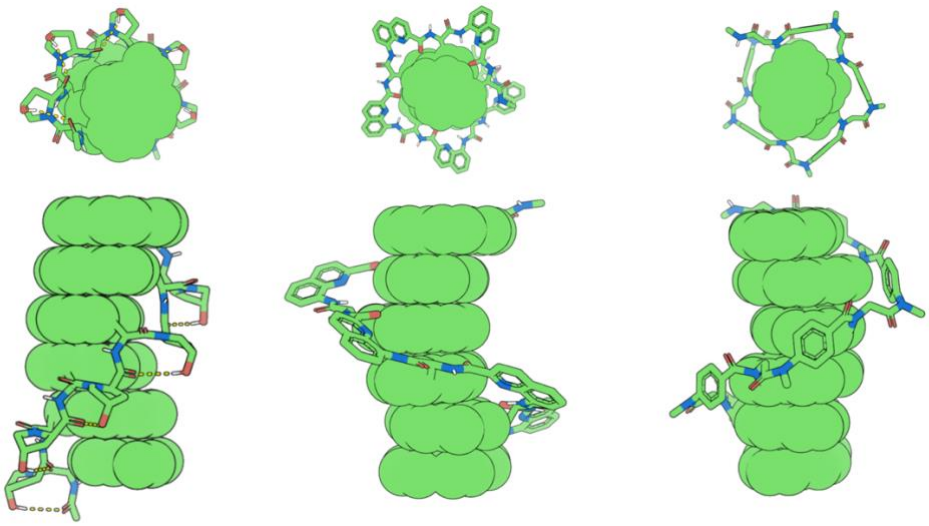


Figure 4.2. Examples of designed helically symmetric polypeptide chains that orient pyrene stacking along the polymer chain. The pyrene column is shown in green spheres. The backbone atoms that bridge the layers are shown in sticks and colored traditionally. All non-polar hydrogens are hidden

4.2 Final Remarks

Molecular engineering is still an engineering field in infancy. However, I believe that the future of molecular engineering will grow in extraordinary and unpredictable directions. Engineering the structure molecules with atomic precision seemed impossible not long ago. Like the progression of the various other fields of engineering, given 100 years of development, the reach of molecular engineering will expand to all aspects of life. For example, when the lightbulb was first invented, no one could imagine computer chips that followed. I suspect the same is true for molecular engineering.

