The Rational Discovery and Development of Disordered Protein Ligands

David Baggett

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2019

Reading Committee:
Abhinav Nath, Chair
William Atkins
Rheem A. Totah

Program Authorized to Offer Degree:
Medicinal Chemistry

University of Washington

Abstract

The Rational Discovery and Development of Disordered Protein Ligands

David Baggett

Chair of the Supervisory Committee:
Abhinav Nath
Department of Medicinal Chemistry

Intrinsically disordered proteins play vital roles in biology and their dysfunction contributes to many major disease states, making them appealing pharmacological targets. These proteins are challenging targets for rational ligand discovery or drug design because they are highly dynamic and fluctuate through a diverse set of conformations, frustrating structure-based approaches. This dissertation describes the unique properties of disordered proteins that challenge ligand design strategies, then details methodologies designed to address these challenges. Chapter 2 details the *in silico* and *in vitro* methods used to identify and validate ligands of the disordered protein tau. Chapter 3 then further examines these compounds and how they interact with tau, as well as utilizing analogous compounds to understand the structure/function relationships that dictate their activity. Chapter 4 illustrates the adaptability of these approaches by utilizing similar methodologies to identify ligands of a different disordered protein system, phenol soluble modulins.

Table of Contents

Acknowledgements

There are a number of people who deserve recognition for helping me get to this point:

Abhinav Nath: Is everything I could've needed in an advisor. I have always been impressed by his limitless patience, understanding, and dedication to science. In many ways I want to model my scientific career and the ways I work with others on the examples he has given.

Rheem Totah: Rheem was one of the first people to encourage me to find a work life balance but has also been a model of hard work and dedication. She is a great example of how to support someone to take care of themselves while simultaneously putting in the hours needed to complete a project.

William Atkins: Bill has been a source of support and a refreshing reminder to face challenges with a perspective that includes a bit of humor. I hope to keep that with me as long as I'm in science.

My lab mates, past and present: The Nath lab has always been a place of support and innovation. Lab meetings were instrumental for my ability to complete projects and digest new ideas.

My cohort (Hannah Baughman, Eric Evangelista, and Dennis Goulet): Not only do I appreciate their help when completing coursework, but I recognize that because of them many of my perspectives on life have changed for the better.

My family: My family's encouragement and support are what allowed me to achieve what I have. The foundations that they gave me are what allowed me to travel to Washington and take on this project.

Michelle Murphy: Michelle has helped me survive some of the most challenging times of my life and has been joy in my life. She ensured that when I felt despondent and drained, I took care of myself and recovered. We plan to get married soon, and I look forward to spending the rest of my life with her.

Chapter 1
The Importance and Challenges of Studying Disordered Proteins

1.1) Introduction to disordered proteins

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) (collectively called disordered protein in this chapter) are proteins and sub-domains of proteins that lack well folded and stable conformations. They have garnered a great amount of attention due to the fact that they defy the structure-function paradigm that shaped thoughts about proteins and their behavior for decades.[1] Until recently, disordered proteins were largely ignored and dismissed as niche proteins that defied the general rules governing protein function. Due to their dynamic structures in solution, disordered proteins are not easily observable using common techniques to determine protein structures, and as a result were mostly undetected for decades. Recent advances in bioinformatics and biophysical techniques have elucidated their roles, and the structure-function paradigm has expanded to a spectrum instead of a rule to accommodate the new insights derived from studies of intrinsic disorder.[2] Disordered proteins display interesting and dynamic functional behavior, and a number of unique properties that make them valuable tools for the cell, but also susceptible to aberrant behavior that leads to different pathologies.

A brief history of intrinsically disordered proteins and the structure-function paradigm

The structure function-relationship hypothesis has defined how the scientific community relates a protein's 3D structure to its biological applications. Hypothesized by Emil Fischer in 1894, the "lock and key" analogy, equated proteins and their ligands partners to specifically shaped locks designed to interact with particular keys.[3,4] This theory maintained that in order for a protein to perform a reliable biological function such as catalysis, it needed to have a specific and singular conformation that was structured to enact such a function. This theory was

widely accepted for a time and reinforced by multiple published examples of proteins whose functions could only be performed by exact placement of ligands relative to specific residues. Examples of proteins that suggested a biological function without a reliable structure were dismissed as rare exceptions to the rule.[5]

Near the end of the twentieth century, a growing list of publications demonstrated examples of proteins that defied the rigid structure-function paradigm. They showed similar phenomena, but authors used markedly different terminology to make their points, making it challenging to connect the studies with each other to show a larger phenomenon instead of isolated case-studies.[6–9] Many groups questioned the notion that a well-defined structure is a necessity for a protein to enact function, but the scientific community at large ignored these challenges and held to the idea that structure defined function.[10–13] It was not until a series of publications that focused on the prevalence of IDPs across biology,[14–18] as well as an increased understanding of the specific types of protein sequences that resulted in functionally unfolded proteins[19] that a change in the widespread perception of disordered proteins and their utility matured.

The realization that disordered proteins are not rare but are instead remarkably common in many biological systems is what changed their widespread perception from niche and novel to integral to the function of biological systems. The first step of this shift in perception required a method of identifying these obscure protein sequences. To do that, the Dunker group developed a set of computational predictors that identified disordered regions in proteins with reasonable accuracy.[14] Their pioneering work employed two methods: a rule-based approach developed by rationally examining known regions of disorder of calcineurin, and a neural network approach that employed a learning algorithm that generated a set of predictors by identifying features

common to a set of manually identified disordered regions across many peptides. The neural network method examined features such like individual amino acid composition, but also more generalized parameters such as hydrophobicity and flexibility of peptides to identify the features that suggest disorder. The presence of 8 specific amino acids (A, R, G, Q, S, P, E, and K), as well as hydropathy and chain flexibility, were the features with the strongest correlation to observed disorder. These methods were able to predict disorder with accuracies from 59 to 74%, depending on the way the method was built and what it was tested against.[14,20] Early analysis of 5 eukaryotic genomes using these methods showed that 25-41% of the encoded proteins examined had significant disordered regions. This illustrates that if the disordered portion of a proteome is disregarded or ignored, it leaves the understanding of biology "seriously incomplete".[16]

The Dunker group and others improved upon and applied this methodology to identify large swaths of proteins across different systems that are natively disordered. Improvements came from adjusting the predictive algorithms well as improving the datasets that were fed into computer algorithms. As more disordered regions were identified and confirmed, more trends became apparent, such as amino acid compositions between well folded/compact proteins and the more extended/disordered proteins were more specific than just enrichment of the 8 amino acids originally identified. Instead of a defined class of "order promoting" or "disorder promoting" residues, it is more appropriate to note that they exist on a spectrum. Arranged from order-promoting to disorder-promoting residues, the ranking is: W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P.[21] When the patterns identified with these studies were applied to curated sets of proteins, the neural network predictor used found that proteins associated with cancer or cell-signaling had notably higher amounts of protein disorder than a control set of

eukaryotic proteins. Although less than the two enriched sets, the control set of unspecified proteins predicted about half of them included had significant regions of disorder (>30 amino acid chains).[18] This not only showed that disordered proteins are important to our overall understanding of biology, but also very relevant to human health and disease.[18] Later analysis of almost 3500 proteomes showed a correlation between proteome complexity and the proportion of disordered content in the proteome (Figure 1.1).[22] It is possible that the correlation between disordered content and the complexity of a proteome is related to the prevalence of disordered protein in cellular signaling pathways. An analysis of different classes of proteins found that over 60% of the almost 2500 signaling proteins were predicted to have segments of disorder at least 30 residues long. The only examined category predicted to be more disordered than signaling was the class defined as proteins whose mutations are strongly correlated with the onset of cancer. Almost 80% of these proteins have significant segments of intrinsic disorder in the wild-type sequence.[18]

Comparing Ordered and Disordered Proteins

Almost immediately after the recognition of how prevalent IDPs and IDRs were in biology, research focused on understanding what factors encourage protein disorder. The methods used to identify protein disorder identified commonalities between disordered proteins and helped understand these factors that drive disorder. IDPs and IDRs contain notably fewer hydrophobic residues (lower hydropathy), and much higher net charges at physiological pH than well folded proteins (Figure 1.2).[19] Low hydrophobicity removes the entropic encouragement separate hydrophobic regions from the surrounding aqueous environment by folding. A high net charge encourages disorder because like-charged residues repel each other and force the protein conformation to expand. This is akin to how many well-folded proteins unfold at extreme pH,

where residues are more likely to have the same charge.[23] Not only does net charge affect the behavior of protein extension and unfolding, but also the charge distribution plays an important role. Computational modeling of simplistic proteins containing the same net charge, but differing charge distributions showed that charge distribution is a critical factor that determines how "random" these peptides behaved and affects global properties such as radius of gyration (Figure 1.3).[24] The understanding of what determines relative order of a protein continues to evolve, and work is ongoing as to how best to classify them and combine different metrics of disorder.[25–28]

A popular way to think about the possible confirmations of proteins and a useful way to contrast the behavior of well-folded proteins against disordered ones is in terms of energy landscapes. Levinthal's paradox (named after the researcher who posited the question) asks: out of all of the possible configurations of a given protein sequence, how does a (well-folded) protein rapidly fold into to a comparatively narrow region of conformations in which it performs its function.[29] A random search where the torsion angles between amino acids arbitrarily shift and fluctuate until they find the proper conformation is too time consuming and impractical for biological purposes. Instead nature performs a biased search where environmental forces guide the protein towards the "correct" structure.[30,31] This theory matured to take into account specific forces such as entropy, van der Waals forces and electrostatics, and the model took the form of an energy landscape where conformations were represented by their relative energies and an energy funnel guides the conformation of an ordered protein towards its well-folded set of conformations.[32]

In contrast to well-folded proteins, where energy landscapes are relatively smooth and guide sequences towards a specific structure, intrinsically disordered proteins have rugged

landscapes without any global minima structure to navigate towards (Figure 1.4).[33,34] Another important difference between them is that while the energy landscape of folded-proteins is stable under physiological conditions, the landscape of a disordered protein is highly sensitive to the environment of the protein and can drastically change with protein concentration, ionic strength, and other factors. This has implications on their different roles in biology, a point elaborated on later in this chapter. This feature was considered as a possibility even before the concept of an intrinsically disordered protein was widely accepted as a potential answer to the differences between the speed of folding and the different dynamics exhibited by different classes of proteins.[35]

1.2) The utility of disorder

Bioinformatic analysis has revealed that IDPs and IDRs are remarkably prevalent. In many eukaryotic systems, >30% of the proteome is predicted to have substantial disordered content, and some classes of proteins have higher percentages of disorder.[36] The biological purposes of these regions have not always been clear, but improved techniques in biophysics and bioinformatics have recently shed light on disordered proteins and the advantages they have compared to folded protein.[37] Notably, disordered proteins are highly represented in signaling cascades and facilitate many protein-protein interactions where cellular mechanics take advantage of the unique properties that come with disorder.[18,38,39]

Advantages of conformational flexibility

The most immediately accessible advantage of IDPs is their ability to change conformation to perform a function.[40] For example: a disordered protein or disordered region can change its conformation in order to interact with potential protein partners. Adopting a specific conformation for a specific protein interaction is hypothesized to happen through a

process called coupled folding and binding where the disordered protein is encouraged to fold by interactions with a protein partner that stabilize a particular conformation.[41–44] Alternatively, some interactions do not require disordered proteins to take a specific conformation upon binding to a target, instead utilizing one of many possible configurations. This ensemble of possible protein-protein conformations is collectively called a "fuzzy complex".[45] It is thought that the combination of these two phenomena: the ability to take on a specific fold when binding a protein target, and the ability to bind without needing a specific fold, are factors that allow IDPs and IDRs the ability to serve as critical intermediaries in signaling cascades.[40,46] They often serve as hubs in complex networks of protein interaction, facilitating interactions between multiple sets of proteins to propagate a cellular response.

One of the advantages thought to be inherent to disordered proteins is their ability to increase the rate of association by what is called the "fly-casting" method. This hypothesis states that since disordered proteins are extended, they have a greater reach than their well-folded counterparts to encounter their binding partner, and therefore take less time to associate. Once the protein makes a partial contact, an induced folding event then occurs resulting in a favorable binding event. In contrast, the chances of a folded protein making contact with its partner in the correct orientation is much lower.[47] Association rates for IDP's are often very fast,[43,48,49] lending weight to the fly-casting hypothesis, but it is difficult to directly test the hypothesis as altering the order/disorder dynamic is difficult to do without affecting other factors.[50] Examination of the fly-casting hypothesis with computational and mathematical modeling suggest that the mechanism is possible, but the resulting rate-enhancement is comparatively small: only 1.6-3 fold faster.[51] An alternative explanation is that disordered proteins are able to eliminate the requirement of correct-orientation that limits the rate constants of ordered proteins with a

mechanism dubbed “dock and coalesce”. In this mechanism, only part of the protein is required to find its proper position, and the flexible nature of a disordered protein then coalesces around it, eliminating the need for completely aligned orientations.[51]

Beyond the ability to bind, the folding-binding events of disordered proteins gives them another advantage that is well utilized in cellular mechanisms. The interactions between IDPs and their protein partners are highly specific, but this specificity comes at a thermodynamic cost as the enthalpic stabilization from the protein-protein interaction is tempered by the entropic penalty of restricting a disordered protein’s conformation.[52] These low affinities paired with fast association rates result in fast disassociation-rates. [51] This dynamic allows disordered proteins to elicit highly specific interactions to trigger signaling events, but quickly dissociate from the partner. This is advantageous to processes such as signaling which require a fast response, but one that is easily reversed as to not leave the signal active for too long.

Regulation of disordered protein activity

Beyond the ability of IDPs to interact with the requisite players of a signaling cascade, the ability of IDP/IDR behavior to be modulated by the cell is a necessity for their utility as signal messengers. A useful signal is one that can not only be turned on and off, but also have a gradient of responses and activities.

The first opportunity for IDP behavior to be regulated is through regulating its production and degradation. Improper regulation of IDP production is hazardous to the cell, evidenced by the fact that gene populations that are harmful to cells when overexpressed tend to be high in disordered content.[53] Multiple mechanisms control the abundance of IDP concentrations in the cell, primarily increased transcript clearance and degredation.[54–56] It is critical for the cell to maintain a tight regulatory control on disordered protein populations, as high concentrations can

initiate improper molecular interactions that aberrantly trigger signal cascades with often disastrous consequences to the cell.[53] Not only how and when IDPs are formed but also where they are formed is highly controlled by the cell. As opposed to ordered proteins which tend to be synthesized then transported afterwards, disordered proteins have their mRNA transported to specific regions of the cell and the resulting proteins are created there.[57,58] This allows a cell to maintain an overall low concentration of a specific protein but have a localized high-concentration to enact a specific function when needed, while at the same time minimizing the chance that it inadvertently makes an erroneous connection while being transported.[58]

In some cases, the localization of disordered proteins creates dynamic and tunable sub-compartments within a cell. Liquid-liquid phase separation (LLPS) occurs when soluble proteins self-associate and create a phase-separated boundary within an aqueous environment. These liquid environments are separated from the rest of the cytoplasm are sometimes called non-membrane-bound organelles (NMBO). This phenomenon is central for proper function of cellular processes such as the nuclear pore complex.[59,60] LLPS droplets can continue to mature and transform from liquid condensates to solids, leading to aggregation of proteins. Factors that facilitate phase separation include multivalent binding as well as amino acid sequences that encourage regions of extreme flexibility. These properties are ones that IDPs exemplify, and it comes as no surprise that IDPs are often critical components in these phase separating systems. Misfunction of the disordered proteins regulating NMBOs can lead to pathologies such as amyotrophic lateral sclerosis.[61,62]

An additional mechanism for cells to regulate IDP behavior is by modulating their energy landscape. A common and direct method to change the energy landscape is to change the physical properties of an IDP with a post translational modification (PTM). Disordered proteins

and segments are more enriched in PTMs than well-folded proteins and regions.[63,64] Post translational modifications can change energy landscapes by altering the charge distribution of a protein or by introducing steric hindrances.[65] This change in energy landscape causes some conformations to be favored and other conformations to be disfavored, which in turn alters the IDPs binding affinity for other proteins and the kinetics of those interactions.[66,67]

Ionic strength of the aqueous environment of IDPs is another mechanism by which cells regulate IDP activity. IDPs are often high in net charge which encourages extended conformations, but they also rely on electrostatic interactions to generate specificity with oppositely charged grooves in protein partners.[68] By modulating the ionic strength of the solution surrounding IDPs, a cell effectively changes the strength of these electrostatic forces and regulates IDP conformation and behavior.[69] Similarly, the energy landscape of an IDP can also be changed by the presence of specific ions, protein crowding, and the presence osmolytes. [70–72] These factors change across different regions of the cell, giving IDPs drastically different conformational ensembles based on their specific situation which results in drastically different behaviors.

1.3) Characterizing the structures and biophysical properties of IDPs

Biophysical studies focusing on disordered proteins and disordered regions are frustrated by their structural heterogeneity. Most notably, X-ray crystallography (XRC) is incapable of providing information on IDPs or IDRs in their native state. Because XRC relies on reconstituting a single repeated conformation that has been organized in a crystal lattice, it cannot provide any information on the heterogenous conformational ensemble populated by disordered proteins. Disordered proteins have been crystalized and structures derived from XRC, but in these cases their structure is uniformly stabilized often through binding with another

partner, or extreme chemical environments and is not representative of the protein under natural conditions.[5,73] Despite being unable to use the gold standard of protein structure determination, scientists have been able to gather valuable information on disordered protein structure through the innovative use of other biophysical techniques.

Nuclear Magnetic Resonance studies of IDPs

Nuclear Magnetic Resonance (NMR) has proven to be one of the most useful biophysical techniques for studying protein disorder, despite analytical challenges. In NMR, samples are examined under the effect of a constant strong magnetic field. When these samples are then probed with a weaker magnetic field, the reaction to the weaker field can be measured. Each atom's nucleus will respond differently to the field based on what type of atom it is and the chemical environment it is experiencing. NMR only reports on atoms with an odd number of nuclear particles (protons and neutrons), so samples often need to be made with radioactive isotopes to make them are observable, making NMR a costly procedure to run. NMR is often used for well folded proteins to derive their structure in non-crystalline environments, giving a better understanding of their conformation and dynamics in realistic conditions. by IDPs and IDRs frustrate standard NMR methods to study protein structure because they often only display extremely low chemical shifts which makes interpretation of the data difficult.[74]

Despite challenges associated with NMR and disordered proteins, it is still the most used technology when studying biophysical characteristics of IDPs.[73,75] Even relatively simple experiments such as $^{1}H–^{15}N$ heteronuclear single-quantum correlation (HSQC) are useful for informing on IDPs. Even when these experiments cannot determine specific conformations, changes in chemical shifts report on events such such as binding and other structural changes.[76,77] Despite limitations, HSQC is heavily used because it can provide information on

specific residues of an IDP, and can be relatively inexpensive compared to some other NMR approaches.[78]

NMR data can be used to generate structural ensembles of disordered proteins by back calculating structures and weighting them appropriately.[79] While these chemical shifts are useful for determining local factors such as secondary structure of residues and atoms in IDPs, it isn't uncommon that the information provided by these studies isn't sufficient to fully understand the structure and dynamics of an IDP. In these cases, interpretation of NMR data is enhanced by the addition of other sources of information that provide orthogonal information. Paramagnetic relaxation enhancement (PRE) experiments often provide a great deal of insight when added to NMR studies. By incorporating specific paramagnetic probes into specific points of the disordered protein and then observing the reduction of the NMR peaks of spatially close residues, it is possible to determine the long-range (35 Å) contacts experienced by the probed residue.[80] In one recent example, NMR studies utilizing PRE experiments have been used to show where protein chaperones interact with disordered ligands. By placing spin-labels on the disordered N-terminal region of the chaperone protein HSPB1, NMR studies were able to show interactions between the disordered N-terminal with rest of the protein, providing insight into the domain functions of HSPB1.[81]

Small angle X-ray Scattering and Single-molecule fluorescence studies

Small angle X-ray scattering (SAXS) is another technique that has provided a great deal of insight towards the structures of disordered proteins in solution and is often used synergistically with NMR studies.[82,83] SAXS curves can be interpreted to report on factors that would be otherwise inaccessible to NMR, such as size and shape, as well as oligomeric state.[78] Different analyses yield information on the radius of gyration as well as the flexibility of the

protein can be extrapolated.[84] SAXS data is often used as a method of validation for other methods of approximating the behavior of a disordered protein such as computational studies[85], but combinations with other techniques such as NMR allow the generation of theoretical conformational ensembles from SAXS data itself.[86,87] These ensembles then provide valuable insight to the behavior and functions of IDPs and their complexes.

Single molecule fluorescence techniques have been useful tools in probing the protein dynamics and afford unique perspectives that are advantageous towards studying IDPs. In particular the ability to examine a single molecule at a time is helpful for determining discrete parameters as opposed to ensemble averages, which is one of the factors limiting NMR data. Förster resonance energy transfer (FRET) works through the energy transfer of two distinct probes and can be used to probe the distances between the different probes and how those distances fluctuate. Single molecule FRET (smFRET), applies this same principle to a single protein molecule at a time. This specificity is achieved by either exciting one molecule at a time, or restricting the area observed such that only one molecule is being examined by the technique. By focusing on one molecule at a time, rare or exceptionally fast conformational changes and fluctuations that would normally be obscured in ensemble techniques can be observed.[88] This technique has been effectively used to identify a drastic change in the energy landscape of α-synuclein in response the presence of ligands and osmolytes and notable conformational changes in the disease associated IDP tau in the presence of heparin. smFRET also showed that the disordered C-terminal region of the NMDA receptor extended upon phosphorylation.[72,89,90]

Molecular Dynamics and Monte Carlo Studies of IDPs

The biophysical techniques that are used to examine disordered proteins provide invaluable insight to the function and dynamics of IDPs. Unfortunately, this information is

generally representative of average states of the protein as opposed to specific states, or the information is low resolution and therefore insufficient to gain atomic-level insight pertaining to a protein's structure. Molecular dynamics (MD) and Monte Carlo simulations help meet the need for higher resolution information by generating structural ensembles that help interpret and predict experimental data.[91] Molecular dynamics simulations predict the behavior of molecular environments by calculating the energies between particles using force-field equations and then using those energies to determine how these particles should move. Monte Carlo simulations work in a similar fashion, although instead of directly calculating energies into movement of particles, Monte Carlo simulations use these energies to calculate the chances that a protein will move from one state to another. These methods have classically been used and developed to understand physical properties such as folding, allosteric effects and the mechanisms of action of well-folded protein systems. They are appealing because they allow a level of control that is impossible in most experimental techniques, but also are able to report of the movement and precise location of each particle in the simulation even if that information is an estimate.[92] Since these techniques approximate the atomic positions of proteins with explicit detail, they are appealing methods of getting conformational details of disordered proteins.

Molecular dynamics approaches have been developed to understand the dynamics and mechanics of well folded proteins. Initial forays of applying these techniques to disordered systems noted that parameters derived from their simulations didn't agree with experimental results.[93] The force-fields and solvent models used to calculate energies were based on data and theory pertaining to well-folded proteins and the differences between the two classes of protein required different parameters. Since identifying this issue, force fields and water models have been optimized for use with disordered proteins and regions.[94,95]

A separate issue pertaining to disordered proteins and molecular dynamics is computational cost. Molecular dynamics simulations routinely perform calculations between hundreds of thousands of particles, and the number of calculations scales with the number of particles making them often computationally quite expensive. This is already notable when dealing with well-folded proteins and the conformational space to be explored is relatively small. When dealing with disordered proteins, the relevant conformational landscape is much larger, and the timescales needed to sufficiently explore them with conventional methods are prohibitively expensive.[96] To circumvent this problem various enhanced sampling techniques have been developed which apply artificial potentials or forces to overcome the energy barriers that separate conformations. One such enhanced sampling techniques is replica exchange molecular dynamics (REMD) which functions by simultaneously running multiple simulations at various temperatures, and periodically exchanges the conditions between simulations. This allows the simulations to use the higher temperature conditions to overcome the energy barriers between relevant states that would otherwise hamper the conformational exploration.[97,98] Enhanced sampling methods are often used to generate vast conformational libraries, which are then mined for relevant and insightful structural information. [99–103]

Other methods that modify the way calculations are handled do so either to reduce the complexity of the system, or to restrict it such that it agrees with experiment. Coarse-grained techniques simplify the nature of a simulation by representing complex regions like functional groups, residues or even entire protein domains as simplified "beads". This approach reduces the computational load of a simulation by reducing the number of interactions that need to be calculated, but does so at the cost of detail. Despite the loss of detail, this method has been used

to simplify protein representations and gain insight into what drives complex phenomena such as protein folding and aggregation.[104,105]

In order to improve the structural ensembles generated from computational approaches without performing a computationally intensive search of the entire energy landscape, a useful approach is to restrict simulations such that they are confined within regions that agree with experimental values from *in vitro* experiments.[96] This can be done either by removing generated conformations that disagree with experiment, or by creating customized force fields that a MD or Monte Carlo simulation can use to impose an energetic cost for moving towards conformations that disagree with experimental data. This has been successfully performed using data from smFRET, NMR, and SAXs experiments, and in theory most sources of reliable data could be used to bias computational experiments[34].

1.4) Amyloid aggregates

A class of diseases where IDP involvement is notable is degenerative misfolding diseases.[106] In these diseases proteins misfold into a specific type of aggregate called amyloid fibrils. Amyloid fibrils are characterized by a helical stack of beta-sheets parallel to the fiber's axis.[107] The concept of amyloid structure was originally established in 1854 by the physician scientist Rudolph Virchow who noticed strange deposits in brain tissue that appeared to be some sort of protein. The name "amyloid" means starch-like, and comes from the fact that it was identified by iodine staining and reacted in a similar manner that starch does.[108] The more biophysical definition currently used to define amyloid was established by an early protein structure experiment that determined that egg-white proteins formed the characteristic stacked beta-sheets when in amyloid fibrils.[109] Amyloid structure is in some ways an interesting

counterpart to disordered proteins: it is remarkably stable and the cores are structurally similar regardless of the sequence of the constituent protein.

Amyloid formation, stability, and structure

Originally amyloid was thought to be a phenomenon exhibited by a small set of proteins prone to such behavior. Recent studies have shown that many proteins are capable forming amyloid structures if subjected to the correct conditions, although many of these conditions are quite extreme and not likely to be encountered physiologically. [110,111] There are two factors that seem to be critical for the formation of amyloid: protein concentration and the generation of a structural template to build upon.

A protein's concentration has little effect on the stability of its well-folded concentration, but it can change the stability of amyloid conformations. When a protein's concentration is low, amyloid configurations are unstable when compared to the folded state and the protein has no impetus to reconfigure to amyloid. As protein concentration increases amyloid configurations become more stable. Amyloid configurations are dependent on intermolecular bonds which are more likely when the protein exists in higher concentrations, as opposed to the intramolecular bonds that dictate the stability of native configurations which are mostly independent of protein concentration. After the protein concentration passes the critical concentration at which the amyloid state is as stable as the native state, the protein is thermodynamically driven to change structure and begin forming amyloid. When the protein concentration is beyond the critical concentration, the protein may still remain in its native state if there is a sufficiently high energy barrier that prevents its conversion.[110,111] Such a system is said to be "kinetically trapped".

If a protein is at a sufficiently high concentration that amyloid structures are stable, it may still be kinetically trapped until it experiences a conformation that is capable of aggregation,

a process called nucleation.[112] This is evidenced by the characteristic lag-phase of amyloid formation kinetics.[113] The fact that the lag phase can be eliminated by adding a small amount of pre-formed aggregate shows that this delay is the time it takes for a protein in solution to form that template.[114] For well-folded proteins, a precursor to this step is at least partial unfolding of the protein region that forms the amyloid core. For IDPs/IDRs, the protein requires no unfolding but still has to search conformational space in order form the correct structure to serve as a seed for aggregation.[115]

Amyloid in biology

The presence of amyloid protein deposits is indicative and even diagnostic of a number of human diseases.[106,116,117] However despite this correlation, the source of pathology in these diseases is still unclear. In some systemic amyloidosis pathologies, astoundingly large quantities of amyloid aggregates (sometimes even kilograms) are deposited in tissues and interfere with biological function, leading to biological system failures.[118] In other pathologies the causative agents are less clear. For example, Alzheimer's disease is predominantly characterized by amyloid deposits of two types of proteins: Aβ and tau. Early studies showed that the presence of Aβ amyloid was correlated with cell death, suggesting that its aggregation was the causative factor in neurodegeneration.[119] This hypothesis was questioned when analysis showed that disease progression is not strongly correlated with the quantity of insoluble aggregate, as would be expected if it was the toxic species. Instead disease progression has been shown to be correlated with soluble concentrations of aggregating protein.[120] Additionally mouse models show disease symptoms before insoluble aggregates are formed.[121] These observations gave rise to a new hypothesis that suggests that a pre-fibrillar state is responsible for cell toxicity in amyloid pathologies.[122] Upon further examination, multiple studies have shown that although

fibrils can be toxic, pre-fibrillar species are often more toxic and are thought to be the most destructive element for many proteins capable of forming amyloid aggregates.[123–125]

Beyond its role in disease, there are many examples of functional amyloid. Functional amyloid structures have been found in bacteria, fungi, spiders, and even humans.[122] The human hormone system makes use of amyloid structure as a way to store and release peptide hormones in a controlled manner.[126] Ironically one of the more notable examples of functional amyloid is a major issue for human disease. Bacterial biofilms are a surface-attached bacterial community, organized in such a way that they can channel necessary nutrients to different layers.[127,128] This is an issue for human disease as these biofilm communities are often resistant to different antibacterial agents. A class of peptides called phenol soluble modulins (PSMs) have been shown to be involved in structuring biofilms and have amyloid characteristic while doing so, suggesting a role for amyloid in healthy bacterial colonies.[129–131]

1.5) Drug discovery for disordered or misfolded Proteins

Because of their widespread utility in disease and healthy biological functions, IDPs have been identified as promising protein targets for therapeutic intervention.[132,133] The predominant focus around drugging these proteins is to modulate their protein-protein interactions and make these interactions either more or less favorable by stabilizing or destabilizing the protein-protein complex. The interactions between IDPs and their protein partners are highly specific and often have many downstream effects. As such, a change in affinity due to a small molecule could have specific yet drastic ramifications on cell behavior.

Current methods of drug development are frustrated by the dynamic nature that defines disordered protein, as many medicinal chemistry techniques rely on utilizing an established and reliable binding pocket to understand factors that determine ligand affinity and effect. Despite

this complication, modulation of disordered protein behavior with small-molecules is appealing. The following segments will discuss methodologies that have been used to identify and refine potential therapeutics intended to modulate the activity of disordered proteins.

High throughput screening and chemical analoging

High throughput screening is a viable, albeit costly and time-consuming method of discovering ligands for disordered proteins. As with all attempts of ligand searching, the first step in using such an approach is to develop an assay capable of detecting binding in an easily scalable manner. Unfortunately, the dynamic nature of disordered proteins presents challenges for many of the methods routinely used with well-folded targets. For example, methods that require the protein to be attached to a surface, like surface plasmon resonance, cannot be used on IDPs because the action of capturing the protein can be enough to drastically modify the dynamics and affect interactions with potential ligands.

A good example of successful library screening for IDP modulating compounds focuses on the interaction between c-Myc and Max. The c-Myc and Max interaction is a well studied coupled folding and binding event. When c-Myc binds to Max it undergoes a disorder to order transition, as evidenced by changes in Circular Dichroism spectra.[134] Myc proteins are involved in signaling and transcription pathways, and their activity has ramifications on metabolism, apoptosis, and other aspects of cell growth, making it a promising target for pharmaceutical invervention.[135] To develop a scalable screening assay to identify ligands targeting the disordered protein c-Myc, researchers adapted a yeast transcription system so that it only activated when these two proteins associated. They then used this to screen compounds. By measuring the effect of the compound on beta galactosidase production, a downstream result of the transcription factor activated by the interaction between c-myc and Max, they were able to

identify numerous low-molecular weight compounds that inhibited the interaction.[136] Subsequent studies explored similar chemical space of compounds identified in the pilot study to improve upon the efficacy of these compounds, showing that once ligands are found they can be further optimized using canonical medicinal chemistry techniques such as structure-activity relationship studies.[137] Further examination of this interaction showed that the activity of these molecules was a result of their affinity towards c-Myc, which confined the conformations of the protein to the disordered state, preventing the structural rearrangement necessary for its binding with its partner Max.[138] This shows the importance of conformational flexibility towards IDP function, and also illustrates how modulating the conformational behavior of IDPs can be a valid strategy for influencing their behavior in disease states.

Computational modeling and *in silico* docking

Another approach for developing ligands for disordered proteins utilizes the high resolution data gathered from molecular dynamics to use as targets for computational docking. The initial step for such a study is to generate a conformational library of different possible structures of the protein. Zhu et. al did this for the aggregation prone disordered protein alpha synuclein using replica exchange molecular dynamics. They generated data using a range of temperatures spanning 276 to 376°K and then searched the data generated at 309.4°K for highly populated structures using clustering analysis. The structures identified with clustering analysis were then used in a computational docking screen to find binding hotspots using a fragment library. They also screened against then known ligands curcumin and Congo red to show that they bound in these identified hot-spots and followed up with molecular dynamics simulations initiated with these bound conformations to suggest that the compounds kept their association in these hot spots.

Continued studies with this methodology focused on the disordered protein α-synuclein (α-syn).[139] Zhu et al adjusted their method to bias their conformational library by restraining the molecular dynamics simulations with NMR-derived parameters. After generating and searching the conformational library of α-syn they sought out binding pockets and screened a library of fragment probes which led to the identification a lead compound. Examination of this compound found that it rescued α-syn-induced disruption of vesicle trafficking. This served as an early example for the viability of using computational methods to discover viable lead compounds that target disordered proteins.

1.6) Tau as a model system

Tau is an exemplary IDP to study for the purposes of deciphering IDP behavior and then targeting it for therapeutic intervention. It has a flexible and dynamic conformational ensemble, it has relevant involvement in functional biology, and it is involved in a number of pathologies involving amyloid aggregation.

In its natural role, Tau is a microtubule associated protein that is thought to bind and stabilize microtubules and has important roles in neurite outgrowth and axonal trafficking[140]. Its structure is often classified into three distinct regions. The N terminal region is the projection domain that helps regulate microtubule spacing and is involved with membrane anchoring and tau dimerization.[141–143] The Microtubule binding region (MTBR) is a domain composed of 4 imperfect repeats. It is at least partially responsible the association of tau with microtubules with multiple weak binding sites across different repeats. These interactions help drive the assembly of tubulin monomers into filaments, as shown by their effect on tau's ability to associate with two tubulin monomers.[144] Between these two regions is the Proline rich region (PRR), a domain

composed of two sub-regions and highly concentrated in proline residues. This domain also has tubulin binding and polymerization capacity and works synergistically with the MTBR.[145]

Full-length tau protein is 441 residues long, but like many highly regulated disordered proteins, it is subject to differential splicing resulting six different possible isoforms. Different isoforms of tau have between zero to two N-terminal repeats, and the second of the four repeats of the MTBR is also differentially expressed. Proportions of the different isoforms with relative populations of these different isoforms changing throughout different stages of development. In fetal stages of development only the shortest tau isoform is expressed but by adulthood all six isoforms are expressed. The reasons behind this are not fully known.[146–148] Another notable behavior of disordered proteins that tau exemplifies is the dependence of its behavior on phosphorylation. Full length tau has 73 serines capable of being phosphorylated, and there is a complex network of kinases that handles this process and help regulate its natural function.[148] In many disease states tau becomes hyperphosphorylated and this hyperphosphorylation is associated with the detachment of tau from microtubules, and tau that is extracted from post-mortem samples of patients with tauopathies are often hyperphosphorylated.[149,150] The relationship between tau phosphorylation and its function in biology and disease complicated. The behavior of tau is determined not only by whether or not it has been phosphorylated, but by which specific sites have been phosphorylated. There are many sites that when phosphorylated induce aggregation, but there also exist sites that when activated with a kinase inhibit the aggregation of tau, making the relationship between tau, phosphorylation, and the kinases and phosphatases that regulate it hard to decipher.[151,152]

Tau's formation into amyloid aggregates is a diagnostic feature in a suite of pathologies, collectively called tauopathies, and include neurodegenerative diseases such as Alzheimer's

disease and chronic traumatic encephalopathy. These tauopathies closely resemble another class of diseases in which the disordered protein α-synuclein aggregates into amyloid fibrils and is exemplified by the Parkinson's disease. The similar mechanisms involved suggest that these neurodegenerative diseases are related and this hypothesis is further supported by examples of α-synuclein and tau affecting each other's pathologies.[153] α-synuclein aggregates, called Lewy bodies, have been found in half of Alzheimer's patients when identified with α-synuclein antibodies.[154] *In vivo* models have shown that tau co-localizes with α-synuclein and changed the size distribution of α-synuclein aggregates as well as increased their toxicity.[155] The routine co-occurrence of the aggregates as well as *in vitro* evidence that they affect each other has brought into question whether the tau aggregation and α-synuclein disease classes are distinct from each other. It is plausible that these diseases intermingle, and have almost cooperative effects with one disease promoting the pathological mechanisms of another.[153]

Different diseases that display tau aggregation have different fibril morphology. Recently, cryogenic electron microscopy structures were built using samples extracted from Alzheimer's patients, and multiple distinct morphologies were observed.[156] Additionally, fibrils harvested from different diseases such as Pick's disease and Alzheimer's disease have unique morpholgoies.[157] When used to template further aggregation in mice models, different morphologies or "strains" of tau aggregation induce distinct physiological effects. Additionally, when fibril fragments were used to propagate further aggregation in cell models the strains were mostly self-replicating. When the morphologies changed there were some morphologies favored over others. This suggests that there is not only flexibility in fibril formation, but that the formation of some morphologies is preferred over others.[158]

In pathological settings tau takes a very long time to aggregate – often times the timescale is measured in years or decades. It is not known what incites the aggregation, although factors such as hyperphosphorylation and interaction with lipid bilayers are implicated. The unfolded protein response, encompassing molecular chaperones and the ubiquitin/proteasome system, is also instrumental in repressing tau aggregation and defending neurons from tau pathology. For experimental purposes, it is not viable to examine the aggregation of tau over biological timescales. Instead *in vitro* experiments take advantage of a number of known cofactors that shorten the time of aggregation. Of these cofactors, the polyanionic biopolymer heparin is the most widely used to incite the aggregation of amyloid fibrils for proteins such as tau.[159] The presence of heparin reliably incites amyloid formation, but the fibrils formed as a result display a unique fibril morphology not observed in disease conditions, calling into question the utility of using it to aid in biophysical studies.[160]

When studying aggregation processes, there are a number of tau constructs that are useful for studying the mechanics of tau aggregation[161]. Many studies focus on the MTBR as opposed to the full-length protein, as it has been well characterized to recapitulate the aggregation and pathological membrane binding activity of full-length tau. Under otherwise identical conditions, the aggregation of the MTBR of tau is an order of magnitude faster than the full-length construct and thereby serves as a useful model construct in studies focusing on tau aggregation.

. Attempts to understand the role of disordered proteins in biology often come from bioinformatics approaches that look for trends in large sets of data, or in biochemical studies that alter their sequence or expression. While these techniques are obviously valuable and give insight to the roles that disordered proteins have they are limited. Bioinformatics approaches identify insightful trends such as involvement in cancer or signaling pathways, but cannot ascribe

specific roles in those pathways. Biochemical methods that change the sequence of IDPs or ablate their expression can point towards the specific actions or roles of disordered proteins, but struggle to show incremental differences or the effects of temporary changes. Modulation of protein behavior with a small molecule ligand would allow for more fine-grained and temporally defined methods of determining their roles in biology. This in turn would be useful for our understanding of their role in disease, but also in other more fundamental ways such as probing the transcription network or understanding how protein disorder impacts protein-protein interactions in the cell.

At the moment there is a lack of methods sufficiently attuned to the nature of IDPs that can rapidly and efficiently develop small molecule ligands targeting them. This thesis project seeks to develop tools necessary for identifying and developing small molecule ligands targeting disordered proteins. This work is built upon the hypothesis that despite the dynamic nature of disordered proteins, there are structures, either local or global, that can be targeted by small molecules and change the behavior of the disordered protein. Chapter 2 details a method that takes advantage of in silico and in vitro methods to determine locally persistent structures and efficiently identify ligands that bind to the MTBR of tau. Chapter 3 then expands upon that work to elucidate structural activity relationships of these compounds and expand upon them finding more active compounds that are chemically related to the lead compounds identified in chapter 2. Chapter 4 shows that these methods can be adjusted to make use of additional information available depending on the system. By targeting relevant global conformations of phenol soluble modulins and then applying a similar method of computational screening that was developed in chapter 1, multiple ligands were found that affect the membrane interaction and aggregation of short disordered peptides that are involved in mechanisms of antibacterial resistance.

Figures

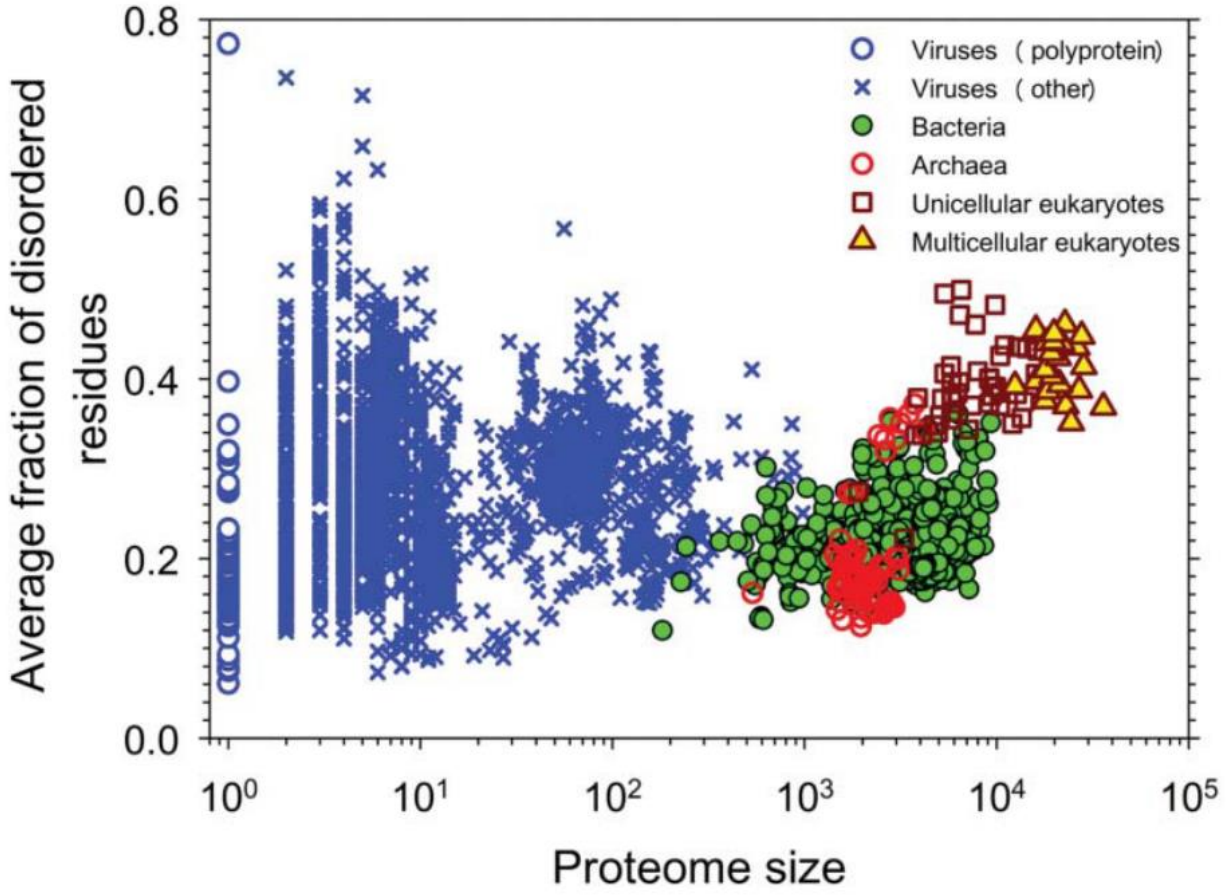


Figure 1.1: More complex proteomes display higher disordered content. Reproduced from Uversky 2015.[162]

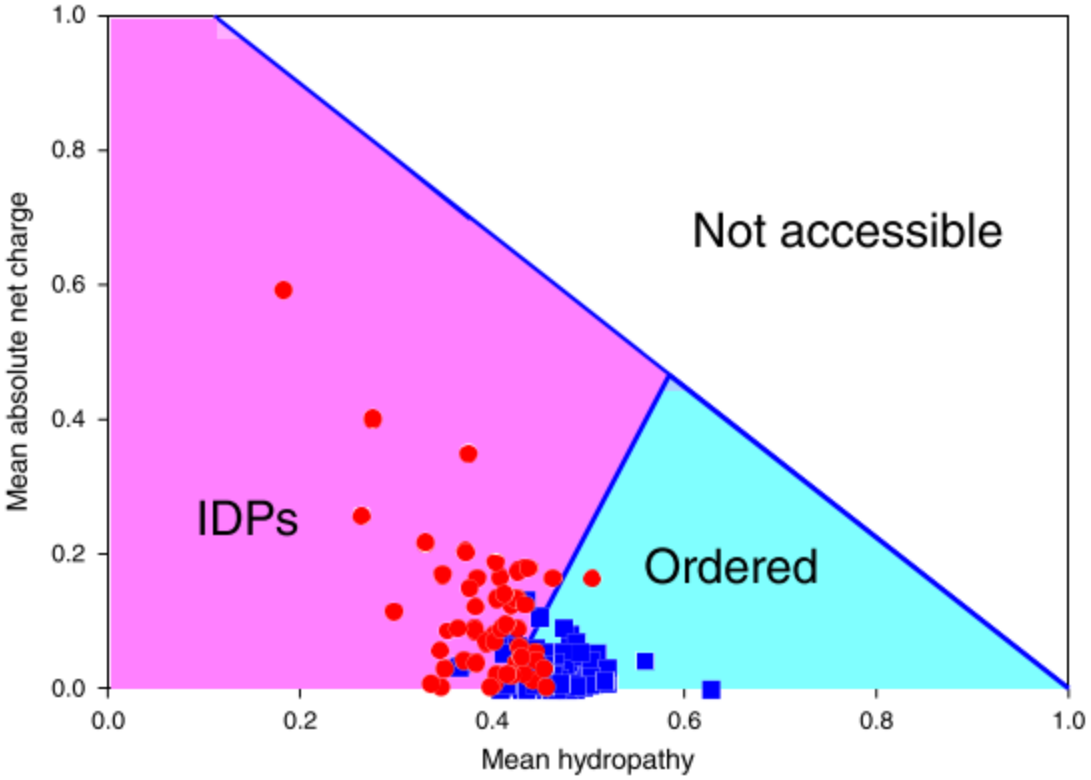


Figure 1.2: There is a relationship between hydropathy and net charge that differentiates the regions occupied by ordered (blue squares) and disordered (red circles) proteins. Reproduced from Uversky 2013.[163]

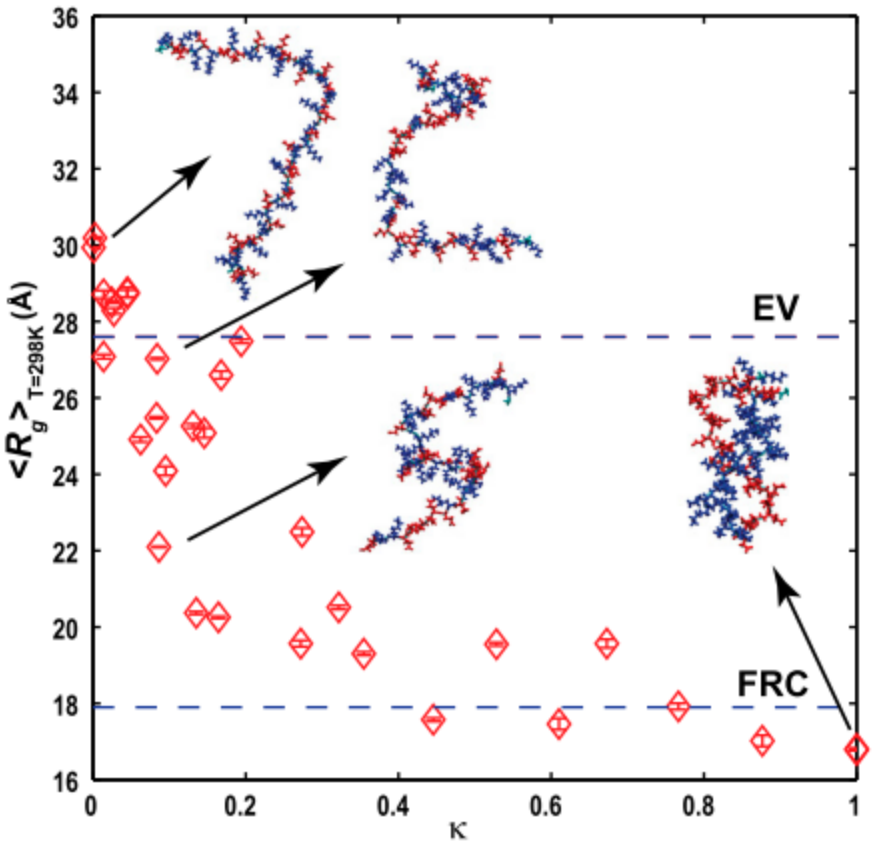


Figure 1.3: Computational studies performed by the Pappu group show *in silico* radius of gyration for different peptides is related to the parameter κ, which is a metric for how well mixed a peptide's charges are. Proteins with a low κ-value have their charge residues well mixed, and tend to be more extended than their more segregated counterparts, and behave more like theoretical random coils as a result.[24]

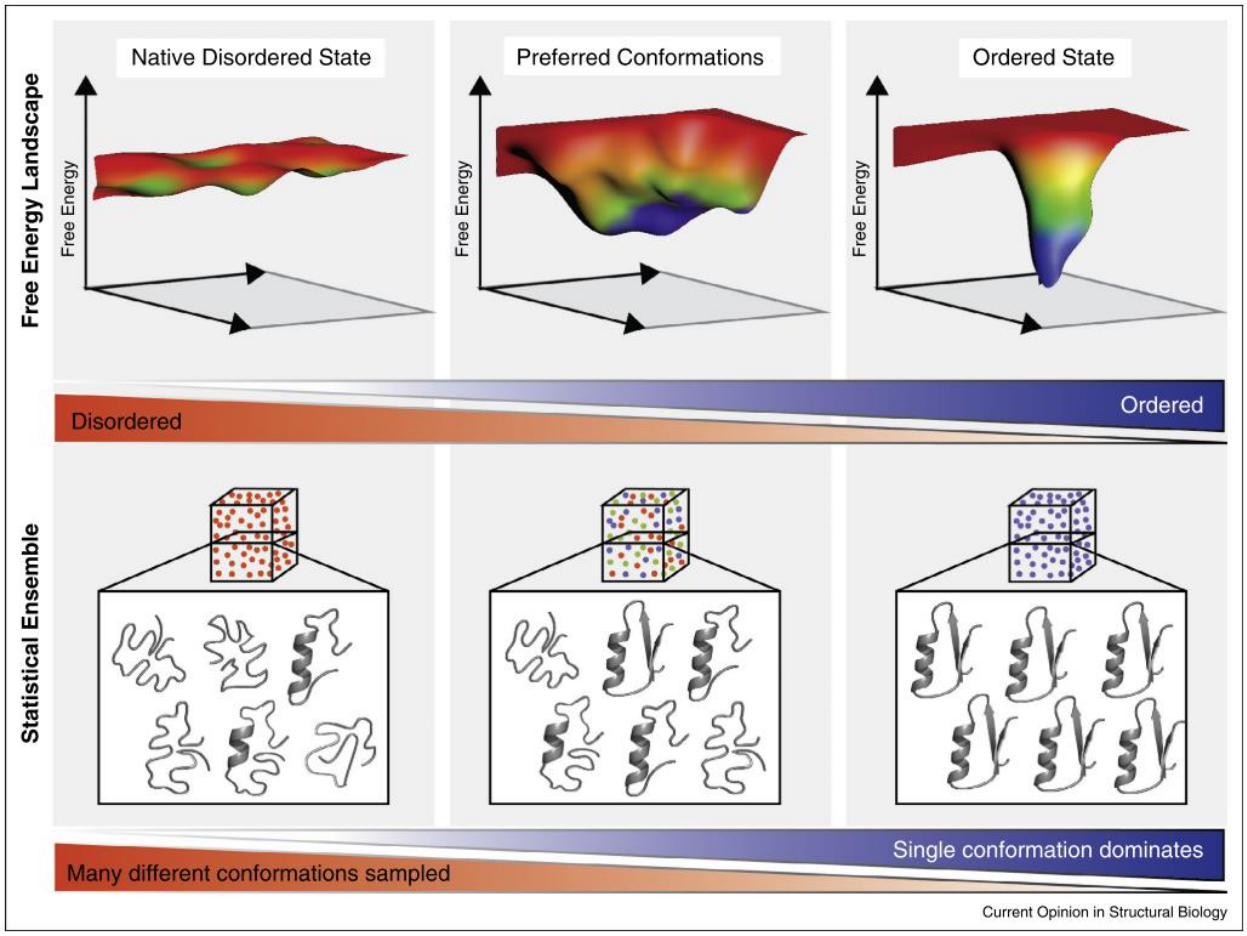


Figure 1.4: The energy landscapes (top) for disordered proteins can change with various chemical environments. Some conditions result in a relatively flat energy landscape where no particular conformation is encouraged and the ensemble of conformations (bottom) is very diverse. As conditions push towards a specific conformation, the energy landscape will demonstrate more prominent global minima and the conformational ensemble with consist of more homologous conformations. Used with permission from Flock, 2014.[52]

References

(1) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293* (2), 321–331.