A full repository of my software can be found at: https://github.com/atom-moyer

BIBLIOGRAPHY

1. D. L. Nelson, M. M. Cox, *Lehninger principles of biochemistry*. (W.H. Freeman ; Macmillan Learning, New York Houndmills, Basingstoke, ed. Seventh edition., 2017), pp. xxxiv, 1172, AS1134, G1120, I1145 pages.
2. D. Voet, J. G. Voet, C. W. Pratt, *Fundamentals of biochemistry : life at the molecular level*. (John Wiley & Sons, Hoboken, NJ, ed. Fifth edition., 2016), pp. 1 volume (various pagings).
3. A. Fersht, *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. Series in structural biology (World Scientific, New Jersey, 2017), pp. xxii, 631 pages.
4. D. J. Barlow, J. M. Thornton, Helix geometry in proteins. *J Mol Biol* **201**, 601-619 (1988).
5. K. A. Bolin, G. L. Millhauser, α and 310: The Split Personality of Polypeptide Helices. *Accounts of Chemical Research* **32**, 1027-1033 (1999).
6. A. A. Adzhubei, M. J. Sternberg, Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol* **229**, 472-493 (1993).
7. B. Brodsky, A. V. Persikov, Molecular structure of the collagen triple helix. *Adv Protein Chem* **70**, 301-339 (2005).
8. B. Wang, W. Yang, J. McKittrick, M. A. Meyers, Keratin: Structure, mechanical properties, occurrence in biological organisms, and efforts at bioinspiration. *Progress in Materials Science* **76**, 229-318 (2016).
9. M. W. Vetting *et al.*, Pentapeptide repeat proteins. *Biochemistry* **45**, 1-10 (2006).
10. T. J. El-Baba, D. R. Fuller, D. A. Hales, D. H. Russell, D. E. Clemmer, Solvent Mediation of Peptide Conformations: Polyproline Structures in Water, Methanol, Ethanol, and 1-Propanol as Determined by Ion Mobility Spectrometry-Mass Spectrometry. *J Am Soc Mass Spectrom* **30**, 77-84 (2019).
11. R. P. Riek, R. M. Graham, The elusive π-helix. *J Struct Biol* **173**, 153-160 (2011).
12. K. Neupert-Laves, M. Dobler, The crystal structure of a K+ complex of valinomycin. *Helv Chim Acta* **58**, 432-442 (1975).
13. B. A. Wallace, K. Ravikumar, The gramicidin pore: crystal structure of a cesium complex. *Science* **241**, 182-187 (1988).
14. M. Ohnishi, D. W. Urry, Temperature dependence of amide proton chemical shifts: the secondary structures of gramicidin S and valinomycin. *Biochem Biophys Res Commun* **36**, 194-202 (1969).
15. L. L. Wang *et al.*, Conformations and molecular interactions of poly-γ-glutamic acid as a soluble microbial product in aqueous solutions. *Sci Rep* **7**, 12787 (2017).
16. L. PAULING, R. B. COREY, H. R. BRANSON, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37**, 205-211 (1951).
17. W. L. Bragg, J. C. Kendrew, M. F. Perutz, Polypeptide chain configurations in crystalline proteins. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **203**, 321-357 (1950).
18. L. PAULING, R. B. COREY, The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* **37**, 251-256 (1951).
19. A. RICH, F. H. CRICK, The structure of collagen. *Nature* **176**, 915-916 (1955).