(2) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (3), 197–208.

(3) Lemieux, R. U.; Spohr, U. How Emil Fischer Was Led To The Lock and Key Concept for Enzyme Specificity. In *Advances in Carbohydrate Chemistry and Biochemistry*; 1994; Vol. 50, pp 1–20.

(4) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme [The Influence of Configuration on the Effect of Enzymes]. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985–2993.

(5) Uversky, V. N.; Dunker, A. K. *Understanding Protein Non-Folding*; 2010; Vol. 1804, pp 1231–1264.

(6) Cast, K.; Damaschun, H.; Eckert, K.; Schulze-Forster, K.; Maurer, H. R.; Miiller-Frohne, M.; Zirwer, D.; Czamecki, J.; Damaschun, G. *Prothymosin a: A Biologically Active Protein with Random Coil Conformation*; 1995; Vol. 34.

(7) Isbell, D. T.; Du, S.; Schroering, A. G.; Colombo, G.; Shelling, J. G. *Metal Ion Binding to Dog Osteocalcin Studied by NMR Spectroscopy ^*; 1993; Vol. 32.

(8) Fisher, W. R.; Taniuchi, H.; And, $; Anfinsen, C. B. *On the Role of Heme in the Formation of the Structure of Cytochrome C*; 1973; Vol. 248.

(9) Dolgikh, D. A.; Gilmanshin, R. I.; Brazhnikov, E. V.; Bychkova, V. E.; Semisotnov, G. V.; Venyaminov, S. Y.; Ptitsyn, O. B. α-Lactalbumin: Compact State with Fluctuating Tertiary Structure? *FEBS Lett.* **1981**, *136* (2), 311–315.

(10) Huber, R.; Bennett, W. S. Functional Significance of Flexibility in Proteins. *Biopolymers* **1983**, *22* (1), 261–279.

(11) Sigler, P. B. Acid Blobs and Negative Noodles. *Nature* **1988**, *333* (6170), 210–212.

(12) Holt, C.; Sawyer, L. Caseins as Rheomorphic Proteins: Interpretation of Primary and Secondary Structures of the As1-, β- and κ-Caseins. *J. Chem. Soc. Faraday Trans.* **1993**, *89* (15), 2683–2692.

(13) Pontius, B. W. Close Encounters: Why Unstructured, Polymeric Domains Can Increase Rates of Specific Macromolecular Association. *Trends Biochem. Sci.* **1993**, *18* (5), 181–186.

(14) Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, J. E.; Dunker, A. K. Identifying Disordered Regions in Proteins from Amino Acid Sequence. *IEEE Int. Conf. Neural Networks - Conf. Proc.* **1997**, *1*, 90–95.

(15) Romero, P.; Obradovic, Z.; Kissinger, C. R.; Villafranca, J. E.; Garner, E.; Guilliot, S.; Dunker, A. K. Thousands of Proteins Likely to Have Long Disordered Regions. *Pac. Symp. Biocomput.* **1998**, 437–448.

(16) Dunker, A. K.; Obradovic, Z.; Romero, P.; Garner, E. C.; Brown, C. J. *Intrinsic Protein Disorder in Complete Genomes*; 2000; Vol. 11.

(17) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; et al. Intrinsically Disordered Protein. *J. Mol. Graph. Model.* **2001**, *3263* (00), 26–59.

(18) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K. Intrinsic Disorder in Cell-Signaling and Cancer-Associated Proteins. *J. Mol. Biol.* **2002**, *323* (3), 573–584.

(19) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why Are “natively Unfolded” Proteins

Unstructured under Physiologic Conditions? *Proteins Struct. Funct. Genet.* **2000**, *41* (3), 415–427.
(20) Williams, R. M.; Obradovi, Z.; Mathura, V.; Braun, W.; Garner, E. C.; Young, J.; Takayama, S.; Brown, C. J.; Dunker, A. K. The Protein Non-Folding Problem: Amino Acid Determinants of Intrinsic Order and Disorder. *Pac. Symp. Biocomput.* **2001**, 89–100.
(21) Campen, A.; Williams, R.; Brown, C.; Meng, J.; Uversky, V.; Dunker, A. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.* **2008**, *15* (9), 956–963.
(22) Xue, B.; Dunker, A. K.; Uversky, V. N. Orderly Order in Protein Intrinsic Disorder Distribution: Disorder in 3500 Proteomes from Viruses and the Three Domains of Life. *J. Biomol. Struct. Dyn.* **2012**, *30* (2), 137–149.
(23) Goto, Y.; Takahashi, N.; Fink, A. L. Mechanism of Acid-Induced Folding of Proteins. *Biochemistry* **1990**, *29* (14), 3480–3488.
(24) Das, R. K.; Pappu, R. V. Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (33), 13392–13397.
(25) Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. Flavors of Protein Disorder. *Proteins Struct. Funct. Genet.* **2003**, *52* (4), 573–584.
(26) Martin, A. J. M.; Walsh, I.; Tosatto, S. C. E. MOBI: A Web Server to Define and Visualize Structural Mobility in NMR Protein Ensembles. *Bioinforma. Appl. NOTE* **2010**, *26* (22), 2916–2917.
(27) Bellay, J.; Han, S.; Michaut, M.; Kim, T.; Costanzo, M.; Andrews, B. J.; Boone, C.; Bader, G. D.; Myers, C. L.; Kim, P. M. *Bringing Order to Protein Disorder through Comparative Genomics and Genetic Interactions*; 2011.
(28) Walsh, I.; Giollo, M.; Domenico, T. A. Di; Ferrari, C.; Zimmermann, O.; Tosatto, S. C. E. E.; Di Domenico, T.; Ferrari, C.; Zimmermann, O.; Tosatto, S. C. E. E. Comprehensive Large-Scale Assessment of Intrinsic Protein Disorder. *Bioinformatics* **2015**, *31* (2), 201–208.
(29) Levinthal, C. *ARE THERE PATHWAYS FOR PROTEIN FOLDING ?*; 1968; Vol. 65.
(30) Koide, T.; Odani, S.; Ono, T.; Laemmli, U. K.; Leung, L. L. K.; Harpel, P. C.; Nachman, R. L.; Rabellino, E. M.; Levine, R. L.; Federici, . M; et al. *Theory for the Folding and Stability of Globular Proteins1"*; Connolly, 1985; Vol. 24.
(31) Zwanzig, R.; Szabo, A.; Bagchi, B. *Levinthal's Paradox*; 1992; Vol. 89.
(32) Englander, S. W.; Mayne, L. The Nature of Protein Folding Pathways.
(33) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annu. Rev. Biophys.* **2008**, *37* (1), 215–246.
(34) Fisher, C. K.; Stultz, C. M. Constructing Ensembles for Intrinsically Disordered Proteins. **2011**.
(35) Dill, K. A.; Chan, H. S. *From Levinthal to Pathways to Funnels*; 1997.
(36) Ward, J. J. J.; Sodhi, J. S. S.; McGuffin, L. J. J.; Buxton, B. F. F.; Jones, D. T. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337* (3), 635–645.
(37) Turoverov, K. K.; Kuznetsova, I. M.; Uversky, V. N. The Protein Kingdom Extended: Ordered and Intrinsically Disordered Proteins, Their Folding, Supramolecular Complex Formation, and Aggregation. *Prog. Biophys. Mol. Biol.* **2010**, *102* (2–3), 73–84.
(38) Kim, P. M.; Sboner, A.; Xia, Y.; Gerstein, M. The Role of Disorder in Interaction

Networks: A Structural Analysis. *Mol. Syst. Biol.* **2008**.
(39) Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N. Flexible Nets. The Roles of Intrinsic Disorder in Protein Interaction Networks. *FEBS J.* **2005**, *272* (20), 5129–5148.
(40) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signaling and Regulation HHS Public Access. *Nat Rev Mol Cell Biol* **2015**, *16* (1), 18–29.
(41) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Showing Your ID: Intrinsic Disorder as an ID for Recognition, Regulation and Cell Signaling. *J. Mol. Recognit.* **2005**, *18* (5), 343–384.
(42) Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K. Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements. *Biochemistry* **2005**, *44* (37), 12454–12470.
(43) Sugase, K.; Dyson, H. J.; Wright, P. E. Mechanism of Coupled Folding and Binding of an Intrinsically Disordered Protein. *Nature* **2007**, *447* (7147), 1021–1025.
(44) Dyson, H. J.; Wright, P. E. Coupling of Folding and Binding for Unstructured Proteins. *Current Opinion in Structural Biology*. 2002, pp 54–60.
(45) Tompa, P.; Fuxreiter, M. Fuzzy Complexes: Polymorphism and Structural Disorder in Protein-Protein Interactions. *Trends Biochem. Sci.* **2008**, *33* (1), 2–8.
(46) Kriwacki, R. W.; Hengst, L.; Tenna, L. *Structural Studies of P21 Waf1/CiPla State: Conformational Disorder i (Cell Cycle Regulation/NMR Spectroscopy/Protein Folding/Kina*; 1996; Vol. 93.
(47) Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. *Speeding Molecular Recognition by Using the Folding Funnel: The Fly-Casting Mechanism*; 2000; Vol. 97.
(48) Shammas, S. L.; Travis, A. J.; Clarke, J. Remarkably Fast Coupled Folding and Binding of the Intrinsically Disordered Transactivation Domain of CMyb to CBP KIX. *J. Phys. Chem. B* **2013**, *117* (42), 13346–13356.
(49) Arai, M.; Ferreon, J. C.; Wright, P. E. Quantitative Analysis of Multisite Protein−Ligand Interactions by NMR: Binding of Intrinsically Disordered P53 Transactivation Subdomains with the TAZ2 Domain of CBP. *J. Am. Chem. Soc* **2012**, *134*, 22.
(50) Mollica, L.; Bessa, L. M.; Hanoulle, X.; Jensen, M. R.; Blackledge, M.; Schneider, R. Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Frontiers in Molecular Biosciences*. 2016, p 52.
(51) Zhou, H. X.; Pang, X.; Lu, C. Rate Constants and Mechanisms of Intrinsically Disordered Proteins Binding to Structured Targets. *Physical Chemistry Chemical Physics*. 2012, pp 10466–10476.
(52) Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M. Controlling Entropy to Tune the Functions of Intrinsically Disordered Regions. *Current Opinion in Structural Biology*. Elsevier Ltd 2014, pp 62–72.
(53) Vavouri, T.; Semple, J. I.; Garcia-Verdugo, R.; Lehner, B. Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell* **2009**, *138* (1), 198–208.
(54) Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Madan Babu, M.; Babu, M. M. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science (80-.).* **2008**, *322* (November), 1365–1368.
(55) Madan Babu, M.; Van Der Lee, R.; Sanchez De Groot, N.; Rg Gsponer, J.; Gough, J.; Dunker, K. Intrinsically Disordered Proteins: Regulation and Disease This Review Comes

from a Themed Issue on Sequences and Topology Edited. *Curr. Opin. Struct. Biol.* **2011**, *21*, 1–9.
(56) Edwards, Y. J.; Lobley, A. E.; Pentony, M. M.; Jones, D. T. Insights into the Regulation of Intrinsically Disordered Proteins in the Human Proteome by Analyzing Sequence and Gene Expression Data. *Genome Biol.* **2009**, *10* (5), 50.
(57) Babu, M. M. The Contribution of Intrinsically Disordered Regions to Protein Function, Cellular Complexity, and Human Disease. *Biochemical Society Transactions*. 2016, pp 1185–1200.
(58) Weatheritt, R. J.; Gibson, T. J.; Babu, M. M. Asymmetric MRNA Localization Contributes to Fidelity and Sensitivity of Spatially Localized Systems. *Nat. Struct. Mol. Biol.* **2014**, *21* (9), 833–839.
(59) Weber, S. C.; Brangwynne, C. P. Getting RNA and Protein in Phase. *Cell*. Elsevier June 8, 2012, pp 1188–1191.
(60) Lemke, E. A. The Multiple Faces of Disordered Nucleoporins. *Journal of Molecular Biology*. Academic Press May 2016, pp 2011–2024.
(61) Mackenzie, I. R.; Rademakers, R.; Neumann, M. TDP-43 and FUS in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Lancet Neurol.* **2010**, *9* (10), 995–1007.
(62) Metskas, L. A.; Rhoades, E. Order–Disorder Transitions in the Cardiac Troponin Complex. *Journal of Molecular Biology*. NIH Public Access July 2016, pp 2965–2977.
(63) Darling, A. L.; Uversky, V. N. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Frontiers in Genetics*. Frontiers May 4, 2018, p 158.
(64) Pejaver, V.; Hsu, W. L.; Xin, F.; Dunker, A. K.; Uversky, V. N.; Radivojac, P. The Structural and Functional Signatures of Proteins That Undergo Multiple Events of Post-Translational Modification. *Protein Sci.* **2014**, *23* (8), 1077–1093.
(65) Madan Babu, M.; Van Der Lee, R.; Sanchez De Groot, N.; Rg Gsponer, J.; Gough, J.; Dunker, K. Intrinsically Disordered Proteins: Regulation and Disease. *Curr. Opin. Struct. Biol.* **2011**, *21*, 1–9.
(66) Bah, A.; Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-Translational Modifications. *J. Biol. Chem.* **2016**, *291* (13), 6696–6705.
(67) Monahan, Z.; Ryan, V. H.; Janke, A. M.; Burke, K. A.; Rhoads, S. N.; Zerze, G. H.; O'Meally, R.; Dignon, G. L.; Conicella, A. E.; Zheng, W.; et al. Phosphorylation of the FUS Low-complexity Domain Disrupts Phase Separation, Aggregation, and Toxicity . *EMBO J.* **2017**, *36* (20), 2951–2967.
(68) Patil, A.; Nakamura, H. Disordered Domains and High Surface Charge Confer Hubs with the Ability to Interact with Multiple Proteins in Interaction Networks. *FEBS Lett.* **2006**, *580* (8), 2041–2045.
(69) Müller-Späth, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Rüegger, S.; Reymond, L.; Nettels, D.; Schuler, B. Charge Interactions Can Dominate the Dimensions of Intrinsically Disordered Proteins.
(70) Wicky, B. I. M.; Shammas, S. L.; Clarke, J. Affinity of IDPs to Their Targets Is Modulated by Ion-Specific Changes in Kinetics and Residual Structure. **2017**, *114* (37).
(71) Levine, Z. A.; Larini, L.; LaPointe, N. E.; Feinstein, S. C.; Shea, J.-E. Regulation and Aggregation of Intrinsically Disordered Peptides. *Proc. Natl. Acad. Sci.* **2015**, *112* (9), 2758–2763.
(72) Moosa, M. M.; Ferreon, A. C. M.; Deniz, A. A. Forced Folding of a Disordered Protein

Accesses an Alternative Folding Landscape. *ChemPhysChem* **2015**, *16* (1), 90–94.
(73) Mittag, T.; Forman-Kay, J. D. Atomic-Level Characterization of Disordered Protein Ensembles. *Curr. Opin. Struct. Biol.* **2007**, *17* (1), 3–14.
(74) Gibbs, E. B.; Cook, E. C.; Showalter, S. A. Application of NMR to Studies of Intrinsically Disordered Proteins. *Arch. Biochem. Biophys.* **2017**, *628*, 57–70.
(75) Jensen, M. R.; Ruigrok, R. W.; Blackledge, M. Describing Intrinsically Disordered Proteins at Atomic Resolution by NMR. *Current Opinion in Structural Biology*. 2013, pp 426–435.
(76) Baughman, H. E. R.; Clouser, A. F.; Klevit, R. E.; Nath, A. Chaperone Effects on Tau Fibril Formation HspB1 and Hsc70 Chaperones Engage Distinct Tau Species and Have Different Inhibitory Effects on Amyloid Formation. *J. Biol. Chem.* **2018**, *293* (8).
(77) Baughman, H. E. R.; Clouser, A. F.; Klevit, R. E.; Nath, A. HspB1 and Hsc70 Chaperones Engage Distinct Tau Species and Have Different Inhibitory Effects on Amyloid Formation. *J. Biol. Chem.* **2018**, *293* (8), 2687–2700.
(78) Gibbs, E. B.; Showalter, S. A. Quantitative Biophysical Characterization of Intrinsically Disordered Proteins. *Biochemistry* **2015**, *54* (6), 1314–1326.
(79) Gong, H.; Zhang, S.; Wang, J.; Gong, H.; Zeng, J. Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data. *J. Comput. Biol.* **2016**, *23* (5), 300–310.
(80) Marius Clore, G.; Iwahara, J. Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. *Chem. Rev.* **2009**, *109* (9), 4108–4139.
(81) Clouser, A. F.; Baughman, H. E.; Basanta, B.; Guttman, M.; Nath, A.; Klevit, R. E. Interplay of Disordered and Ordered Regions of a Human Small Heat Shock Protein Yields an Ensemble of "quasi-Ordered" States. *Elife* **2019**, *8*, 1–31.
(82) Hennig, J.; Sattler, M. The Dynamic Duo: Combining NMR and Small Angle Scattering in Structural Biology. *Protein Science*. Blackwell Publishing Ltd 2014, pp 669–682.
(83) Sibille, N.; Bernadó, P. Structural Characterization of Intrinsically Disordered Proteins by the Combined Use of NMR and SAXS. *Biochemical Society Transactions*. 2012, pp 956–962.
(84) Rambo, R. P.; Tainer, J. A. Characterizing Flexible and Intrinsically Unstructured Biological Macromolecules by SAS Using the Porod-Debye Law. *Biopolymers* **2011**, *95* (8), 559–571.
(85) Bernadó, P.; Svergun, D. I. Structural Analysis of Intrinsically Disordered Proteins by Small-Angle. *Society* **2012**, *8*, 151–167.
(86) Krzeminski, M. L.; Marsh, J. A.; Neale, C.; Choy, W.-Y.; Forman-Kay, J. D. Characterization of Disordered Proteins with ENSEMBLE. *Bioinforma. Appl.* **2013**, *29* (3), 398–399.
(87) Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J. D. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* **2010**.
(88) Leblanc, S. J.; Kulkarni, P.; Weninger, K. R. Single Molecule FRET: A Powerful Tool to Study Intrinsically Disordered Proteins. *Biomolecules* **2018**, *8* (4).
(89) Elbaum-Garfinkle, S.; Rhoades, E. Identification of an Aggregation-Prone Structure of Tau. *J. Am. Chem. Soc.* **2012**, *134* (40), 16607–16613.
(90) Choi, U. B.; Xiao, S.; Wollmuth, L. P.; Bowen, M. E. Effect of Src Kinase

Phosphorylation on Disordered C-Terminal Domain of N-Methyl-D-Aspartic Acid (NMDA) Receptor Subunit GluN2B Protein. *J. Biol. Chem.* **2011**, *286* (34), 29904–29912.
(91) Das, P.; Matysiak, S.; Mittal, J. Looking at the Disordered Proteins through the Computational Microscope. *ACS Cent. Sci.* **2018**, *4* (5), 534–542.
(92) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron*. 2018, pp 1129–1143.
(93) Henriques, J. O.; Cragnell, C.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. **2015**.
(94) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14* (1), 71–73.
(95) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (21), E4758–E4766.
(96) Kasahara, K.; Terazawa, H.; Takahashi, T.; Higo, J. Studies on Molecular Dynamics of Intrinsically Disordered Proteins and Their Fuzzy Complexes: A Mini-Review. *Computational and Structural Biotechnology Journal*. Elsevier B.V. January 1, 2019, pp 712–720.
(97) Qi, R.; Wei, G.; Ma, B.; Nussinov, R. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. In *Methods in Molecular Biology*; 2018; Vol. 1777, pp 101–119.
(98) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.
(99) Qiao, Q.; Bowman, G. R.; Huang, X. Dynamics of an Intrinsically Disordered Protein Reveal Metastable Conformations That Potentially Seed Aggregation. *J. Am. Chem. Soc.* **2013**, *135* (43), 16092–16101.
(100) Nath, A.; Rhoades, E. A Flash in the Pan: Dissecting Dynamic Amyloid Intermediates Using Fluorescence. *FEBS Lett.* **2013**, *587* (8), 1096–1105.
(101) Zhu, M.; De Simone, A.; Schenk, D.; Toth, G.; Dobson, C. M.; Vendruscolo, M. Identification of Small-Molecule Binding Pockets in the Soluble Monomeric Form of the Aβ42 Peptide. *J. Chem. Phys.* **2013**, *139* (3).
(102) De Simone, A.; Derreumaux, P. Low Molecular Weight Oligomers of Amyloid Peptides Display B -Barrel Conformations: A Replica Exchange Molecular Dynamics Study in Explicit Solvent. *J. Chem. Phys.* **2010**, *132* (16).
(103) Naganathan, A. N.; Orozco, M. The Conformational Landscape of an Intrinsically Disordered DNA- Binding Domain of a Transcription Regulator. **2013**, No. iii.
(104) Wu, C.; Shea, J. E. Coarse-Grained Models for Protein Aggregation. *Current Opinion in Structural Biology*. 2011, pp 209–220.
(105) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*. 2016, pp 7898–7936.
(106) Chiti, F.; Dobson, C. M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.
(107) Sunde, M.; Serpell, L. C.; Bartlam, M.; Fraser, P. E.; Pepys, M. B.; Blake, C. C. F. Common Core Structure of Amyloid Fibrils by Synchrotron X-Ray Diffraction. *J. Mol. Biol.* **1997**.

(108) Sipe, J. D.; Cohen, A. S. Review: History of the Amyloid Fibril. *J. Struct. Biol.* **2000**, *130* (2–3), 88–98.
(109) Astbury, W. T.; Dickinson, S.; Bailey, K. The X-Ray Interpretation of Denaturation and the Structure of the Seed Globulins. *Biochem. J.* **1935**, *29* (10), 2351-2360.1.
(110) Baldwin, A. J.; Knowles, T. P. J.; Tartaglia, G. G.; Fitzpatrick, A. W.; Devlin, G. L.; Shammas, S. L.; Waudby, C. A.; Mossuto, M. F.; Meehan, S.; Gras, S. L.; et al. Metastability of Native Proteins and the Phenomenon of Amyloid Formation.Pdf. **2011**, 14160–14163.
(111) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. The Amyloid State and Its Association with Protein Misfolding Diseases. *Nat. Rev. Mol. Cell Biol.* **2014**, *15* (6), 384–396.
(112) Naiki, H.; Hashimoto, N.; Suzuki, S.; Kimura, H.; Nakakuki, K.; Gejyo, F. Establishment of a Kinetic Model of Dialysis-Related Amyloid Fibril Extension in Vitro. *Amyloid* **1997**, *4* (4), 223–232.
(113) Merlini, G.; Bellotti, V. Molecular Mechanisms of Amyloidosis. *New England Journal of Medicine*. 2003, pp 583–596.
(114) Serio, T. R.; Cashikar, A. G.; Kowal, A. S.; Sawicki, G. J.; Moslehi, J. J.; Serpell, L.; Arnsdorf, M. F.; Lindquist, S. L. *Nucleated Conformational Conversion and the Replication of Conformational Information by a Prion Determinant Downloaded From*; 2000; Vol. 289.
(115) Uversky, V. N.; Fink, A. L. Conformational Constraints for Amyloid Fibrillation: The Importance of Being Unfolded. *Biochimica et Biophysica Acta - Proteins and Proteomics*. 2004, pp 131–153.
(116) Eisenberg, D.; Jucker, M. The Amyloid State of Proteins in Human Diseases. *Cell*. NIH Public Access March 2012, pp 1188–1203.
(117) Uversky, V. N. The Triple Power of D3: Protein Intrinsic Disorder in Degenerative Diseases Vladimir. *Front. Biosci.* **2014**, *19*, 181–258.
(118) Pepys, M. B. Amyloidosis. *Annu. Rev. Med* **2006**, *57*, 223–264.
(119) Lorenzo, A.; Yankner, B. A. β-Amyloid Neurotoxicity Requires Fibril Formation and Is Inhibited by Congo Red. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91* (25), 12243–12247.
(120) Lue, L. F.; Kuo, Y. M.; Roher, A. E.; Brachova, L.; Shen, Y.; Sue, L.; Beach, T.; Kurth, J. H.; Rydel, R. E.; Rogers, J. Soluble Amyloid β Peptide Concentration as a Predictor of Synaptic Change in Alzheimer's Disease. *Am. J. Pathol.* **1999**.
(121) Larson, J.; Lynch, G.; Games, D.; Seubert, P. Alterations in Synaptic Transmission and Long-Term Potentiation in Hippocampal Slices from Young and Aged PDAPP Mice. *Brain Res.* **1999**, *840* (1–2), 23–35.
(122) Chiti, Fabrizio., Dobson, C. Amyloid Formation, Protein Homeostasis, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **2017**, *86* (1), 1–42.
(123) Ghag, G.; Bhatt, N.; Cantu, D. V.; Guerrero-Munoz, M. J.; Ellsworth, A.; Sengupta, U.; Kayed, R. Soluble Tau Aggregates, Not Large Fibrils, Are the Toxic Species That Display Seeding and Cross-Seeding Behavior. *Protein Sci.* **2018**, *27* (11), 1901–1909.
(124) Frackowiak, J.; Zoltowska, A.; Wisniewski, H. M. Non-Fibrillar β-Amyloid Protein Is Associated with Smooth Muscle Cells of Vessel Walls in Alzheimer Disease. *Journal of Neuropathology and Experimental Neurology*. 1994, pp 637–645.
(125) Cline, E. N.; Bicca, M. A.; Viola, K. L.; Klein, W. L. The Amyloid-β Oligomer

Hypothesis: Beginning of the Third Decade. *Journal of Alzheimer's Disease*. 2018, pp S567–S610.
(126) Maji, S. K.; Perrin, M. H.; Sawaya, M. R.; Jessberger, S.; Vadodaria, K.; Rissman, R. A.; Singru, P. S.; Peter, K.; Nilsson, R.; Simon, R.; et al. *Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules Downloaded From*; 2009; Vol. 325.
(127) O'toole, G.; Kaplan, H. B.; Kolter, R. *Biofilm Formation As Microbial Development*; 2000.
(128) Jamal, M., Tasneem, U., Hussain, T., & Andleeb, and S. A. *Historical Background of Biofilm*; 2015; Vol. 4.
(129) Otto, M. Phenol-Soluble Modulins. *International Journal of Medical Microbiology*. March 2014, pp 164–169.
(130) Periasamy, S.; Joo, H. S.; Duong, A. C.; Bach, T. H. L.; Tan, V. Y.; Chatterjee, S. S.; Cheung, G. Y. C.; Otto, M. How Staphylococcus Aureus Biofilms Develop Their Characteristic Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (4), 1281–1286.
(131) Cheung, G. Y. C.; Duong, A. C.; Otto, M. Direct and Synergistic Hemolysis Caused by Staphylococcus Phenol-Soluble Modulins: Implications for Diagnosis and Pathogenesis. *Microbes Infect.* **2012**, *14*, 380–386.
(132) Metallo, S. J. Intrinsically Disordered Proteins Are Potential Drug Targets. *Curr. Opin. Chem. Biol.* **2010**, *14* (4), 481–488.
(133) Wang, J.; Cao, Z.; Zhao, L.; Li, S. Novel Strategies for Drug Discovery Based on Intrinsically Disordered Proteins (IDPs). *Int. J. Mol. Sci.* **2011**, *12* (5), 3205–3219.
(134) Hammoudeh, D. I.; Follis, A. V.; Prochownik, E. V; Metallo, S. J. Multiple Independent Binding Sites for Small-Molecule Inhibitors on the Oncoprotein c-Myc. *J. Am. Chem. Soc.* **2009**, *131* (21), 7390–7401.
(135) Dang, C. V. C-Myc Target Genes Involved in Cell Growth, Apoptosis, and Metabolism. *Mol. Cell. Biol.* **1999**, *19* (1), 1–11.
(136) Yin, X.; Giap, C.; Lazo, J. S.; Prochownik, E. V. Low Molecular Weight Inhibitors of Myc–Max Interaction and Function. *Oncogene* **2003**, *22* (40), 6151–6159.
(137) Wang, H.; Hammoudeh, D. I.; Follis, A. V.; Reese, B. E.; Lazo, J. S.; Metallo, S. J.; Prochownik, E. V. Improved Low Molecular Weight Myc-Max Inhibitors. *Mol. Cancer Ther.* **2007**, *6* (9), 2399–2408.
(138) Follis, A. V.; Hammoudeh, D. I.; Wang, H.; Prochownik, E. V.; Metallo, S. J. Structural Rationale for the Coupled Binding and Unfolding of the C-Myc Oncoprotein by Small Molecules. *Chem. Biol.* **2008**, *15* (11), 1149–1155.
(139) Tóth, G.; Gardai, S. J.; Zago, W.; Bertoncini, C. W.; Cremades, N.; Roy, S. L.; Tambe, M. A.; Rochet, J.-C. C.; Galvagnion, C.; Skibinski, G.; et al. Targeting the Intrinsically Disordered Structural Ensemble of α-Synuclein by Small Molecules as a Potential Therapeutic Strategy for Parkinson's Disease. *PLoS One* **2014**, *9* (2), e87133.
(140) Drubin, D. G.; Kirschner, M. W. Tau Protein Function in Living Cells. *J. Cell Biol.* **1986**, *103* (6 Pt 2), 2739–2746.
(141) Chen, J.; Kanai, Y.; Cowan, N. J.; Hirokawa, N. Projection Domains of MAP2 and Tau Determine Spacings between Microtubules in Dendrites and Axons. *Nature* **1992**, *360* (6405), 674–677.
(142) Brandt, R.; Léger, J.; Lee, G. *Interaction of Tau with the Neural Plasma Membrane Mediated by Tau's Amino-Terminal*; 1995; Vol. 131.

(143) Feinstein, H. E.; Benbow, S. J.; LaPointe, N. E.; Patel, N.; Ramachandran, S.; Do, T. D.; Gaylord, M. R.; Huskey, N. E.; Dressler, N.; Korff, M.; et al. Oligomerization of the Microtubule-Associated Protein Tau Is Mediated by Its N-Terminal Sequences: Implications for Normal and Pathological Tau Action. *J. Neurochem.* **2016**, 939–954.
(144) Li, X.-H. H.; Rhoades, E. Heterogeneous Tau-Tubulin Complexes Accelerate Microtubule Polymerization. *Biophys. J.* **2017**, *112* (12), 2567–2574.
(145) McKibben, K.; Rhoades, E. Regulation of Tau's Proline Rich Region by Its N-Terminal Domain. *bioRxiv* **2019**.
(146) Goedert, M.; Spillantini, M. G.; Jakes, R.; Rutherford, D.; Crowther, R. A. Multiple Isoforms of Human Microtubule-Associated Protein Tau: Sequences and Localization in Neurofibrillary Tangles of Alzheimer's Disease. *Neuron* **1989**, *3* (4), 519–526.
(147) Goedert, M.; Jakes, R. Expression of Separate Isoforms of Human Tau Protein: Correlation with the Tau Pattern in Brain and Effects on Tubulin Polymerization. *EMBO J.* **1990**, *9* (13), 4225–4230.
(148) Buée, L.; Bussière, T.; Buée-Scherrer, V.; Delacourte, A.; Hof, P. R. Tau Protein Isoforms, Phosphorylation and Role in Neurodegenerative Disorders. *Brain Res. Rev.* **2000**, *33* (1), 95–130.
(149) Chong, F. P.; Ng, K. Y.; Koh, R. Y.; Chye, S. M. Tau Proteins and Tauopathies in Alzheimer's Disease. *Cellular and Molecular Neurobiology*. Springer New York LLC July 1, 2018, pp 965–980.
(150) Lee, G.; Leugers, C. J. Tau and Tauopathies. *Prog. Mol. Biol. Transl. Sci.* **2012**, *107*, 263–293.
(151) Strang, K. H.; Sorrentino, Z. A.; Riffe, C. J.; Gorion, K. M. M.; Vijayaraghavan, N.; Golde, T. E.; Giasson, B. I. Phosphorylation of Serine 305 in Tau Inhibits Aggregation. *Neurosci. Lett.* **2019**, *692*, 187–192.
(152) Wang, Y.; Mandelkow, E. Tau in Physiology and Pathology. *Nat. Rev. Neurosci.* **2015**, *17* (1), 22–35.
(153) Moussaud, S.; Jones, D. R.; Moussaud-Lamodière, E. L.; Delenclos, M.; Ross, O. A.; McLean, P. J. Alpha-Synuclein and Tau: Teammates in Neurodegeneration? *Molecular neurodegeneration*. 2014, p 43.
(154) Hamilton, R. L. Lewy Bodies in Alzheimer's Disease: A Neuropathological Review of 145 Cases Using α-Synuclein Immunohistochemistry. *Brain Pathol.* **2006**, *10* (3), 378–384.
(155) Badiola, N.; de Oliveira, R. M.; Herrera, F.; Guardia-Laguarta, C.; Gonçalves, S. A.; Pera, M.; Suárez-Calvet, M.; Clarimon, J.; Outeiro, T. F.; Lleó, A. Tau Enhances α-Synuclein Aggregation and Toxicity in Cellular Models of Synucleinopathy. *PLoS One* **2011**, *6* (10).
(156) Fitzpatrick, A. W. P.; Falcon, B.; He, S.; Murzin, A. G.; Murshudov, G.; Garringer, H. J.; Anthony Crowther, R.; Ghetti, B.; Goedert, M.; Scheres, S. H. W. Cryo-EM Structures of Tau Filaments from Alzheimer's Disease. *Nature* **2017**, *547*, 185–190.
(157) Falcon, B.; Zhang, W.; Murzin, A. G.; Murshudov, G.; Garringer, H. J.; Vidal, R.; Crowther, R. A.; Ghetti, B.; Scheres, S. H. W.; Goedert, M. Structures of Filaments from Pick's Disease Reveal a Novel Tau Protein Fold. *Nature* **2018**.
(158) Sharma, A. M.; Thomas, T. L.; Woodard, D. R.; Kashmer, O. M.; Diamond, M. I. Tau Monomer Encodes Strains. *Elife* **2018**, *7*, 1–20.
(159) Goedert, M.; Jakes, R.; Spillantini, M. G.; Hasegawa, M.; Crowther, R. A. Assembly of Microtubule-Associated Protein Tau into Alzheimer-like Filaments Induced by Sulphated

Glycosaminoglycans. *Nature*. 1996, pp 550–553.
(160) Zhang, W.; Falcon, B.; Murzin, A. G.; Fan, J.; Anthony Crowther, R.; Goedert, M.; Scheres, S. H. Heparin-Induced Tau Filaments Are Polymorphic and Differ from Those in Alzheimer's and Pick's Diseases. **2019**.
(161) Gustke, N.; Trinczek, B.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. Domains of τ Protein and Interactions with Microtubules. *Biochemistry* **1994**, *33* (32), 9511–9522.
(162) Uversky, V. N. *Paradoxes and Wonders of Intrinsic Disorder: Prevalence of Exceptionality*; Taylor and Francis Inc., 2015; Vol. 3.
(163) Uversky, V. N. Unusual Biophysics of Intrinsically Disordered Proteins. *Biochimica et Biophysica Acta - Proteins and Proteomics*. Elsevier May 1, 2013, pp 932–951.

Chapter 2
The Rational Discovery of Tau-binding Ligands

Introduction

Current approaches to rational drug design and discovery are frustrated by the highly dynamic nature of many protein targets relevant to major human diseases. Specifically, conventional approaches rely on the twin assumptions that a protein populates a single, well-defined structure, and that a small molecule ligand compatible with that structure can bind specifically to the target so as to interrupt pathological activity or restore normal function. However, a significant fraction (estimated to be 30-50%) of human proteins are completely or partially unstructured.[2] These intrinsically disordered proteins (IDPs) rapidly fluctuate between an ensemble of structures instead of adopting a single well-defined conformation, and cannot be characterized using classical structural biology methods such as X-ray crystallography.[3–5] Consequently, it has been challenging to rationally develop new therapies for diseases linked to IDP function and dysfunction.

IDPs and intrinsically disordered regions (IDRs) are involved in many widespread and severe diseases for several reasons.[6,7] First, IDPs tend to play important roles in cell signaling pathways. Due to their remarkable structural plasticity, IDPs can bind transiently yet specifically to a variety of different partners, serving as hubs for the flow of information in the cell. Aberrant signaling (due to mutation, toxins, stress- or age-related post-translational modification) can drastically impair cellular function, accounting for the fact that most cancers involve the dysfunction of particular IDPs, including the tumor suppressors p53 and BRCA1.[8,9]

Second, the unique physical and chemical properties of IDPs enable them to play vital roles as tunable structural components of subcellular macromolecular assemblies such as non-membrane-bound organelles (NMBOs), the nuclear pore complex, and cytoskeletal architecture.[10–12] For instance, NMBOs such as stress granules and P-bodies form by reversible co-aggregation or co-acervation of specific proteins and RNA resulting in liquid-liquid phase separation and droplet formation within the cytoplasm.[10,13] These droplets appear to perform important metabolic and physiological functions, with their formation being switchable and regulated. Defects in particular IDPs can cause disorders such as amyotrophic lateral sclerosis (TDP-43, FUS etc.) and hypertrophic cardiomyopathy (troponin-C).[14,15]

Third, many IDPs are key players in degenerative amyloid disorders such as Alzheimer's disease (AD), type II diabetes, and Parkinson's disease.[16,17] In these conditions, particular IDPs convert from soluble, monomeric native states into a heterogeneous and dynamic ensemble of toxic intermediates and then into highly ordered β-sheet-rich amyloid aggregates. Amyloid toxicity frequently involves the remodeling and disruption of membranes by partially structured monomeric or oligomeric species.

There is therefore an urgent need for drug discovery strategies targeting IDPs, exemplified by the focus of this study: microtubule-associated protein tau. In its physiological role, tau stabilizes microtubules, regulates axonal trafficking in neurons, and is involved in neurite outgrowth.[18] Tau remains predominantly disordered even when bound to microtubules or tubulin, outside of the residues that interact with these partners.[12,19,20] Amyloid formation by tau (often accompanied by hyperphosphorylation and oxidative modifications) is a hallmark of a range of neurodegenerative disorders, including Alzheimer's disease (AD) and chronic traumatic encephalopathy, that are collectively called tauopathies.[21] AD is also characterized by the

aggregation of the IDP amyloid-β into senile plaques; conversely, tau neurofibrillary tangles are frequently observed in synucleinopathies such as Parkinson's disease (PD) and dementia with Lewy bodies. Membrane interactions are central to the pathological activity of tau: tau causes membrane leakage, disrupts vesicles, and forms lipid-protein co-aggregates.[22–24] Tau occurs in six different isoforms generated by alternative splicing, up to 441 residues long. The central microtubule binding region (MTBR) of tau, consisting of three or four repeats, is responsible both for interacting with tubulin, and for pathological self-association and cytotoxicity.[22,25,26] Two hexapeptide repeats dubbed PHF6 and PHF6* are important drivers of aggregation *in vitro*, and (along with surrounding portions of the MTBR) form the structured core of tau fibrils.[27–29] We focus our studies on the tau4RD construct (residues 244–372, comprising all four repeats of the MTBR), which recapitulates tau amyloidogenesis and membrane interactions.[24,30] Tau is a particularly intriguing model system because it displays the full gamut of IDP behavior: disorder is important to its native function, and it remains largely unstructured even when bound to microtubules;[12,20] its conformational ensemble is very sensitive to environment and binding partners;[31,32] it forms amyloid aggregates and causes membrane leakage;[22,26] it forms lipid-protein co-aggregates,[23,24] and also undergoes liquid-liquid phase transitions.[33]

The discovery and design of drug-like small molecules capable of ameliorating aggregation by tau and similar IDPs has been a long-standing goal.[4,34] Despite substantial effort, our toolkit of amyloid-modulating compounds had until recently been limited to relatively confined regions of chemical space: a handful of dyes and their analogs, polyphenols like tannic acid and epigallocatechin gallate (EGCG), and other natural products such as curcumin.[35–42] This generation of active compounds has generally been discovered serendipitously or by high-

throughput screening, and has yielded valuable insight into the conformation and aggregation mechanisms of diverse IDPs.

Although IDPs are undeniably challenging targets, there has been exciting recent progress towards the goal of rationally discovering or designing active ligands. Advances in spectroscopic techniques including NMR, single-molecule Förster resonance energy transfer (FRET), time-resolved FRET, tryptophan triplet state quenching, and electron paramagnetic resonance have all provided invaluable insights into IDP conformation and dynamics.[5,31,43–46] Simultaneously, the development of enhanced sampling simulation techniques such as replica exchange molecular dynamics (REMD) and metadynamics, as well as optimized Monte Carlo methods, have generated unprecedented molecular insight into the conformational ensembles sampled by IDPs.[31,47–51] Tau and tau4RD have been modeled using Monte Carlo and molecular dynamics approaches.[52–54] Several groups have been able to leverage this conformational insight into monomeric, oligomeric and bound states of IDPs to generate new, biologically active ligands for a variety of disordered targets.[4,55–60] For example, novel ligands have been discovered or designed for amyloid states of tau and other IDPs,[59,60] for membrane-bound dimers of islet amyloid polypeptide (IAPP or amylin),[57] and to disrupt p53/MDM binding.[61] Broad-spectrum amyloid modulators have been developed based on oligothiophene,[62] oligoquinolone,[63,64] (D,L)-peptide[65,66] and molecular tweezer[67] architectures. Moreover, monomeric disordered states of targets such as c-Myc, amyloid-β and α-synuclein have been successfully targeted by identifying transiently sampled, druggable structures.[55,56,58] In these works, experimentally constrained or filtered simulations were used to define an ensemble of target states, which were then subjected to computational screening followed by experimental validation.

Here, we have been able to improve on these efforts using new enhanced sampling simulation protocols for IDPs, methods to identify preferentially sampled local structures within a larger disordered ensemble, and a combination of conventional computational docking and machine learning (ML)-based screening. Using this approach (schematized in Figure 2.1), examining thousands of compounds computationally and just ten *in vitro*, we have identified one novel compound that inhibits tau aggregation as well as a second that binds to fibrils without altering the extent of aggregation.

Experimental Procedures

Molecular Dynamics Simulations

Molecular dynamics simulations, including ReSA simulations were done using GROMACS 4.6.5[68] with the Amber99SB forcefield. Ten starting states for the MTBR were generated, one using a conformation from FRET-restrained Monte Carlo simulations,[31] and the remaining nine using randomized Ramachandran accessible phi-psi angles and correcting for clashes. Systems were solvated using the tip4p-ew water model, with ten chlorine atoms were added to neutralize charge. Starting structures were energy minimized using a steepest-descent algorithm, then allowed to relax for 0.25 ps with a 0.05 fs timestep before starting ReSA simulations. ReSA was performed using PME electrostatic calculations, NVT ensemble with v-rescale thermostat, and a timestep of 5 fs. Simulations were stopped every 2 ns, then restarted, alternating the reference temperature between 300 and 500K each time.

Clustering and Small Molecule Docking

Simulation data from the 300K simulations was taken and used for clustering analysis. Clusters were determined for all 25 segments of the peptide using all-atom RMSD values. Various clustering algorithms and criteria were examined, and the GROMOS algorithm[69] using

an all-atom RMSD cutoff of 0.2 nm was determined to give the best representative structures based on structure homology and population of the clusters. Median structures from each of the 10 most populated clusters of each segment were converted to receptor models for docking in AutoDock Vina.[70] Structures of the ChemBridge CNS collection were obtained from PubChem and parameters were calculated using OpenBabel.[71] Ligand models were built using AutoDock's ligand preparation utility. CACTVS fingerprints[72] were obtained from PubChem. Ligand docking was done using AutoDock Vina, with a search area encompassing the space of a cube extending 2 nm beyond the receptor model in each dimension.

Protein Expression

Plasmids containing the His-tagged tau4RD or full-length tau2N4R genes with a TEV cleavage site were received as a gift from the Rhoades lab at the University of Pennsylvania. Plasmid was transfected into two *E. coli* lines: DH-5α for long-term stocks, and BL-21 for protein expression. To express protein, BL21 (DE3) cells were incubated in 1 L of Luria-Bertani (LB) medium in baffled flasks. Cells were grown in at 37°C, 220 RPM in a New Brunswick Innova 44/44R incubator until the OD reached 0.6–0.7. IPTG was then added to a final concentration of 0.4 mM, the temperature was dropped to 16°C and the culture was allowed to grow overnight. The following day, cells were harvested by centrifugation and resuspended in lysis buffer: 50 mM Tris, 500 mM NaCl, and 10 mM imidazole, pH 8 @ 4°C. Halt Protease inhibitor cocktail (Life Technologies; 300 µL), 0.1 mg/mL DNAse, 0.1 mg/mL RNAse, and 300 µL of a saturated solution of PMSF in ethanol was added per 1 L of growth before lysis via French press. The lysate was then centrifuged at 9000 *g* for 45 minutes. The supernatant was passed through 0.8 and 0.4 µm filters, before being loaded onto a 5 mL Ni-NTA agarose column. The protein was eluted by increasing the concentration of imidazole in the solvent from 10 to

250 mM. The eluent was buffer exchanged back into 10 μM imidazole buffer, while incubating with 1mM dithiothreitol and 250 μL of TEV protease at 2.8 mg/mL overnight at 4°C. This solution was then run over the Ni-NTA column and the flowthrough was collected and concentrated to 1–2 mL using a 3 kilodalton molecular weight cutoff centrifugal filter. The concentrate was fractionated using a 25 mL S200 extended gel-filtration column with buffer containing 20 mM Tris HCL, 50 mM NaCl, 1 mM TCEP pH 7.4 as the mobile phase. Purity was confirmed by SDS-PAGE. Protein was concentrated, aliquoted and flash frozen before being stored at –80°C. Concentrations of protein were obtained using Pierce BCA Protein Assay Kit (Thermo Fisher Scientific).

Preparation and Storage of Selected Compounds

Selected compounds were identified from the ChemBridge CNS small molecule collection, and dry samples were obtained through Hit2Lead (ChemBridge, San Diego, CA). Samples were dissolved in DMSO to 1 mM stock solutions and stored at -80°C. The identity of active compounds was confirmed with Quantitative time of flight mass spectroscopy using an Agilent Technologies 6520 Q-TOF (Figure 2.10). Dynamic light scattering (DLS) was used to confirm that at experimental conditions, artifacts due to compound insolubility[73] were unlikely to contribute to observed activity (Figure 2.11). DLS assays of compound aggregation were performed using a DynaPro NanoStar Laser Photometer and accompanying software. DMSO control and compound samples were incubated overnight at room temperature or at 37°C, and then immediately assayed.

Aggregation Assays

All fibrillization reactions were carried out at 37°C at pH 7.4, in buffer containing 20 mM Tris HCL, 50 mM NaCl, 1 mM TCEP, and 50 μM thioflavin T (ThT). 100 μL aliquots of tau4RD

and ThT at twice the desired final concentration were prepared in black-walled polystyrene 96-well plates (Corning Life Sciences, Tewksbury, MA). Equal volumes of unfractionated ~3KDa heparin sodium (Acros Organics, Morris Plains, NJ) at twice the desired final concentration were added to the wells to initiate the reactions. Final concentrations of tau4RD and heparin were 5 μM and 3 μM respectively. Reaction progress was monitored by increase in Thioflavin T fluorescence with λ_{ex} = 440 nm and λ_{em} = 485 nm in a BioTek Synergy HTX (BioTek, Winooski, VT) plate reader. Reads were taken every 5 minutes after 1 minute of agitation (for tau4RD), or every 10 minutes with continuous agitation (for full-length tau). Intrinsic fluorescence from compounds was in all cases ≤ 5% of the background, which was corrected for by baseline subtraction. Experiments were performed with at least 3 technical replicates.

Tyrosine Fluorescence Assays

Tau4RD contains a single native tyrosine, which provides an intrinsic spectroscopic probe of concentration. Tyrosine fluorescence assays were performed to measure tau4RD concentration in solution using an Agilent Cary Eclipse spectrophotometer. Samples were excited at 276 nm, and fluorescence was measured from 300–310 nm. A standard curve was constructed using buffer-matched samples of tau4RD, and used to estimate soluble tau levels. Samples were incubated under the same condition as the aggregation assays listed above, save that duplicate samples were made without ThT. Initial concentrations were measured immediately after the addition of heparin. Additionally, after 3 hours of incubation, samples were centrifuged (30 minutes at 21,100 *g*) at 4°C, which was sufficient to separate large aggregates and bona fide fibrils from monomer and small, soluble oligomers. After centrifugation the intrinsic tyrosine fluorescence of tau in the supernatant was measured. Each experiment was run with 4 technical replicates, and the results averaged.