20. D. W. Urry, The gramicidin A transmembrane channel: a proposed pi(L,D) helix. *Proc Natl Acad Sci U S A* **68**, 672-676 (1971).
21. S. H. Gellman, Foldamers: A Manifesto. *Accounts of Chemical Research* **31**, 173-180 (1998).
22. C. W. Hume, USPTO, Ed. (EI Du Pont de Nemours and Co, USA, 1937), vol. 2130948.
23. **S. L. Kwoleck**, USPTO, Ed. (**E. I. du Pont de Nemours and Company**, USA, 1974), vol. **3,819,587**.
24. A. Violette *et al.*, N,N'-linked oligoureas as foldamers: chain length requirements for helix formation in protic solvent investigated by circular dichroism, NMR spectroscopy, and molecular dynamics. *J Am Chem Soc* **127**, 2156-2164 (2005).
25. R. Günther, H. J. Hofmann, Hydrazino peptides as foldamers: an extension of the beta-peptide concept. *J Am Chem Soc* **123**, 247-255 (2001).
26. X. Li, Y. D. Wu, D. Yang, Alpha-aminoxy acids: new possibilities from foldamers to anion receptors and channels. *Acc Chem Res* **41**, 1428-1438 (2008).
27. K. P. Fears *et al.*, High-performance nanomaterials formed by rigid yet extensible cyclic β-peptide polymers. *Nat Commun* **9**, 4090 (2018).
28. R. A. Turner, A. G. Oliver, R. S. Lokey, Click chemistry as a macrocyclization tool in the solid-phase synthesis of small cyclic peptides. *Org Lett* **9**, 5011-5014 (2007).
29. W. J. Moree, G. A. van der Marel, R. M. J. Liskamp, Peptides containing a sulfinamide or a sulfonamide moiety: New transition-state analogues. *Tetrahedron Letters* **32**, 409-412 (1991).
30. P. J. Almhjell, C. E. Boville, F. H. Arnold, Engineering enzymes for noncanonical amino acid synthesis. *Chem Soc Rev* **47**, 8980-8997 (2018).
31. T. A. Martinek, F. Fülöp, Peptidic foldamers: ramping up diversity. *Chem Soc Rev* **41**, 687-702 (2012).
32. S. Hecht, I. Huc, *Foldamers : structure, properties, and applications*. (Wiley-VCH, Weinheim, 2007), pp. xxii, 434 p.
33. G. Guichard, I. Huc, Synthetic foldamers. *Chem Commun (Camb)* **47**, 5933-5941 (2011).
34. L. K. Pilsl, O. Reiser, α/β-Peptide foldamers: state of the art. *Amino Acids* **41**, 709-718 (2011).
35. D. J. Hill, M. J. Mio, R. B. Prince, T. S. Hughes, J. S. Moore, A field guide to foldamers. *Chem Rev* **101**, 3893-4012 (2001).
36. C. M. Goodman, S. Choi, S. Shandler, W. F. DeGrado, Foldamers as versatile frameworks for the design and evolution of function. *Nat Chem Biol* **3**, 252-262 (2007).
37. D. Seebach, J. Gardiner, Beta-peptidic peptidomimetics. *Acc Chem Res* **41**, 1366-1375 (2008).
38. R. P. Cheng, S. H. Gellman, W. F. DeGrado, beta-Peptides: from structure to function. *Chem Rev* **101**, 3219-3232 (2001).
39. W. S. Horne, S. H. Gellman, Foldamers with heterogeneous backbones. *Acc Chem Res* **41**, 1399-1408 (2008).
40. F. Bouillère, S. Thétiot-Laurent, C. Kouklovsky, V. Alezra, Foldamers containing γ-amino acid residues or their analogues: structural features and applications. *Amino Acids* **41**, 687-707 (2011).
41. K. Möhle, R. Günther, M. Thormann, N. Sewald, H. J. Hofmann, Basic conformers in beta-peptides. *Biopolymers* **50**, 167-184 (1999).