Gel Densitometry

Aggregation reactions were allowed to proceed for 3 hours and spun down as described above. 15 μL of the supernatant was boiled with 3 μL of loading dye and then run on an Invitrogen Bolt 12% Bis-Tris 15 well gel, along with T=0 controls and standard samples of 1, 3 and 5 μM tau4RD. Gels were scanned with a Li-Cor Odyssey CLx gel-scanner, and band intensity was analyzed using Fiji.[74]

Results

Our approach is depicted schematically in Figure 2.1. We began by performing enhanced sampling molecular dynamics on tau4RD to sample its native conformational ensemble. We then used clustering methods to identify preferentially sampled, locally structured states within this ensemble. These preferentially sampled structures were then used in computational screening procedures to identify promising compounds from a library of ~5×10^4 small molecules. ML techniques were used to refine and guide molecular docking calculations. Finally, we assayed ten selected compounds in terms of their effects on tau4RD amyloid formation *in vitro*.

While replica exchange molecular dynamics (REMD) simulations are physically rigorous and are useful for smaller IDPs, they are impractical for an IDP as large as tau4RD. (For example, using explicit solvent, approximately 200 replicas would be required to ensure an exchange probability of 0.01.[75]) To overcome this limitation, we developed and implemented an alternative technique we term Repeated Simulated Annealing (ReSA). In ReSA, a single MD trajectory is periodically (every 2 ns) exchanged from a low temperature (300 K) to a high temperature (500 K) and then returned to 300K. Upon raising the temperature, the target protein rapidly samples conformational space. When the temperature is lowered, the protein dynamics slow down and the simulation samples the local conformational space of the protein (Figure

2.2A). We ran ten 100ns simulations from 10 distinct starting positions, then sampled the 300 K portions of the ReSA trajectory every 50 ps (thereby omitting most structures sampled as tau4RD relaxed from 500 K to 300 K), resulting in 500 ns of usable conformational data. As expected, ReSA samples a broader conformational space in a given time period than conventional MD (Figure 2.2B). Moreover, ReSA more closely recapitulates the radius of gyration (R_g) derived from SM-FRET[31] and small-angle X-ray scattering[76] than conventional MD does (Figure 2.3D). NMR chemical shifts calculated from the ReSA ensemble are also in good agreement with experimental measurements (Figure 2.2). Both the predicted chemical shifts and the contact map (Figure 2.4A) are very similar to those obtained by extensive (4.8 μs) REMD simulations.[54]

After generating a library of conformations using ReSA, we used a local clustering approach to identify potential docking targets. One of the challenges in identifying targets from an IDP conformational ensemble is determining which conformations are the most frequently sampled in solution. More popular conformations are desirable as targets: the more frequently sampled a conformation is, the greater the chance it encounters a ligand. However, identifying well-sampled conformations for dynamic protein targets is not straightforward. Conformational fluctuations in distant, unrelated parts of the protein can cause the clustering algorithm to classify two conformations as distinct and relatively uninteresting even if they both displayed the same local structural motif. Because our conformational library displayed few long-range contacts, we expect local structures to be relevant targets (Figure 2.3). Attempts to generate clusters based on global conformational similarity (e.g., Cα RMSD for the entire tau4RD sequence) would overlook *bona fide*, preferentially sampled local structures. Instead, we performed clustering on overlapping segments of tau4RD extracted from the larger ReSA ensemble. We tested segment lengths of 5, 10 and 20 residues, and found that ten-residue segments (Figure 2.5) provided the

best compromise between cluster resolution and cluster population. This approach allowed us to model the conformations of tau4RD in the context of the complete molecule, yet also identify relevant local structures. After experimentation with different clustering methods and all-atom RMSD cutoff values, we clustered our structures using the GROMOS geometric clustering method[69] using a RMSD cutoff of 0.2 nm. This allowed for the identification of well-resolved preferentially sampled structures for each of the 10-residue segments (Figure 2.6).

We used these structures as docking targets for virtual screening of small molecules. Each of the 10 most populated clusters for each of the 25 segments was used as a target for computational docking. Each cluster was representative of 0.5% to 4% of the conformational library. A pilot set of 1000 compounds of the ChemBridge CNS collection was docked against the 250 targets using AutoDock Vina. These results showed that compounds preferentially bound to particular regions of the tau4RD (Figure 2.7A), suggesting that these segments are essentially the most "druggable" portions of the protein. We also found that similar docking results are obtained when targets were generated from either half of the ReSA 300K dataset (Figure 2.6C), which suggests that the ReSA conformational ensemble has converged sufficiently for our purpose of ligand discovery.

It would be computationally prohibitive to dock larger ($> 10^4$ compound) libraries to all 250 targets. Therefore, we used ML to mine large compound collections and prioritize small molecules for more detailed computational screening. ML techniques can robustly and efficiently predict diverse aspects of chemical function from simplified descriptions of a compound's structure.[57,77–80] We used CACTVS chemical fingerprints[81] to quantify the presence of chemical features of the compounds in the CNS collection, then used partial least squares regression (PLSR) analysis to correlate the presence and absence of chemical features with docking scores.

A CACTVS fingerprint is an 882-bit binary string, with a 0 or 1 at each position corresponding to the absence or presence of a particular chemical feature in a compound of interest. PLSR is a versatile technique capable of identifying the relationships between two matrices, one of which is dependent in some way on the other.[77] PLSR simultaneously decomposes those matrices to identify patterns of covariance to build a regression matrix that, when multiplied by the independent matrix, approximates the dependent matrix. In our case, the independent matrix F consists of the CACTVS fingerprints for each compound in a training set, while the dependent matrix S consists of docking scores for each compound in that training set to each target in a panel of ReSA-derived clusters. PLSR yields a prediction matrix P such that $F \cdot P \approx S$. If a vector f consists of the fingerprint of a new compound not in the training set, then the vector $s = f \cdot P$ will contain the predicted docking score of that compound for each target in the panel. PLSR thus approximates the results of multiple, computationally intensive docking calculations with a single matrix multiplication step.

We validated our PLSR predictions using a subsampling approach. Briefly, 10 randomly selected molecules were withheld from the training set, and the model was trained with the remaining 990 (Figure 2.7 B). We then compared the predicted scores of the 10 test compounds with docking scores generated by Autodock Vina. We repeated this subsampling ten times, each with a different set of 10 test compounds. The Pearson correlation coefficient between PLSR predictions and docking scores were 0.60 ± 0.29 as compared to 0.64 for the training set, while the Spearman rank-order correlation coefficient was 0.58 ± 0.20 for the test sets vs. 0.67 for the training set. This indicated that PLSR was able to predict the docking scores of unknown compounds with reasonable accuracy.

We then applied our validated PLSR protocol to predict docking scores for all 50,000 compounds in the ChemBridge CNS library, and selected the top 1000 molecules for another round of docking. For computational efficiency, all subsequent docking calculations were performed only with the segments of tau4RD identified to be druggable (Figure 2.7A). The docking results of these 1000 compounds were used to update the PLSR training set, and the PLSR predictions were repeated. The 1000 compounds predicted to have the best docking scores were again selected for docking, and this process was repeated, at which point the proportion of new compounds (i.e., not identified in previous rounds) dropped to ~40% and the docking scores of the best compounds stayed approximately constant. At this stage, we considered the ML screen to have converged, and ranked all compounds in terms of their docking score for any segment of tau4RD (Figure 2.7C). High-scoring compounds tended to be hydrophobic (mean clogP of the top 1% of compounds was 3.8, as compared to 3.1 for the entire compound library), and to feature heteroatomic conjugated systems. Selected docked poses are shown in Figure 2.8.

We selected 10 compounds (see Table 2.1) for experimental testing based on docking score, targeted segment, and chemical structure, so as to prioritize high-affinity ligands that bound different regions of tau4RD and sampled diverse regions of chemical space. These 10 compounds were tested for their effects on the pathological aggregation of tau4RD using a kinetic assay that relies on the enhancement of ThT fluorescence when bound to amyloid structure. We compared the aggregation of tau4RD in the presence and absence of each compound and found that 1 and 6 markedly altered the observed trajectory (Figure 2.9). Compound 1 delayed and decreased the ThT fluorescence in a dose-dependent manner, and compound 6 increased the magnitude of ThT fluorescence in a dose dependent manner (In contrast, 0 of 10 compounds randomly selected from the ChemBridge CNS library showed any

activity). To corroborate the results of the ThT assay, we centrifuged aggregation reactions after 3 hours and measured the amount of soluble tau4RD remaining in the supernatant using intrinsic tyrosine fluorescence. Compound 1 consistently decreased the amount of aggregated protein (confirmed by gel densitometry, Figure 2.12), while 6 had no effect. These results suggest that 1 is a *bona fide* aggregation inhibitor, while 6's effects on the ThT fluorescence are likely photophysical in origin. Since this signal results from fibril-bound ThT, the enhancement of fluorescence suggests that 6 does interact with tau4RD fibrils without affecting the rate or extent of aggregation. In contrast, 1 is a fairly potent aggregation inhibitor, capable of delaying aggregation substoichiometrically. However, despite this effect on kinetics, neither 1 nor 6 appears to affect the total amount of aggregate formed once the reaction has plateaued (see 96 h data in Figure 2.13).

The tau4RD construct has been extensively studied *in vitro*, in part because it forms fibrils much more readily than full-length tau. In order to determine whether 1 and 6 retain their activity towards biologically relevant forms of tau, we studied the aggregation of full-length (2N4R) tau in the presence and absence of each compound. The response of full-length tau to these modulators mirrored that of tau4RD, albeit on much slower timescales: 1 delayed the onset of aggregation, while 6 increased ThT fluorescence intensity without seeming to affect the kinetics of aggregation in any systematic way. This suggests that 1 and 6 could be useful reagents in cellular and animal models of tau pathology, whether studying full-length tau or more experimentally tractable constructs.

Importantly, both 1 and 6 are chemically quite distinct from any other compound known to affect amyloid formation by tau or any other protein. This suggests that our approach has indeed enabled us to explore new regions of chemical space for tau aggregation inhibitors.

Discussion

The role of computation in ligand discovery is to reduce the amount of *in vitro* testing needed to identify active compounds. High throughput screening for IDP ligands is often challenging because binding or other functional outcomes can be difficult to detect. Rates of success for these undirected screening are as low as 0.04%.[82] As such, computational approaches are appealing in that much of the selection process can be shifted to the relatively cheap and quick *in silico* methods. Recent attempts have aimed to identify targets (often using computational methods guided by *in vitro* measurements), and then use computational docking approaches to identify a limited number of promising compounds. For example, Yu *et al*. targeted the IDP c-Myc based on models built from molecular dynamics biased with a known inhibitor. Their computational screening identified 273 compounds from which they identified 7 novel ligands.[58] Vendruscolo *et al*. used clustering methods to identify the most populated global conformations in REMD simulations of α-synuclein, and then used fragment mapping to identify binding hotspots. From computational docking to these hotspots, the authors identified 89 compounds for *in vitro* characterization, and reported one biologically active compound capable of rescuing dopaminergic cells in a Parkinson's cell model.[56]

The process we described here distinguishes itself in three major ways. Firstly, appropriate and efficient enhanced sampling molecular dynamics methods explored conformational space and generated a valuable conformational library. Secondly, the segmented clustering approach allowed the identification of well-populated local conformations that serve as appealing ligand targets. Thirdly, machine learning techniques identified and utilized chemical characteristics that focused screening attempts and drastically increased the efficiency of computational screening. Our approach identified two novel tau ligands with promising

activity – 1 as an aggregation inhibitor, and 6 as a fibril-binding ligand – out of just ten compounds characterized *in vitro*. This suggests that our methodology successfully identified promising targets within a dynamic conformational ensemble and used this information to select compounds likely to interact with an IDP target. Future studies will examine the biological activities of 1, 6 and related compounds in cellular and *in vivo* assays of tau pathology. Importantly, while the compounds identified here are novel and interesting, they will require substantial further development to become lead compounds in drug development efforts. Bioavailability and other pharmacokinetic/pharmacodynamic parameters will need to be optimized, and so will selectivity for tau over other amyloid-forming proteins. Nevertheless, we believe that the development of 1 and 6 represents an exciting advance in the collective effort to improve diagnosis and treatment of AD and other tauopathies. We anticipate that screening larger compound collections using the methods described here will yield multiple, distinct modulators and probes of tau aggregation.

The work presented here provides a framework that can easily be applied to other disordered protein targets. ReSA is easily adaptable to work on other systems, and does not have extensive computational requirements beyond those of canonical molecular dynamics. We stress that ReSA is not likely to generate a complete, physically rigorous landscape of an IDP conformational ensemble, but nevertheless is evidently able to model an IDP at sufficient resolution and accuracy for rational ligand discovery. The segmentation and clustering approach is also well suited for other larger IDPs, although the parameters for clustering will need to be adjusted to accommodate for the flexibility and local structural diversity of the system being examined. Our hope is that this work provides valuable tools to the research community that aid in the development of ligands capable of modulating IDP function or dysfunction, so as to gain

insight into the mechanisms of toxicity, and to advance the eventual development of therapeutics and diagnostics for IDP-mediated pathologies.

Figures

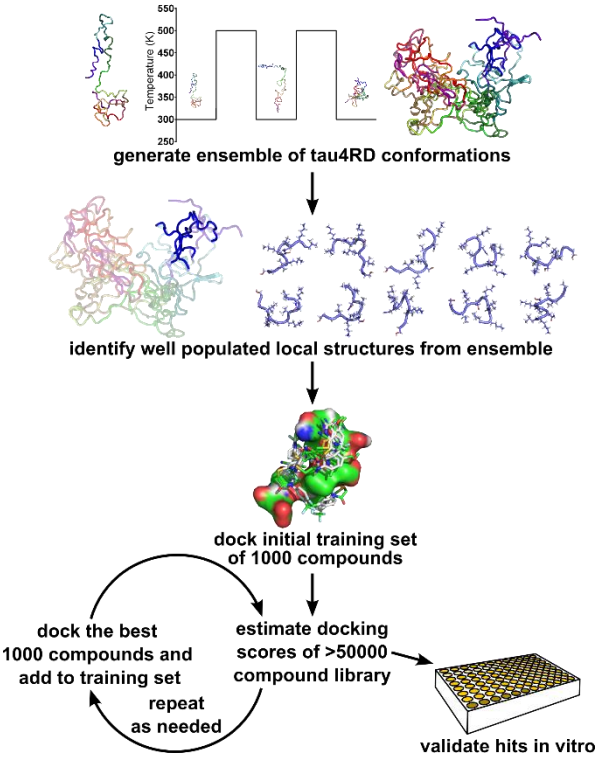


Figure 2.1: Schematic of our combined computational and experimental strategy to discover novel ligands for IDPs.
Initially a conformational library is generated with repeated simulated annealing, then well populated local structures are established from this ensemble using computational methods. These structures are used as targets for molecular docking simulations. Molecular docking and machine learning are iteratively used to determine docking scores of a subset of compounds, and then estimate scores of an entire chemical library. Finally, selected compounds are tested with *in vitro* methods.

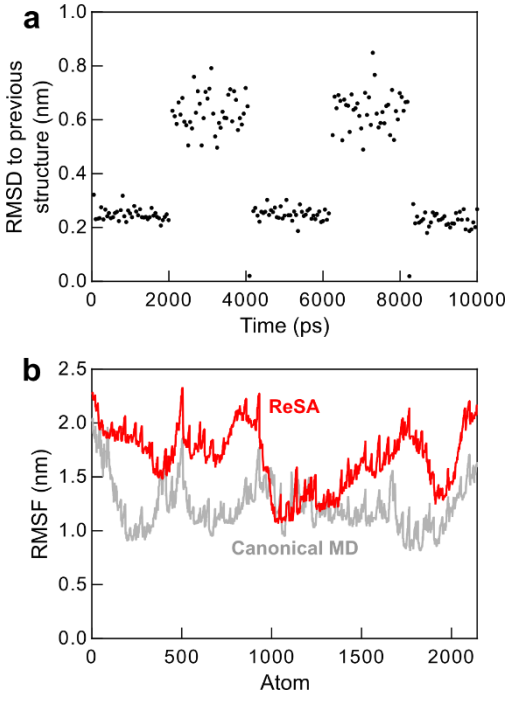


Figure 2.2: ReSA explores relevant conformational space more efficiently than canonical molecular dynamics.
a) In ReSA, rates of change are increased during periods of high temperature, as indicated by higher RMSD values when comparing to structures 50ps prior in the simulation.
b) Root-mean-squared fluctuation (RMSF) values of atoms in ReSA simulation are greater than or equal to the RMSF of those atoms in canonical MD of the same length generated from the same starting structure, indicating that in ReSA the protein is more dynamic.

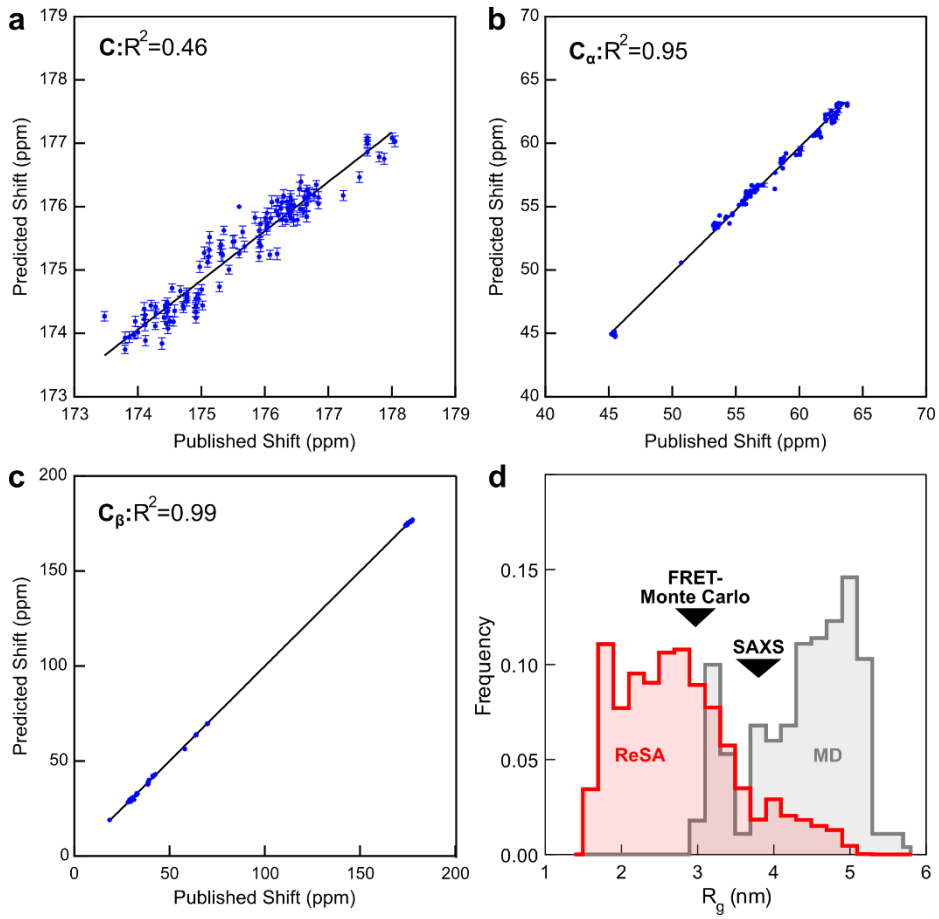


Figure 2.3: Experimental validation of the ReSA-derived conformational ensemble.
^{13}C chemical shifts of carbonyl C (a), Cα (b), and Cβ (c) predicted using SPARTA+ [83] from the ReSA ensemble align well with published values [84].
d) The conformational ensemble generated by ReSA (500 ns) more closely recapitulates radius of gyration (R_g) values derived from FRET-constrained Monte Carlo simulations and small-angle X-ray scattering[76] than that generated by canonical molecular dynamics (50 ns).

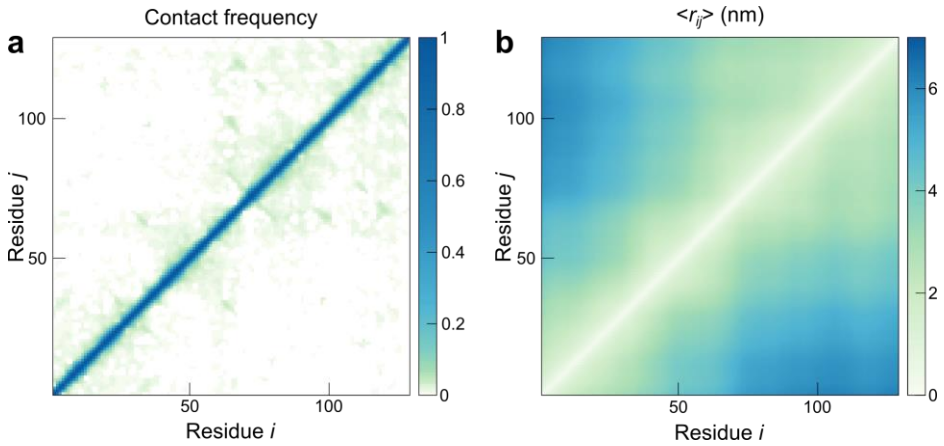


Figure 2.4: Inter-residue contacts in the ReSA-derived conformational ensemble
a) Frequency of ≤ 0.6 nm contacts between residues and (b) average inter-residue distances (nm) for the entire ReSA conformation ensemble show that while long-range contacts do occur, they are rare relative to contacts within 10 residues.

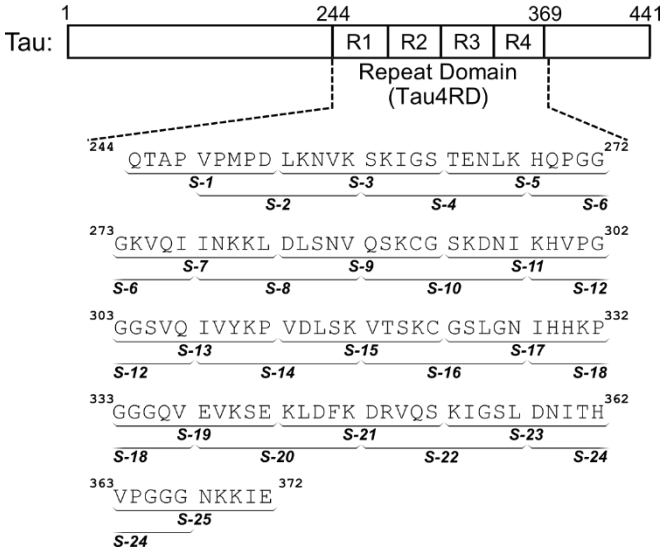


Figure 2.5: Sequence of tau4RD and segments used for local clustering. Tau4RD spans the 4-repeat microtubule-binding domain of the IDP tau, and models the full length protein's capacity for aggregation and membrane interaction. To identify well-populated local structures within the conformational ensemble of tau4RD, the sequence was divided into 25 overlapping segments (denoted S-1 through S-25) and each segment was clustered independently.

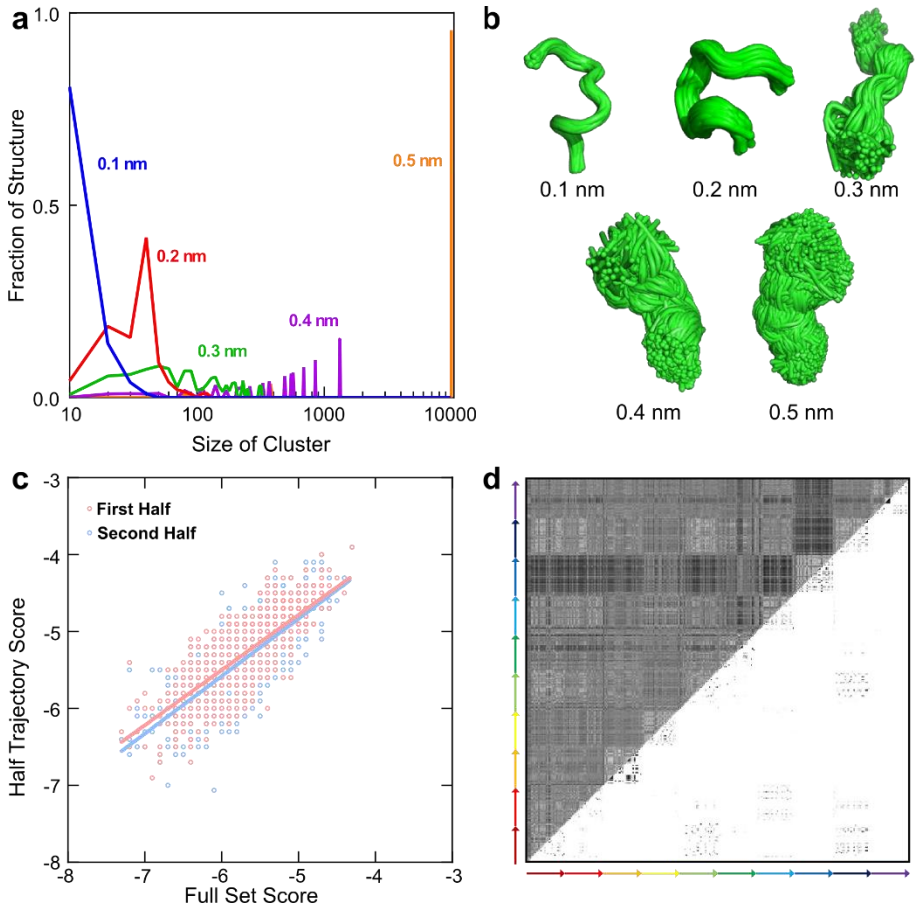


Figure 2.6: Cluster analysis and demonstrations of convergence

a) Cluster population distributions vary greatly with the RMSD cutoff. Clusters generated with overly large RMSD cutoffs are too large and poorly defined. Cutoffs that are too small do not generate well-populated clusters.

b) Representative clusters with RMSD cutoffs between 0.1 and 0.5 nm.

c) Docking targets generated from either the first or the second half of the conformational ensemble of tau4RD yielded similar docking as targets generated from the entire set of conformations, further suggesting that the ReSA simulations have converged for our purposes. Pearson correlation coefficients are 0.78 and 0.76 for the first and second half targets, respectively.

d) A RMSD matrix of Segment 1 compares each structure of all 10 independent ReSA simulations (denoted with different colored arrows) against every other structure. The top half shows structural similarity with darker pixels signifying higher RMSD values. On the bottom half, black pixels represent when two structures fall into the same cluster based on an all-atom RMSD criteria of 0.2 nm. The off-diagonal points on this half show that many clusters are sampled in non-contiguous parts of a given simulation as well as in multiple independent simulations.

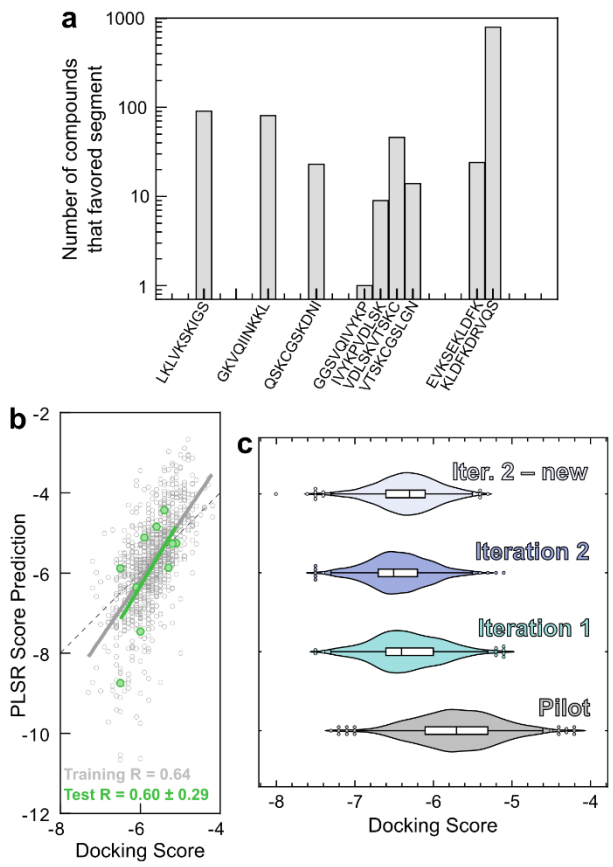
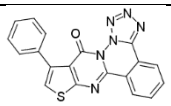
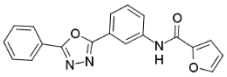
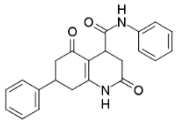
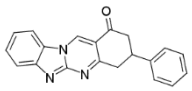
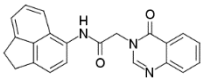
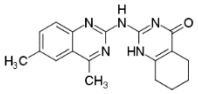
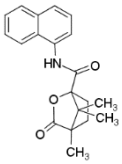
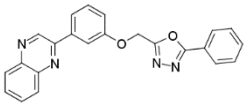
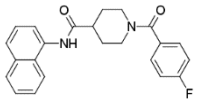
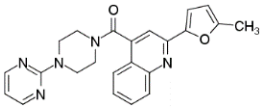


Figure 2.7: Discovery of potentially active compounds through iterative machine learning. a) A pilot set of 1000 compounds randomly chosen from the CNS library was docked against the 10 most populated clusters from each of the 25 segments. The histogram of each compound's most favored target illustrates that some segments of tau4RD are better able to accommodate small molecule binding. b) PLSR trained on 990 compounds from the pilot set was able to predict the docking scores of the remaining 10 compounds with reasonable accuracy. Pearson correlation coefficients are shown at the bottom of the panel. c) PLSR was used to select compounds from the diverse set of small molecules in the CNS library. Iteration 1 yielded 1000 compounds based on the results and fingerprints of the pilot set. Iteration 2, based on the docking scores and fingerprints of the 2000 compounds from the pilot set and Iteration 1, yielded primarily compounds that had already been examined and docked in Iteration 1. When these were excluded from the library, the next 1000 compounds ("Iter. 2 – new") displayed worse docking scores on average. This indicated that the PLSR search had largely converged.

Table 2.1: Compounds selected for in vitro screening, based on a combination of docking score, chemical diversity and sequence coverage.

Compound #	Chembridge ID	PubChem CID	Best Score	Best Target (Segment-Cluster)	Structure
1	7579671	1184213	-8	3-1	
2	6399586	659810	-7.5	21-6	
3	7999316	2988410	-7.4	21-6	
4	7987291	2983350	-7.3	3-1	
5	7358814	1077875	-7.3	21-7	
6	7282131	707162	-6.9	16-9	
7	6952140	3151892	-6.9	19-9	
8	9011432	2994718	-6.9	14-2	
9	6983997	1072807	-6.9	14-2	
10	7233320	1001781	-6.7	10-7	

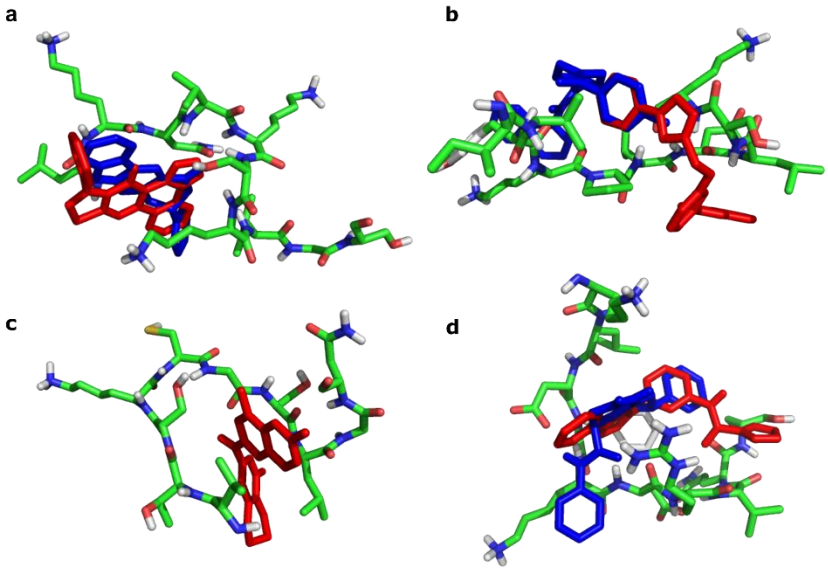


Figure 2.8: Examples of compound docking positions on favored targets.
a) Compound 1 (Red) and Compound 4 (Blue) docked against Segment 3, Cluster 1
b) Compound 8 (Red) and Compound 9 (Blue) docked against Segment 14, Cluster 2
c) Compound 6 (Red) docked against Segment 16, Cluster 9
d) Compound 2 (Red) and Compound 3 (Blue) bound to Segment 21, Cluster 6

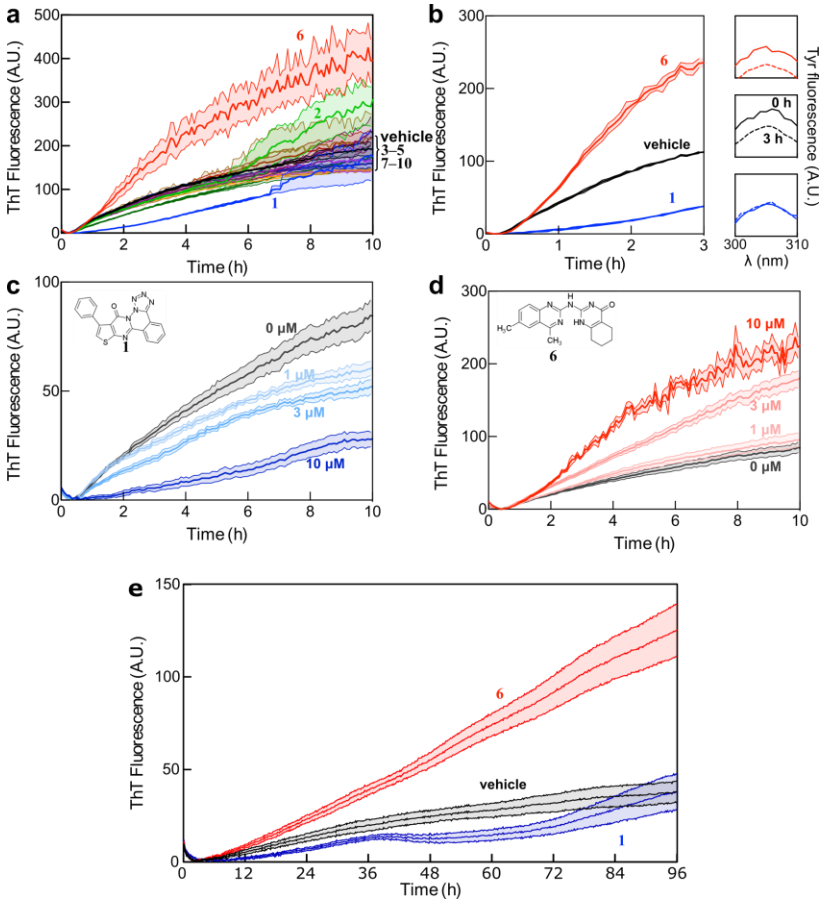


Figure 2.9: Experimental validation of selected compounds. a) ThT-monitored tau4RD fibril formation kinetics in the presence of 10 compounds selected based on computational screening. Compounds 1 and 6 consistently altered the observed fluorescence kinetics relative to vehicle control. Traces show mean ± SEM of 3 technical replicates, and are representative of three independent experiments. b) Measurement of soluble tau content after 3 hours of aggregation indicates that 1 but not 6 affects the extent of aggregation. The main panel shows ThT-monitored aggregation, while the panels on the right show the amount of tau remaining in solution after centrifugation, measured by intrinsic tyrosine fluorescence. Treatment with Compound 1 (blue) shows that soluble tau4RD content is unchanged over the course of the experiment. In contrast, treatments with either Compound 6 (red) or vehicle (black) show a similar loss of fluorescence intensity. c) Compound 1 extends the lag phase and decreases ThT fluorescence intensity in a dose-dependent manner. d) Compound 6 increases ThT fluorescence in a dose-dependent manner, but does not dramatically affect the lag phase. For B and C, traces show mean ± SEM of 4 technical replicates, and are representative of three independent experimentse) Compounds 1 and 6 exhibit similar effects on ThT-monitored aggregation of full-length tau, with 1 delaying aggregation and 6 appearing to raise the final level of fluorescence. For D and E, traces show mean ± SEM of 3 technical replicates, and are representative of at least three independent experiments. Concentration of compounds was set at 10 µM unless otherwise indicated and the concentration of DMSO for all samples was 1%.

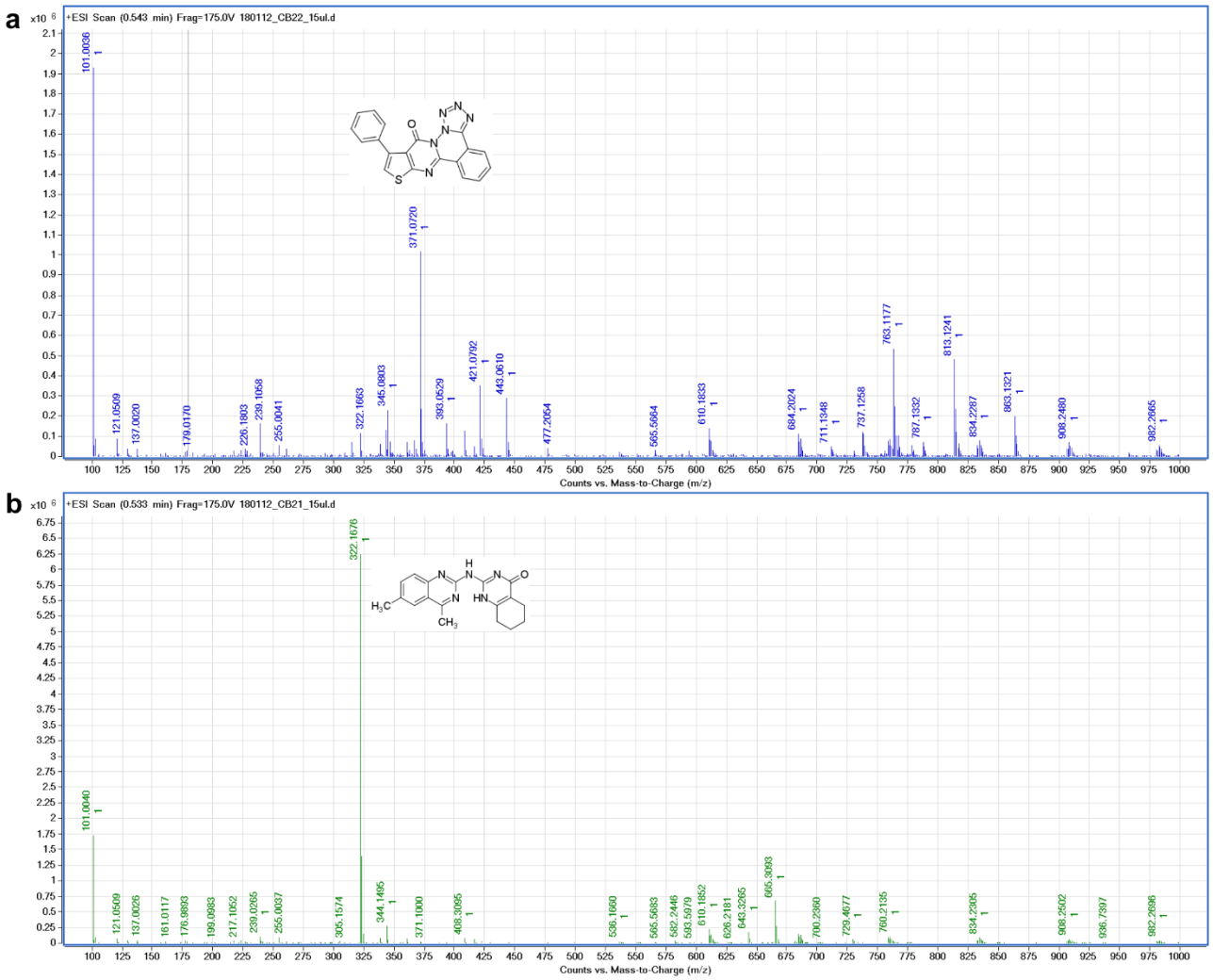


Figure 2.10: Mass Spectrometry
Mass spectrometry confirms the molecular weights of Compound 1 (a) and Compound 6 (b).

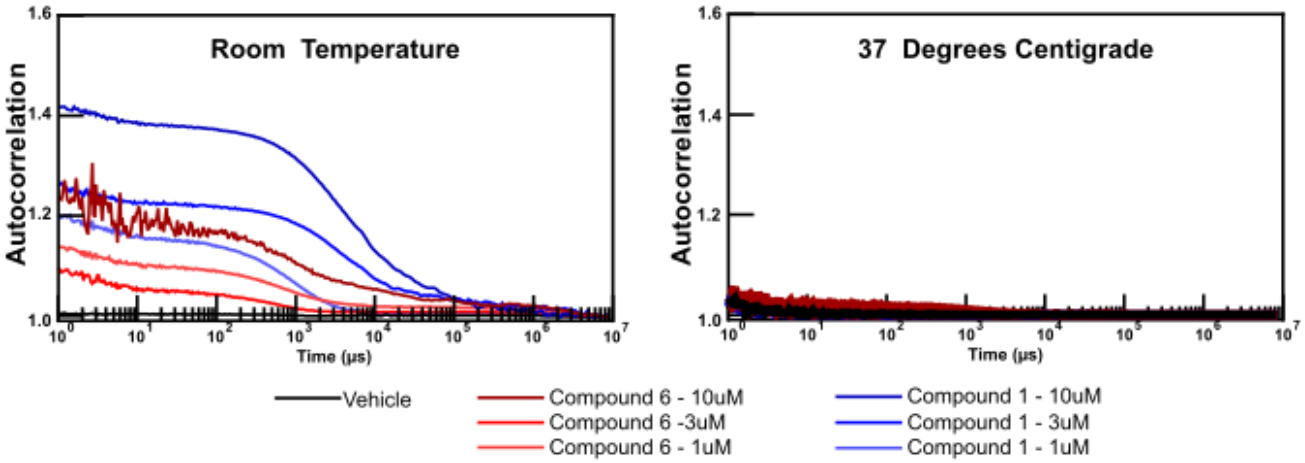


Figure 2.11: Dynamic Light Scattering
a) DLS shows that at room temperature, particles are present at micromolar concentrations of both compound 1 and compound 6. After incubation at 37 degrees C, these particles dissipate.

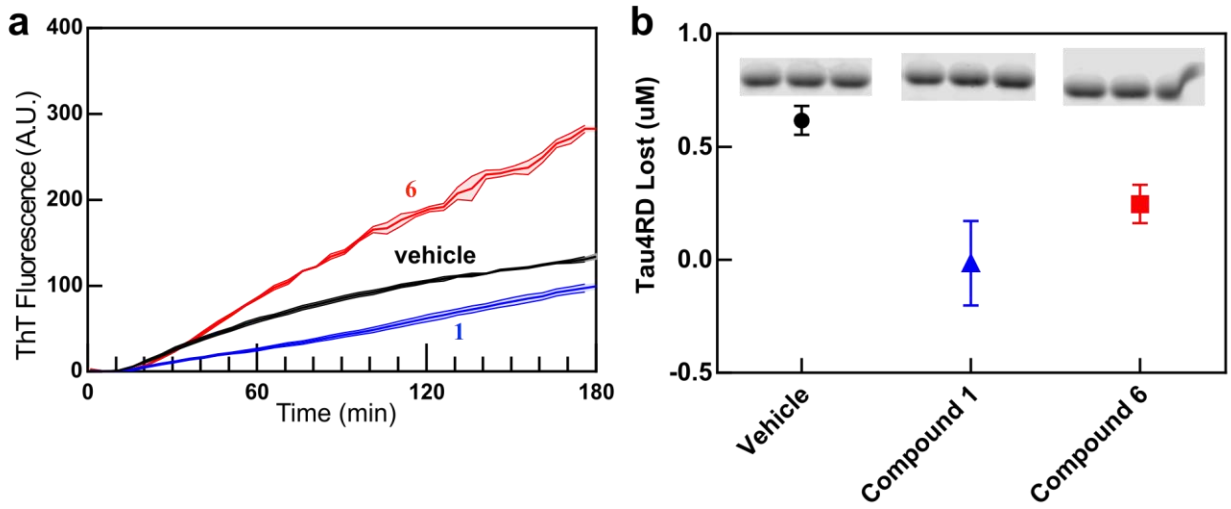
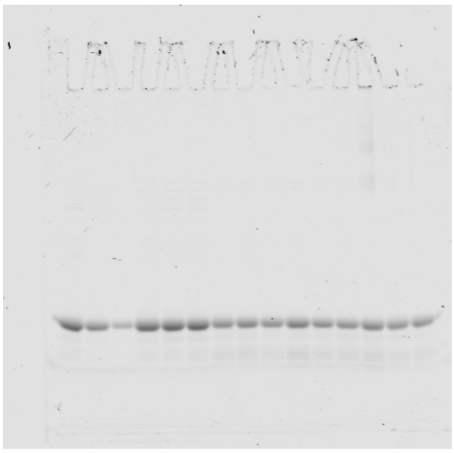


Figure 2.12: Gel Densitometry
After 3 hours of aggregation monitored by ThT fluorescence (a), densitometry (b) confirms the aggregation-delaying effect of Compound 1, and suggests that compound 6 does not strongly affect the aggregation of Tau4RD, as does Tyr-fluorescence.



Lane	Sample
1	5uM Tau4RD
2	3 uM Tau4RD
3	1 uM Tau4RD
4	Vehicle T0
5	Compound 6 T0
6	Compound 1 T0
7	Vehicle Supernatant 1
8	Vehicle Supernatant 2
9	Vehicle Supernatant 3
10	Compound 6 Supernatant 1
11	Compound 6 Supernatant 2
12	Compound 6 Supernatant 3
13	Compound 1 Supernatant 1
14	Compound 1 Supernatant 2
15	Compound 1 Supernatant 3

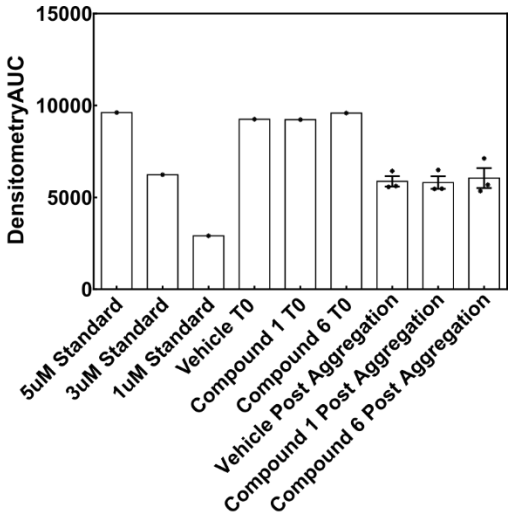


Figure 2.13: Tau4RD 96 Hour Gel Densitometry
After 96 hours of aggregation, densitometry results suggest that there is not a significant difference in the amount of soluble tau remaining after aggregation.

References

1. Baggett, D. W. & Nath, A. The Rational Discovery of a Tau Aggregation Inhibitor. *Biochemistry* 57, 6099–6107 (2018).
2. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annu. Rev. Biophys.* 37, 215–246 (2008).
3. Cheng, Y. *et al.* Rational drug design via intrinsically disordered protein. *Trends Biotechnol.* 24, 435–42 (2006).
4. Ambadipudi, S. & Zweckstetter, M. Targeting intrinsically disordered proteins in rational drug discovery. *Expert Opin. Drug Discov.* 11, 65–77 (2016).
5. Nath, A. & Rhoades, E. A flash in the pan: Dissecting dynamic amyloid intermediates using fluorescence. *FEBS Lett.* 587, 1096–1105 (2013).
6. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 272, 5129–48 (2005).
7. Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol* 98, 85–106 (2008).
8. Uversky, V. N. p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. *Int. J. Mol. Sci.* 17, (2016).
9. Mark, W.-Y. *et al.* Characterization of Segments from the Central Region of BRCA1: An Intrinsically Disordered Scaffold for Multiple Protein–Protein and Protein–DNA Interactions? *J. Mol. Biol.* 345, 275–287 (2005).
10. Weber, S. C. & Brangwynne, C. P. Getting RNA and protein in phase. *Cell* 149, 1188–1191 (2012).
11. Lemke, E. A. The Multiple Faces of Disordered Nucleoporins. *J. Mol. Biol.* 428, 2011–2024 (2016).
12. Melo, A. M. *et al.* A functional role for intrinsic disorder in the tau-tubulin complex. *Proc. Natl. Acad. Sci. U. S. A.* 113, 14336–14341 (2016).
13. Stroberg, W. & Schnell, S. On the origin of non-membrane-bound organelles, and their physiological function. *J. Theor. Biol.* 434, 42–49 (2017).
14. Mackenzie, I. R., Rademakers, R. & Neumann, M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *Lancet Neurol.* 9, 995–1007 (2010).
15. Metskas, L. A. & Rhoades, E. Order-Disorder Transitions in the Cardiac Troponin Complex. *J. Mol. Biol.* 428, 2965–77 (2016).
16. Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. *Cell* 148, 1188–1203 (2012).
17. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* 15, 384–96 (2014).
18. Drubin, D. G. & Kirschner, M. W. Tau protein function in living cells. *J. Cell Biol.* 103, 2739–46 (1986).
19. Kadavath, H. *et al.* Tau stabilizes microtubules by binding at the interface between tubulin heterodimers. *Proc. Natl. Acad. Sci. U. S. A.* 112, (2015).