42. C. Baldauf, R. Günther, H.-J. Hofmann, Helix Formation in α,γ- and β,γ-Hybrid Peptides: Theoretical Insights into Mimicry of α- and β-Peptides. *The Journal of Organic Chemistry* **71**, 1200-1208 (2006).
43. C. Baldauf, R. Günther, H.-J. Hofmann, Helix Formation and Folding in γ-Peptides and Their Vinylogues. *Helvetica Chimica Acta* **86**, 2573-2588 (2003).
44. P. G. Vasudev, S. Chatterjee, N. Shamala, P. Balaram, Structural chemistry of peptides containing backbone expanded amino acid residues: conformational features of β, γ, and hybrid peptides. *Chem Rev* **111**, 657-687 (2011).
45. D. Seebach *et al.*, Probing the Helical Secondary Structure of Short-Chain β-Peptides. *Helvetica Chimica Acta* **79**, 2043-2066 (1996).
46. D. H. Appella *et al.*, Residue-based control of helix shape in beta-peptide oligomers. *Nature* **387**, 381-384 (1997).
47. T. D. W. Claridge *et al.*, 10-Helical conformations in oxetane β-amino acid hexamers. *Tetrahedron Letters* **42**, 4251-4255 (2001).
48. S. Abele, P. Seiler, D. Seebach, Synthesis, Crystal Structures, and Modelling of β-Oligopeptides Consisting of 1-(Aminomethyl)cyclopropanecarboxylic Acid: Ribbon-Type Arrangement of Eight-Membered H-Bonded Rings. *Helvetica Chimica Acta* **82**, 1559-1571 (1999).
49. T. Hintermann, K. Gademann, B. Jaun, D. Seebach, γ-Peptides Forming More Stable Secondary Structures than α-Peptides: Synthesis and helical NMR-solution structure of the γ-hexapeptide analog of H-(Val-Ala-Leu)2-OH. *Helvetica Chimica Acta* **81**, 983-1002 (1998).
50. B. Di Blasio *et al.*, Structural characterization of the .beta.-bend ribbon spiral: crystallographic analysis of two long (L-Pro-Aib)n sequential peptides. *Journal of the American Chemical Society* **114**, 6273-6278 (1992).
51. J. L. Kulp *et al.*, Vibrational circular-dichroism spectroscopy of homologous cyclic peptides designed to fold into β helices of opposite chirality. *Biointerphases* **6**, 1-7 (2011).
52. S. H. Choi, I. A. Guzei, L. C. Spencer, S. H. Gellman, Crystallographic characterization of helical secondary structures in alpha/beta-peptides with 1:1 residue alternation. *J Am Chem Soc* **130**, 6544-6550 (2008).
53. Y. H. Shin, S. H. Gellman, Impact of Backbone Pattern and Residue Substitution on Helicity in α/β/γ-Peptides. *J Am Chem Soc* **140**, 1394-1400 (2018).
54. T. A. Martinek *et al.*, Effects of the alternating backbone configuration on the secondary structure and self-assembly of beta-peptides. *J Am Chem Soc* **128**, 13539-13544 (2006).
55. J. R. Stringer, J. A. Crapster, I. A. Guzei, H. E. Blackwell, Extraordinarily robust polyproline type I peptoid helices generated via the incorporation of α-chiral aromatic N-1-naphthylethyl side chains. *J Am Chem Soc* **133**, 15559-15567 (2011).
56. J. A. Crapster, I. A. Guzei, H. E. Blackwell, A peptoid ribbon secondary structure. *Angew Chem Int Ed Engl* **52**, 5079-5084 (2013).
57. R. V. Mannige *et al.*, Peptoid nanosheets exhibit a new secondary-structure motif. *Nature* **526**, 415-420 (2015).
58. X. Li *et al.*, Photoinduced Electron Transfer and Hole Migration in Nanosized Helical Aromatic Oligoamide Foldamers. *J Am Chem Soc* **138**, 13568-13578 (2016).
59. M. Kudo, V. Maurizot, B. Kauffmann, A. Tanatani, I. Huc, Folding of a linear array of α-amino acids within a helical aromatic oligoamide frame. *J Am Chem Soc* **135**, 9628-9631 (2013).
60. M. Kudo, V. Maurizot, H. Masu, A. Tanatani, I. Huc, Structural elucidation of foldamers with no long range conformational order. *Chem Commun (Camb)* **50**, 10090-10093 (2014).
61. R. Zubatyuk, J. S. Smith, J. Leszczynski, O. Isayev, Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci Adv* **5**, eaav6490 (2019).
62. J. S. Smith *et al.*, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **10**, 2903 (2019).
63. J. S. Smith *et al.*, The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci Data* **7**, 134 (2020).
64. J. C. Kromann, C. Steinmann, J. H. Jensen, Improving solvation energy predictions using the SMD solvation method and semiempirical electronic structure methods. *J Chem Phys* **149**, 104102 (2018).
65. C. Bannwarth, S. Ehlert, S. Grimme, GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J Chem Theory Comput* **15**, 1652-1671 (2019).
66. E. Caldeweyher *et al.*, A generally applicable atomic-charge dependent London dispersion correction. *J Chem Phys* **150**, 154122 (2019).
67. S. Riniker, G. A. Landrum, Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J Chem Inf Model* **55**, 2562-2574 (2015).
68. S. Wang, J. Witek, G. A. Landrum, S. Riniker, Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J Chem Inf Model* **60**, 2044-2058 (2020).
69. A. Hjorth Larsen *et al.*, The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017).
70. C. R. Harris *et al.*, Array programming with NumPy. *Nature* **585**, 357-362 (2020).
71. J. C. Cole, O. Korb, P. McCabe, M. G. Read, R. Taylor, Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *J Chem Inf Model* **58**, 615-629 (2018).
72. B. Nijholt, JosephHoofwijk, JornAkhmerov, Anton. (Zenodo, 2021).
73. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261-272 (2020).