20. Santarella, R. A. *et al.* Surface-decoration of microtubules by human tau. *J. Mol. Biol.* 339, 539–53 (2004).
21. Lee, G. & Leugers, C. J. Tau and tauopathies. *Prog. Mol. Biol. Transl. Sci.* 107, 263–93 (2012).
22. Flach, K. *et al.* Tau oligomers impair artificial membrane integrity and cellular viability. *J. Biol. Chem.* 287, 43223–33 (2012).
23. Elbaum-Garfinkle, S., Ramlall, T. & Rhoades, E. The role of the lipid bilayer in tau aggregation. *Biophys. J.* 98, 2722–2730 (2010).
24. Ait-Bouziad, N. *et al.* Discovery and characterization of stable and toxic Tau/phospholipid oligomeric complexes. *Nat. Commun.* 8, 1678 (2017).
25. Barghorn, S., Biernat, J. & Mandelkow, E. Purification of recombinant tau protein and preparation of Alzheimer-paired helical filaments in vitro. *Methods Mol. Biol.* 299, 35–51 (2005).
26. Künze, G. *et al.* Binding of the three-repeat domain of tau to phospholipid membranes induces an aggregated-like state of the protein. *Biochim. Biophys. Acta - Biomembr.* 1818, 2302–2313 (2012).
27. Ganguly, P. *et al.* Tau assembly: The dominant role of PHF6 (VQIVYK) in microtubule binding region repeat R3. *J. Phys. Chem. B* 119, 4582–4593 (2015).
28. von Bergen, M. *et al.* Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif (306VQIVYK311) forming beta structure. *Proc. Natl. Acad. Sci.* 97, 5129–5134 (2000).
29. Fitzpatrick, A. W. P. *et al.* Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* 547, 185–190 (2017).
30. Baughman, H. E. R., Clouser, A. F., Klevit, R. E. & Nath, A. HspB1 and Hsc70 chaperones engage distinct tau species and have different inhibitory effects on amyloid formation. *J. Biol. Chem.* 293, 2687–2700 (2018).
31. Nath, A. *et al.* The conformational ensembles of α-synuclein and tau: combining single-molecule FRET and simulations. *Biophys. J.* 103, 1940–9 (2012).
32. Elbaum-Garfinkle, S. & Rhoades, E. Identification of an aggregation-prone structure of tau. *J. Am. Chem. Soc.* 134, 16607–16613 (2012).
33. Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E. & Zweckstetter, M. Liquid–liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau. *Nat. Commun.* 8, 275 (2017).
34. Campos, H. C. *et al.* The role of natural products in the discovery of new drug candidates for the treatment of neurodegenerative disorders I: Parkinson's disease. *CNS Neurol. Disord. Drug Targets* 10, 239–50 (2011).
35. Masuda, M. *et al.* Small molecule inhibitors of alpha-synuclein filament assembly. *Biochemistry* 45, 6085–94 (2006).
36. Lakey-Beitia, J., Berrocal, R., Rao, K. S. & Durant, A. A. Polyphenols as therapeutic molecules in Alzheimer's disease through modulating amyloid pathways. *Mol. Neurobiol.*

51, 466–479 (2015).
37. Alavez, S., Vantipalli, M. C., Zucker, D. J. S., Klang, I. M. & Lithgow, G. J. Amyloid-binding compounds maintain protein homeostasis during ageing and extend lifespan. *Nature* 472, 226–9 (2011).
38. Ono, A. E., Oyekigho, E. W. & Adeleke, O. A. Isolated systolic hypertension: Primary care practice patterns in a Nigerian high-risk subpopulation. *Sao Paulo Medical Journal* 124, 105–109 (2006).
39. Hauber, I., Hohenberg, H., Holstermann, B., Hunstein, W. & Hauber, J. The main green tea polyphenol epigallocatechin-3-gallate counteracts semen-mediated enhancement of HIV infection. *Proc Natl Acad Sci U S A* 106, 9033–9038 (2009).
40. Akoury, E. *et al.* Inhibition of Tau Filament Formation by Conformational Modulation. *J. Am. Chem. Soc.* 135, 2853–2862 (2013).
41. Ballatore, C. *et al.* Discovery of Brain-Penetrant, Orally Bioavailable Aminothienopyridazine Inhibitors of Tau Aggregation. *J. Med. Chem.* 53, 3739–3747 (2010).
42. Akoury, E. *et al.* Mechanistic Basis of Phenothiazine-Driven Inhibition of Tau Aggregation. *Angew. Chemie Int. Ed.* 52, 3511–3515 (2013).
43. Salmon, L. *et al.* NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132, 8407–18 (2010).
44. Grupi, A. & Haas, E. Time resolved FRET detection of subtle temperature induced conformational biases in ensembles of α-synuclein molecules. *J. Mol. Biol.* 411, 234–47 (2011).
45. Drescher, M. EPR in Protein Science. in *Topics in current chemistry* 321, 91–119 (2011).
46. Soranno, A., Longhi, R., Bellini, T. & Buscaglia, M. Kinetics of contact formation and end-to-end distance distributions of swollen disordered peptides. *Biophys. J.* 96, 1515–28 (2009).
47. Ferrie, J. J. *et al.* Using a FRET Library with Multiple Probe Pairs To Drive Monte Carlo Simulations of α-Synuclein. *Biophys. J.* 114, 53–64 (2018).
48. Scheraga, H. A., Khalili, M. & Liwo, A. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* 58, 57–83 (2007).
49. Leone, V., Marinelli, F., Carloni, P. & Parrinello, M. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* 20, 148–54 (2010).
50. Vitalis, A. & Pappu, R. V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* 30, 673–99 (2009).
51. Vitalis, A. & Pappu, R. V. Methods for Monte Carlo simulations of biomacromolecules. *Annu. Rep. Comput. Chem.* 5, 49–76 (2009).
52. Fisher, C. K., Huang, A. & Stultz, C. M. Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.* 132, 14919–14927 (2010).
53. Ozenne, V. *et al.* Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J. Am. Chem. Soc.* 134, 15138–15148 (2012).

54. Luo, Y., Nussinov, R. & Wei, G. Structural Insight into Tau Protein ' s Paradox of Intrinsically Disordered Behavior, Self-Acetylation Activity, and Aggregation. (2014).
55. Zhu, M. *et al.* Identification of small-molecule binding pockets in the soluble monomeric form of the Aβ42 peptide. *J. Chem. Phys.* 139, (2013).
56. Tóth, G. *et al.* Targeting the intrinsically disordered structural ensemble of α-synuclein by small molecules as a potential therapeutic strategy for Parkinson's disease. *PLoS One* 9, e87133 (2014).
57. Nath, A., Schlamadinger, D. E., Rhoades, E. & Miranker, A. D. Structure-Based Small Molecule Modulation of a Pre-Amyloid State: Pharmacological Enhancement of IAPP Membrane-Binding and Toxicity. *Biochemistry* 54, (2015).
58. Yu, C. *et al.* Structure-based Inhibitor Design for the Intrinsically Disordered Protein c-Myc. *Sci. Rep.* 6, 22298 (2016).
59. Sievers, S. A. *et al.* Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* 475, 96–100 (2011).
60. Seidler, P. M. *et al.* Structure-based inhibitors of tau aggregation. *Nat. Chem.* 10, 170–176 (2017).
61. Yu, X., Narayanan, S., Vazquez, A. & Carpizo, D. R. Small molecule compounds targeting the p53 pathway: are we finally making progress? (2014). doi:10.1007/s10495-014-0990-3
62. Civitelli, L. *et al.* The Luminescent Oligothiophene p-FTAA Converts Toxic $A\beta_{1-42}$ Species into Nontoxic Amyloid Fibers with Altered Properties. *J. Biol. Chem.* 291, 9233–9243 (2016).
63. Kumar, S. & Miranker, A. D. A foldamer approach to targeting membrane bound helical states of islet amyloid polypeptide. *Chem. Commun. (Camb).* 49, 4749–51 (2013).
64. Kumar, S., Brown, M. A., Nath, A. & Miranker, A. D. Folded small molecule manipulation of islet amyloid polypeptide. *Chem. Biol.* 21, 775–781 (2014).
65. Kellock, J., Hopping, G., Caughey, B. & Daggett, V. Peptides Composed of Alternating L- and D-Amino Acids Inhibit Amyloidogenesis in Three Distinct Amyloid Systems Independent of Sequence. *J. Mol. Biol.* 428, 2317–2328 (2016).
66. Bleem, A., Francisco, R., Bryers, J. D. & Daggett, V. Designed α -sheet peptides suppress amyloid formation in Staphylococcus aureus biofilms. *NPJ Biofilms Microbiomes* 1–9 (2017). doi:10.1038/s41522-017-0025-2
67. Sinha, S. *et al.* Lysine-Specific Molecular Tweezers Are Broad-Spectrum Inhibitors of Assembly and Toxicity of Amyloid Proteins. *J. Am. Chem. Soc.* 133, 16958–16969 (2011).
68. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4, 435–447 (2008).
69. Daura, X. *et al.* Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* 38, 236–240 (1999).

70. Trott, O. & Olson, A. J. AutoDock Vina. *J. Comput. Chem.* 31, 445–461 (2010).
71. O'Boyle, N. M. *et al.* Open Babel: an open chemical toolbox. *J. Cheminform.* 3, 1–14 (2011).
72. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213 (2016).
73. Shoichet, B. K. Screening in a spirit haunted world. *Drug Discov. Today* 11, 607–615 (2006).
74. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods* 9, 676–682 (2012).
75. Patriksson, A. & Van Der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* 10, 2073 (2008).
76. Mylonas, E. *et al.* Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* 47, 10345–53 (2008).
77. Abdi, H. Partial Least Squares (PLS) regression. in *Encyclopedia of Social Science Research Methods* (eds. Lewis-Beck, M., Bryman, A. & Futing, T.) 792–795 (Sage, 2003).
78. Plewczynski, D., Spieser, S. A. H. & Koch, U. Performance of machine learning methods for ligand-based virtual screening. *Comb. Chem. High Throughput Screen.* 12, 358–68 (2009).
79. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
80. Nath, A., Zientek, M. A., Burke, B. J., Jiang, Y. & Atkins, W. M. Quantifying and predicting the promiscuity and isoform specificity of small-molecule cytochrome P450 inhibitors. *Drug Metab. Dispos.* 38, (2010).
81. Ihlenfeldt, W.-D. D., Takahashi, Y., Abe, H. & Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* 34, 109–116 (1994).
82. Bulic, B., Pickhardt, M., Mandelkow, E. E.-M. M. & Mandelkow, E. E.-M. M. Tau protein and tau aggregation inhibitors. *Neuropharmacology* 59, 276–289 (2010).
83. Shen, Y. & Bax, A. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* 48, 13–22 (2010).
84. Barre, P. & Eliezer, D. Structural transitions in tau k18 on micelle binding suggest a hierarchy in the efficacy of individual microtubule-binding repeats in filament nucleation. *Protein Sci.* 22, 1037–1048 (2013).

Chapter 3
Structure-Activity Relationships of Novel Tau Ligand Families

Introduction

IDPs are integral in signaling and transcription pathways and are often found as protein scaffolds.[1,2] These functions require versatile binding to multiple partners which makes IDPs and IDRs uniquely suited to fill these roles. As their natural utility has become clearer, so has the involvement of IDPs in pathology.[3] Often these diseases are related to the dysregulation of the natural function of an IDP, leading to aberrant signaling and transcription. With this in mind, it is unsurprising that IDPs are often key players in cancer. Additionally, IDPs are prone to misfolding and aggregation, are hallmarks of proteopathies such as Alzheimer's disease, Parkinson's disease, and Type II diabetes.[4–6] As such, the development of small-molecule ligands that interact with and modulate the activity of IDPs has been a long-standing goal.

Drug development strategies intended to identify and optimize ligands for well folded proteins are frustrated when applied to disordered proteins. The process of drug design begins with identifying lead compounds. These lead compounds can be found either through high-throughput screening or a rational approach based on protein structure. The majority of lead compounds for IDPs have been identified serendipitously or through high-throughput screening (HTS).[7] For example, compounds that target the IDP interaction between cMyc and Max were identified using a HTS method developed using a hybrid-yeast method that reported on this interaction.[8] Although often successful, these approaches are costly in terms of time, effort, and expense. The first step in a rational structure-guided approach to identify a target structure, which yields the information needed to identify suitable ligands. The conformational plasticity of

IDPs makes this a major challenge because an IDP's flexible nature results in a diverse conformational ensemble instead of a single well-populated structure that many techniques that determine structure rely on. After a suitable lead compound is found, the chemical structure is optimized to increase affinity and mitigate other issues such as solubility or off-target effects. The compounds identified targeting cMyc and Max interactions were further optimized by building a pharmacophore model and searching chemicals related to the originally identified lead compound.[9,10] Searching the chemical space surrounding lead compounds can lead to compounds with increased affinity, bioavailability, or mitigated side-effects.

Tau is an IDP that that stabilizes microtubules, regulates neurite outgrowth and also axonal trafficking in its physiological role.[11,12] It is implicated in a family of degenerative diseases called tauopathies, including Alzheimer's disease, frontotemporal dementia, Lewy body dementia and chronic traumatic encephalopathy. In these diseases, tau dissociates from the microtubule and is found as an insoluble amyloid aggregate.[13–15] For both the natural function and disease states, a specific region called the microtubule binding region (MTBR) is responsible for much of tau's binding activity. When expressed independently, the MTBR recapitulates tau's microtubule binding activity as well as its pathological aggregation, and has served as a reliable model system.

Chapter 2 established computational and *in vitro* techniques to identify lead compounds that bind to a MTBR construct, Tau4RD. Repeated simulated annealing molecular dynamics (ReSA-MD), a novel enhanced sampling technique, was used to generate a conformational library of Tau4RD conformers and computational clustering techniques identified locally persistent structure within. These local persistent structures were then used as targets for a molecular docking screen, and from those results ten promising compounds were selected. Of these ten

compounds, two showed distinctly different fluorescence traces from vehicle control in a Thioflavin-T (ThT) based aggregation assay. Compound 1 delayed and decreased the amount of ThT fluorescence, and Compound 6 increased the amount of fluorescence, both in dose-dependent manners.

This chapter builds on this breakthrough by using compounds 1 and 6 as leads to explore nearby chemical space and understand the structure-activity-relationship (SAR) of various analogs. While this has been performed on numerous well-folded proteins and is a crucial step in ligand design in the pharmaceutical industry, there are comparatively few examples of this approach when working with disordered targets. Here we not only find new small-molecule ligands of tau, but also identify SAR for compounds belonging to a novel class of tau aggregation inhibitors.

Experimental Procedures

Chemicals. Experimental compounds were purchased from either Chembridge or Enamine, detailed in Table 3.1. All other reagents were obtained from Sigma-Aldrich and used without further purification.

Protein Expression and Purification. Recombinant Tau constructs were expressed and purified from *E. coli* as described in Chapter 2.

Aggregation Assays. Two methods to determine aggregation were performed in this chapter. Stand-alone aggregation assays examined the effect of Tau4RD aggregation on ThT fluorescence and were performed as in Chapter 2. Comparative assays utilizing duplicate samples, one with ThT and one without. ThT samples served to monitor the aggregation and identify samples with unexpected aggregation kinetics. Samples without ThT were centrifuged at 37°C and 21100 g for 30 minutes and the supernatant extracted. Supernatant samples were then analyzed with one

or both of the tyrosine fluorescence assays, or absorbance assays, detailed in chapter 2 or below, respectively.

Tyrosine fluorescence assays. Tyrosine Fluorescence assays were performed as detailed in chapter 2.

Determination of compound 6 concentrations: Compound 6 concentration was determined by measuring absorbance at 248 nm. Compound 6's absorbance is sufficiently distinguishable from background absorbance of buffer and Tau4RD. Absorbance of samples was recorded between 195 and 345 nm using a Cary 3E spectrophotometer. The reference cell contained 1% DMSO in buffer. Raw absorbance data were baseline adjusted such that the 345 nm absorbance was set to 0. A standard curve was constructed using buffer-matched samples of compound 6 and used to determine soluble compound levels.

Gel Densitometry. Gel Densitometry was used as an orthogonal means to quantify soluble protein after aggregation as described in Chapter 2.

Computational Docking. Three-dimensional models of each compound were downloaded from the Pubchem chemical library and then converted to Autodock's pdbqt format using OpenBabel and Autodock Tools.[16–19] Compounds were then docked against all 250 peptide targets that were established in chapter 2, using the same docking protocol.

Compound affinity model. We derived an equation to describe the equilibrium between soluble and aggregate-bound states of a drug such as compound 6 as follows. Let drug *C* bind to the soluble states *S* of a protein (whether monomeric or oligomeric) with dissociation constant K_s:

$$S + C \rightleftharpoons SC;\ K_s = \frac{[S][C]}{[SC]} \quad \text{.....(1)}$$

Let drug *C* also bind to the insoluble fibrillar states *F* of a protein with dissociation constant K_f:

$$F + C \rightleftharpoons FC;\ K_f = \frac{[F][C]}{[FC]} \quad \text{.....(2)}$$

Then the fraction *f* of compound associated with insoluble states is given by:

$$f = \frac{[FC]}{[C]+[SC]+[FC]}$$

$$= \frac{{}^{[F][C]}/_{K_f}}{[C]+{}^{[S][C]}/_{K_s}+{}^{[F][C]}/_{K_f}}$$

$$= \frac{{}^{[F]}/_{K_f}}{1+{}^{[S]}/_{K_s}+{}^{[F]}/_{K_f}} \quad \text{.....(3)}$$

We begin the development of our pull-down model by establishing two equations representing the affinity of our compound (C) for soluble (S) and fibrillar Tau4RD: K_s and K_f respectively. Then we express the amount of Compound 6 pulled down as a ratio of the compound pulled down with fibrillar tau [FC] and the combined total of compound concentrations: free compound [C], compound associated with soluble tau [SC] and the compound associated with the fibrillar tau [FC]. We then substitute the earlier established affinity relationships in place of the [SC] and [FC] terms and simplify the ratio to get the final equation used to evaluate our data.

Note that [S] and [F] refer to the concentration of *unbound* soluble and fibrillar species respectively, not including any tau bound to the compound. As long as the dissociation constants

K_f and K_s are higher than the total concentrations of tau, then the difference between unbound and total concentrations of soluble or fibrillar tau is negligible. Substituting $[S] \approx [S_{tot}]$ and $[F] \approx [F_{tot}]$ into equation 3, we can then fit the spin-down data to estimate the values of K_f and K_s. However, this procedure yielded best-fit values of $K_f = 5.2 \pm 0.9$ μM and $K_s = 8.2 \pm 5.0$ μM. These values are close to the tau concentrations used (5–10 μM), and so the assumptions that $[S] \approx [S_{tot}]$ and $[F] \approx [F_{tot}]$ do not hold. Instead, we used numerical simulations of equations 1 and 2 implemented in CopasiUI (http://copasi.org). This approach allows the explicit computation of *f*, [F] and [S] at any given concentration of $[F_{tot}]$ and $[S_{tot}]$. Fitting the numerical model to experimental data, assuming 1:1 binding of tau and compound, yielded best-fit parameter values of $K_f = K_s = 0.5 \pm 1.5$ μM.

Results

Chapter 2 identified two compounds of interest, designated compound 1 and compound 6 (Figure 3.1). Compound 1 decreases and delays fluorescence in a ThT-based Tau4RD aggregation assay, which suggests it inhibits the aggregation of Tau4RD. In contrast, Compound 6 increased the amount of fluorescence in that assay. When probed further with orthogonal methods, we showed that Compound 1 was genuinely decreasing the amount of aggregation, but Compound 6's increase in fluorescence in the ThT assay was likely photophysical in origin. In this work we probe the structure-activity relationship of Compound 1 with chemical analogs. Then we investigate the interaction between Compound 6 and tau4RD and finally we explore the chemical space surrounding that compound.

Compounds from the ChemBridge and Enamine chemical libraries that shared chemical similarities to parts or all of compound 1 were purchased and examined with computational and in vitro methods. (Figure 3.1). We have taken the liberty of labeling the 6 rings of compound 1

to facilitate discussion of chemical differences between it and related molecules. Compounds 1.1, 1.2, and 1.3 share similarity with the "left" half of compound 1, consisting of the A, B, and C rings, but lack the remaining three. Compound 1.4 and 1.5 are very similar to the lead compound, differing only in the substituents of the B-ring. In place of a phenyl ring (ring A), Compound 1.4 has a hexene, and compound 1.5 has two methyl groups. Finally, Compounds 1.6, 1.7, and 1.8 probe the impact of the ring systems on the opposite end of the molecule (rings D, E, and F). Contrasting the activities of compounds 1.6 and 1.7 against compound 1.4 provide insight to rings E and F, as they are otherwise identical to compound 1.4. Compound 1.8, when compared to Compound 1, provides information on the importance of the tetrazole ring (ring E) of the lead compound. This specifically informs on the impact of the tetrazole as compound 1.8 differs from compound 1 by a single nitrogen to carbon substitution in that ring. To assess the predicted behavior of these compounds, we used the same computational screening procedure that was used to identify compound 1 (Supplementary Table 3.1). Computational results predicted most of the compounds to bind in a similar position as the lead compound (Compound 1) with good docking scores.

After identifying these compounds their effects on aggregation were tested using the established Thioflavin T assay and their activity was compared to vehicle control and the aggregation in the presence of Compound 1. Apart from the lead compound, only compounds 1.4, 1.5 and 1.7 had notable and consistent activity when compared to vehicle control. For the compounds that showed activity, this activity was verified with both dose-dependence experiments as well as orthogonal methods of examining aggregation. Our dose dependent studies showed a direct relationship between dose and effect on aggregation as shown by ThT fluorescence assays (Figure 3.2: B-D) and we also found them to inhibit the aggregation of the

full-length tau construct (Figure 3.3). Tyrosine fluorescence assays of protein in solution corroborate the ThT data and show that compounds 1, 1.4, 1.5 and 1.7 inhibit amyloid formation of Tau4RD, and gel densitometry confirmed the activity of compounds 1 and 1.7 (Figure 3.4).

We sought to determine how Compound 6 was amplifying the fluorescence of the ThT assay. It appears to bind to the Tau4RD aggregate and increase the fluorescence of ThT through that interaction, but we could not discount the possibility that it somehow amplifies the ThT fluorescence without binding Tau4RD. To establish this, we first sought to determine if the compound was associating with aggregated Tau4RD. We determined the amount of compound 6 in solution by measuring the solution's absorbance at 248 nm. By comparing the amount in the supernatants of samples that were induced to aggregate with heparin to those that did not aggregate establish that samples with Tau4RD aggregation had notably depleted concentrations of compound 6.(Figure 3.2).

For this study, two amounts of initial Tau4RD were used (5μM and 10 μM) in combination with a variety of heparin concentrations. Increasing amounts of heparin increase the amount of aggregation and therefore precipitated protein at the end of the assay, although not in a reproducible manner. Multiple amounts of starting tau allowed us to probe the effect of differing amounts of soluble Tau4RD in the presence of similar amounts of fibrillar protein. This information combined with knowledge of the amount of compound removed from solution due to its association with the fibrillar tau allowed us to estimate the affinity of Compound 6 towards Tau4RD. The simplest model that fit the data well yielded an apparent dissociation constant of 0.5 ± 0.15 μM for Compound 6 binding to either soluble or fibrillar tau (Figure 3.7). The resolution of the data was not sufficient to determine whether Compound 6 bound with different affinities to soluble vs. fibrillar forms of tau. Conservatively, we can conclude that Compound 6

binds to both soluble and aggregated tau with an apparent dissociation constant lower than about 5 μM.

Once we established that Compound 6 is a genuine ligand of tau aggregate, we sought out other potential ligands related to it. We obtained 3 compounds chemically similar to compound 6 and tested them with our ThT assay (Figure 3.2D). All of the compounds showed a notable increase in ThT fluorescence, but like Compound 6, gel densitometry showed that they were not affecting the amount of aggregation occurring and the increase of ThT fluorescence is likely a photophysical effect (Figure 3.8). We were unable to build absorbance assays like those used to quantify Compound 6 binding, as the absorbance for the analogs was insufficient.

Discussion

Disordered proteins are a growing focus as they are increasingly shown to be involved in number of diseases, but their disordered nature frustrates attempts to design ligands to modulate their activity in disease states. Their disordered nature obfuscates structure-based ligand design as well as assays often used to assess ligand binding. We utilized computational methods in our previous work to accelerate the search for small molecule ligands of Tau4RD. Our computational efforts led us to 10 compounds, of which 2 showed interesting activity in our initial ThT screening. We investigated these compounds and found that Compound 1 was inhibiting the aggregation of Tau4RD, and Compound 6 was not changing the aggregation, but was amplifying the fluorescence. This work sought to expand our understanding of both compounds.

We also investigated the chemical space surrounding the other lead compound identified in our computational work, Compound 1. With this lead compound we were able to identify a diverse set of chemically related compounds and tested them with our computational method and ThT fluorescence assay. In contrast to the molecules related to compound 6, the compounds

related to compound 1 had a more diverse range of scores, with the highest scores concentrated around two segments of the peptide. This information suggests that targeting segments 3 and 21 may lead to the identification of more active compounds. When we obtained the compounds and tested them with our ThT assay, we were able to identify the contributions of different parts of Compound 1. Compounds 1.1, 1.2, and 1.3 had no activity and thus illustrated that the A, B, and C rings of Compound 1 are not sufficient for activity. Compounds 1.4 and 1.5 show that the substituents extending from ring B are not critical for activity but do influence it. Those two compounds have consistent activity, but are less effective at slowing aggregation than the lead compound. In our computational studies, these compounds were predicted to have a higher affinity to Tau4RD than compound 1, but the ThT traces suggest that their effect is weaker. It is unclear whether the lower potency is due to weaker binding, or a difference in binding mode of these compounds relative to compound 1 that decreases their effects on aggregation. It is likely that steric hindrance is a factor, and the B-ring substituents alter how well the compounds bind to Tau4RD.

Furthermore, the effects of compounds it appears as though the tetrazole (Ring E) is critical and the phenyl ring (Ring F) is detrimental. These results show that once a lead compound is found, classical medicinal chemistry approaches such a deciphering structure-activity-relationships (SAR) can be used to refine ligand activity, even for disordered proteins for which characterizing the “structure” of the protein/ligand complex is difficult or impossible.

Our results show that although compound 6 does not alter the aggregation of Tau4RD, it does bind to the aggregate and amplify the fluorescence of ThT as evidenced by the loss in absorbance at 248 nm (Figure 3.6). As the amount of aggregated tau varies (based on the amounts of heparin and protein used), the ratio of compound loss to aggregate provides insight towards the

compound's affinity towards Tau4RD. In order to estimate the affinity of compound 6 towards Tau4RD, we needed to first develop a model that didn't depend on knowing the amount of compound associated with soluble protein. Our model is based on the minimal assumption that the association of compound 6 to Tau4RD reaches equilibrium by the time we examine it, and that compound bound to the fibrillar species is spun down with the aggregated protein. This leaves the amount in solution the sum of the unbound compound and the compound that is bound to soluble species of Tau4RD. Samples of compound 6 without any fibrillar tau (prepared by incubating them without any heparin) behaved the same as samples with no tau, and neither displayed any loss of compound (data not shown). Our model takes these basic assumptions and manipulates them (detailed in the Methods section) to estimate the affinity of compound 6 to Tau4RD. Comparing the data to simulated curves made by calculating the expected amount of compound pulled down for given affinities towards soluble and fibrillar Tau4RD suggests that the affinity is in the low micromolar range (Supplementary Figure 3). The resolution of the data is not sufficient to make any determinations about the relative affinities of the compound towards the soluble and fibrillar forms of Tau4RD.

With this we have shown that the original computational screen was successful when it identified compound 6 as a potential Tau4RD ligand in chapter 2. The original goal of that search was to identify compounds that bind to tau whether or not they change how tau aggregates. To instead specify the search for compounds that affect aggregation, we would need a better understanding of what areas of Tau4RD drive aggregation. This would enable us to limit our search to those regions, searching for compounds that interfere with the intramolecular interactions that drive aggregation. When we searched the chemical space near compound 6, we found compounds with similar activity and as such developed a novel class of amyloid ligand.

Additionally, when we look at the binding profiles of Compound 6 and its analogs, we see that while they seem to have a preferred location on Tau4RD to dock onto, they can viably dock across many segments of the protein. Small molecule ligands have previously been identified whose affinity towards disordered proteins through multiple weak interactions as opposed to a single specific interaction.[20] Therefore, if compound 6 and its structural analogs do not have a specific binding site it would not be unprecedented.

This chapter took the next logical step in ligand development after identifying lead compounds in chapter 2. We explored the chemical space surrounding those lead compounds using computational and *in vitro* methods and discovered more active compounds. Much like efforts that predated high-resolution protein structures were readily available, the activity and structure of the ligands guides our understanding of their interactions without the need for high resolution representations of binding interactions.

Figures

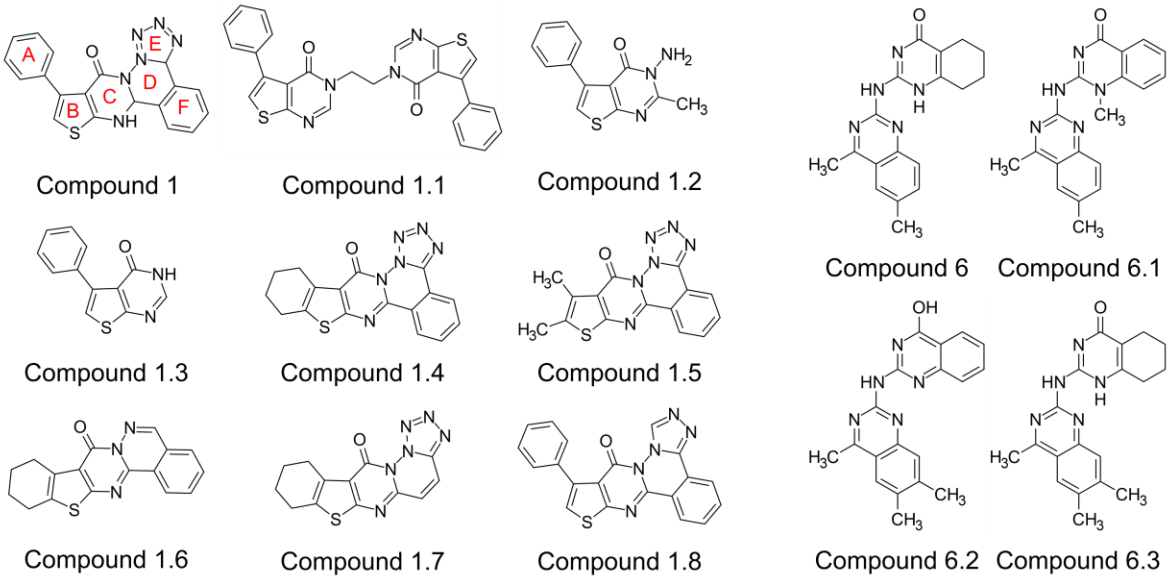


Figure 3.1: Structures of Compounds 1 and 6, as well as examined analogs. Compound 1's rings are labeled with red letters to facilitate discussion on chemical analogs.

Compound	Best Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Compound 1 CBID: 7579671 PCID: 1184213	-8	-5.4	-5.7	-8	-6.2	-5.4	-5.6	-5.7	-6.4	-5.3	-6	-6	-5.6	-5.6	-6.2	-6.3	-5.9	-6	-5.3	-5.5	-5.6	-6.9	-6.2	-5.8	-5.4	-5.4
Compound 1.1 CBID: 7912026 PCID: 1573149	-6.2	-5.1	-5.8	-5.7	-5.6	-5.5	-5.2	-5.7	-6.2	-5.3	-5.5	-5.7	-5.4	-5.8	-6.1	-5.7	-5.6	-5.7	-5.9	-5.4	-5.9	-6.1	-5.9	-5.3	-5.7	-5.4
Compound 1.2 ENID: EN300-03607 PCID: 206184	-5.8	-3.9	-4.8	-5.8	-4.3	-4.3	-4.1	-4.4	-5.2	-4.1	-4.8	-4.5	-4.3	-4.2	-5	-5.2	-4.6	-4.6	-4.6	-4.5	-4.7	-5.6	-4.5	-4.4	-4.5	-4.5
Compound 1.3 ENID: Z2574932101 PCID: 215639	-5.5	-3.7	-4.4	-5.1	-4.2	-4.2	-4.1	-4.5	-4.8	-4.1	-4.8	-4.2	-3.9	-4.1	-4.5	-4.9	-4.6	-4.2	-4.4	-4.4	-4.6	-5.5	-4.5	-4.5	-4.3	-4.4
Compound 1.4 ENID: Z56881708 PCID: 829637	-8.3	-5.4	-6.1	-8.3	-5.5	-5.7	-6.2	-5.8	-6.5	-5.6	-6.2	-6	-5.5	-5.7	-6.2	-6.2	-5.9	-6	-5.3	-5.7	-5.8	-6.8	-6.2	-5.8	-5.3	-5.6
Compound 1.5 ENID: Z56881738 PCID: 2943210	-8.2	-5.2	-5.6	-8.2	-5.3	-5.6	-6	-5.4	-5.9	-5.3	-5.9	-5.5	-5.4	-5.2	-5.7	-5.9	-5.8	-5.8	-5.1	-5.2	-5.5	-6.7	-5.9	-5.6	-5	-5.2
Compound 1.6 ENID: Z55163993 PCID: 705880	-6.2	-4.8	-5.4	-6.2	-5	-5	-5.6	-5.5	-6.1	-4.8	-5.8	-5.5	-4.8	-5.2	-5.5	-5.6	-5.2	-5.2	-5	-4.9	-5.7	-6	-5.8	-5	-4.8	-5.3
Compound 1.7 CBID: 7202175 PCID: 710970	-7	-4.8	-5.7	-7	-5.2	-5	-5.1	-5.2	-5.8	-4.9	-5.8	-5.4	-5.1	-4.7	-5.5	-6	-5.2	-5.5	-4.8	-5.1	-5.4	-6.2	-5.5	-5.3	-5	-5.2
Compound 1.8 ENID: Z56830444 PCID: 985278	-7	-5.2	-5.5	-7	-6.2	-5.2	-5.6	-5.5	-6.2	-5	-5.7	-5.6	-5.2	-5.6	-6.3	-6	-5.7	-5.8	-5.2	-5.3	-5.7	-6.9	-5.9	-5.5	-5.2	-5.4
Compound 6 CBID: 7282131 PCID: 135401693	-6.7	-5.1	-6	-6.7	-5.5	-5.7	-5.6	-5.9	-6.3	-5.3	-6.4	-5.8	-5.5	-5.7	-5.8	-6.6	-6.7	-5.8	-5.9	-5.8	-6.2	-6.5	-5.9	-5.7	-5.6	-5.8
Compound 6.1 CBID: 6589563 PCID: 647882	-6.7	-5.2	-5.9	-6.7	-5.3	-5.7	-5.6	-6.1	-6.1	-5.5	-6.2	-5.6	-5.6	-5.6	-6	-6.4	-6	-5.6	-5.9	-5.5	-5.8	-6.7	-5.8	-5.9	-5.6	-5.5
Compound 6.2 CBID: 6105364 PCID: 135408340	-7.1	-5.2	-6.2	-6.9	-5.6	-6	-5.7	-6.3	-6.6	-5.6	-6.3	-5.8	-5.7	-5.9	-6.2	-7	-7.1	-6.2	-6	-5.8	-6.3	-6.8	-6	-6	-5.8	-6
Compound 6.3 CBID: 7284101 PCID: 135401694	-6.9	-5.2	-6.1	-6.8	-5.6	-5.9	-5.7	-6.3	-6.6	-5.6	-6.2	-5.9	-5.7	-5.9	-6.1	-6.9	-6.9	-6.1	-6	-5.8	-6.3	-6.8	-6.1	-5.9	-5.7	-5.9

Table 3.1: Each compound is listed along with the supplier's identification number (Chembridge's CBID or Enamine's ENID). The best docking scores for each compound examine in this study for each of the 25 segments of Tau4RD. Data shown is the best score against any of the 10 most prominent configurations for each given segment and is color coded (Predicted poor interactions are red and predicted favorable interactions are green).

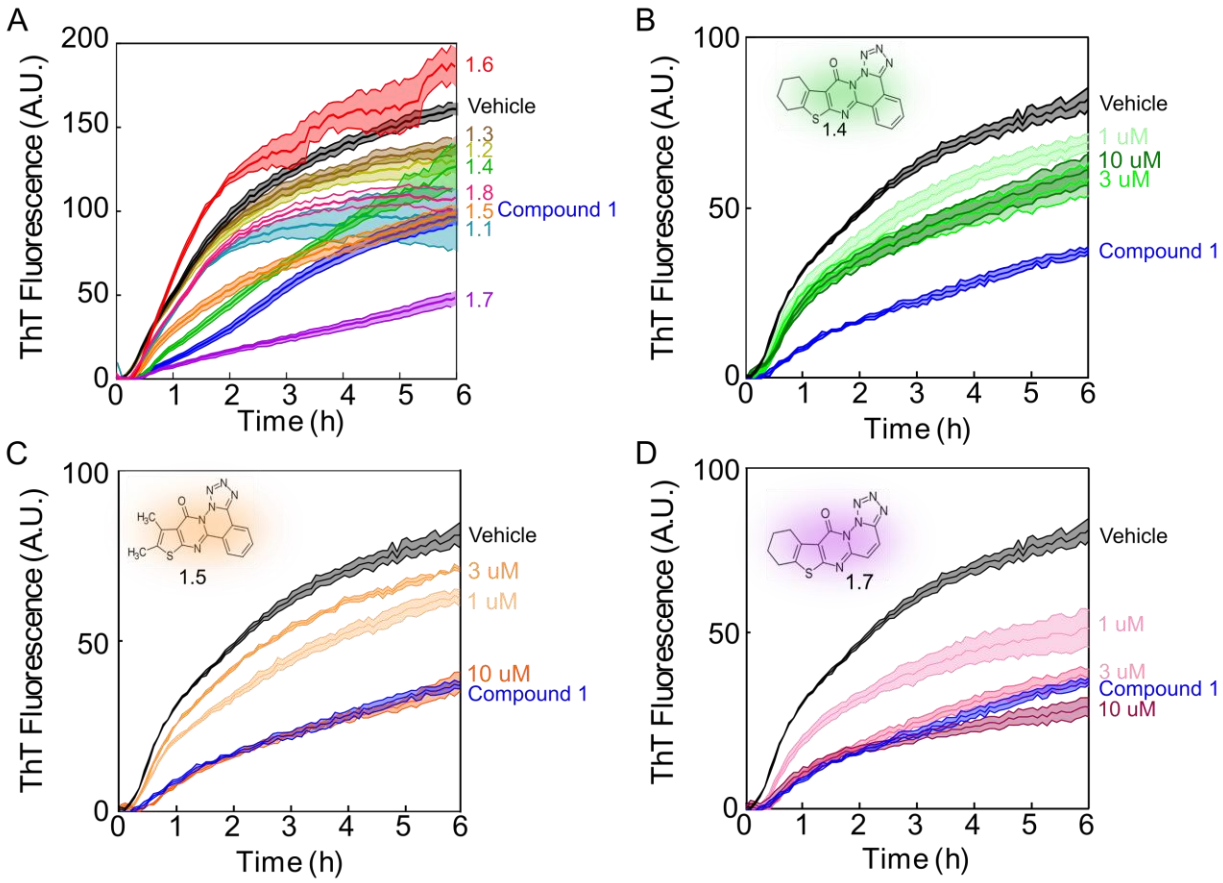


Figure 3.2: A) ThT fluorescence based aggregation assay showing molecules related to Compound 1 display different activities at 10uM. B) Active Compounds display dose-dependent effects at 1, 3, 10uM when compared to vehicle control (black) and 10uM Compound 1(blue). Data shows three technical replicates + SEM and is representative of three independent experiments.

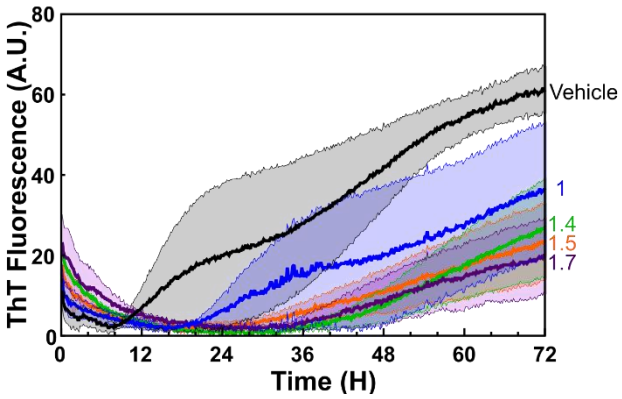


Figure 3.3: ThT Fluorescence assay demonstrates that compound 1 and selected analogs inhibit full length tau aggregation at 10uM concentration. Data shows three technical replicates + SEM and is representative of three independent experiments.

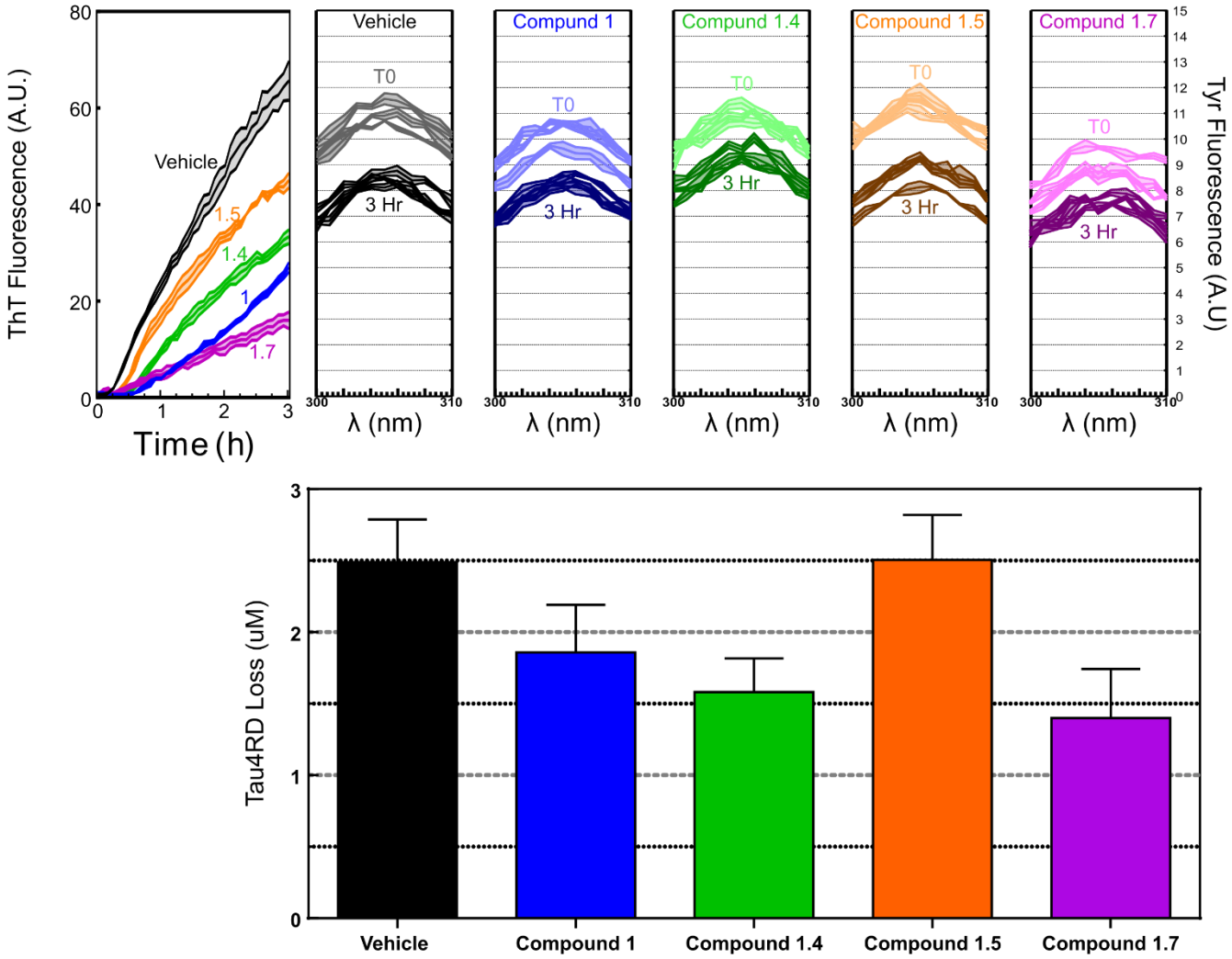


Figure 3.4: Tyrosine Fluorescence Assays shows that significantly less Tau4RD is precipitated out of samples containing 10uM Compound 1, 1.4, or 1.7. This corroborates the ThT data that was run simultaneously. ThT data shows means ± SEM of 3 technical replicates, and Tyr fluorescent data explicitly shows 3 technical replicates of pre aggregate sample compared against 3 technical replicates of non-aggregated sample. Values of Tau4RD precipitation were extrapolated using a standard curve and losses determined for this experiment are shown below tyrosine fluorescence traces. Data shown is representative of three independent experiments

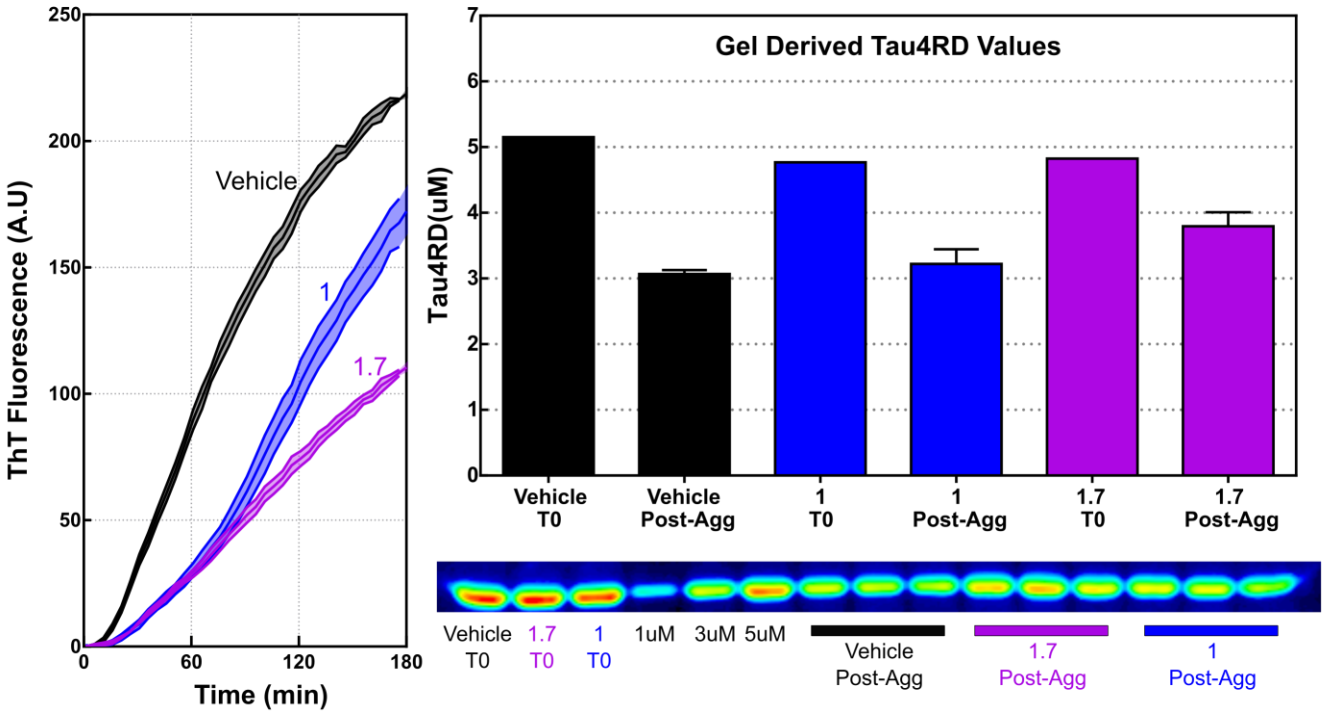


Figure 3.5: After aggregating for 3 hours, Gel densitometry data corroborates trends of ThT Fluorescence Data, showing less Tau4RD precipitated in samples containing 10uM compound 1.7 or 10µM Compound 1 when compared to vehicle Control. ThT Data shows three technical replicates + SEM. This data is representative of three independent experiments

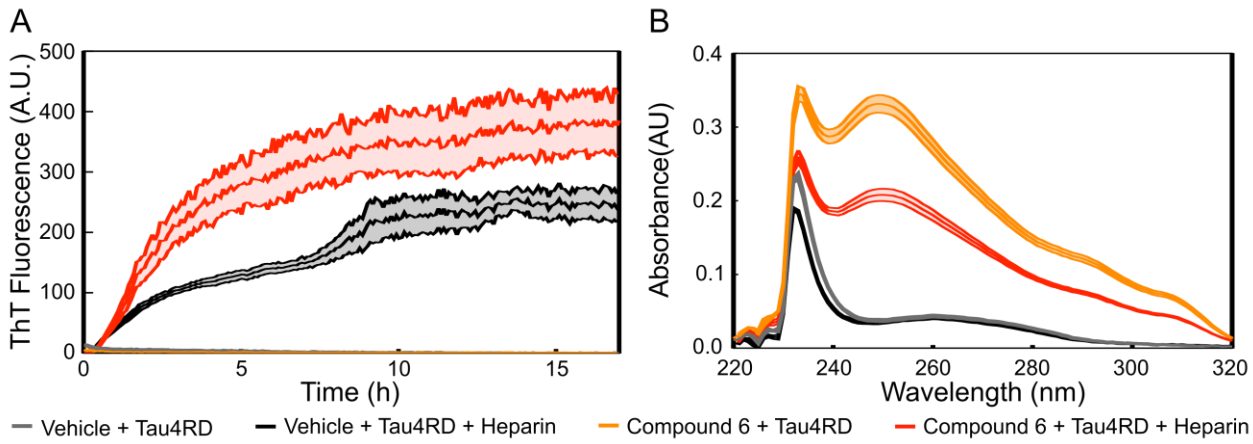


Figure 3.6: A) Compound 6 increases the ThT fluorescence signal associated with amyloid fibrils only when the protein aggregates (induced by the presence of heparin). B) After overnight aggregation, the absorbance of compound 6 is notably decreased.

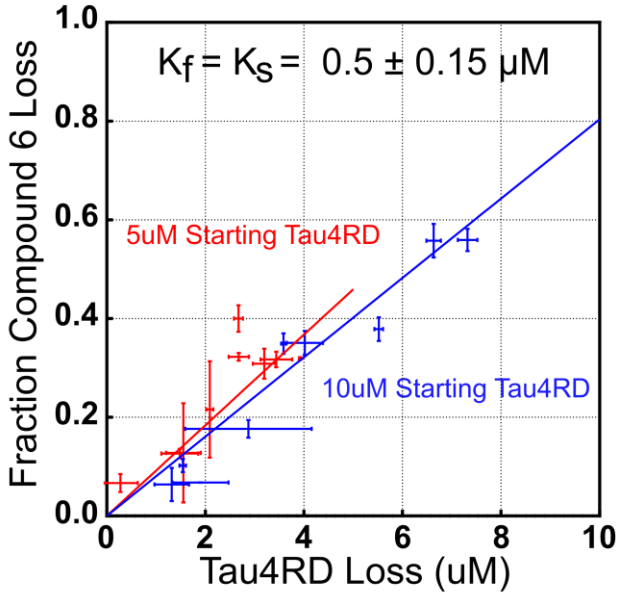


Figure 3.7: Experimentally derived amounts of Tau4RD Loss and Compound 6. The solid lines show fits to equation 3. Fitted values for K_s (affinity of the compound towards soluble Tau4RD) and K_f (affinity of the compound towards Tau4RD fibril) are shown at the top of the graph.

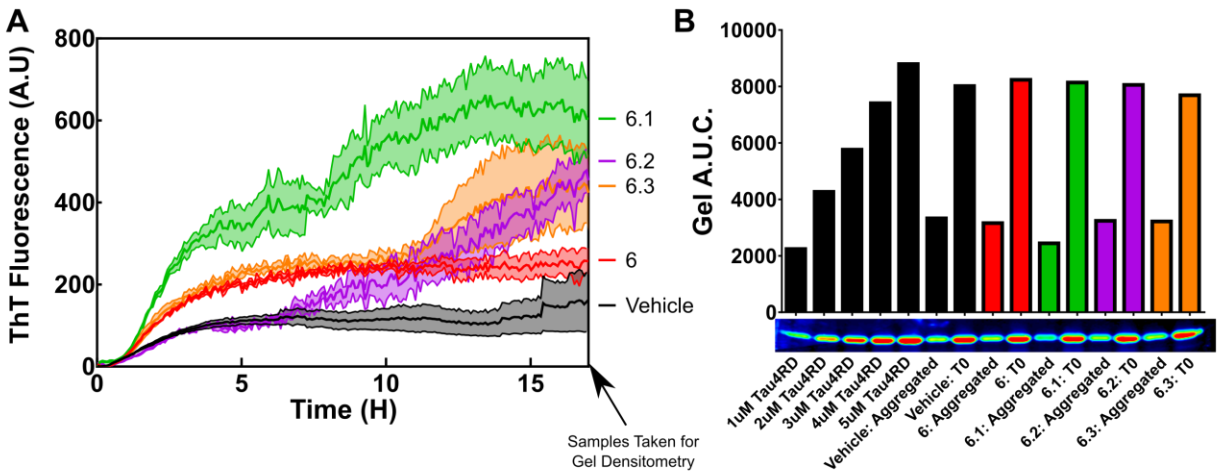


Figure 3.8: A) ThT traces of Tau4RD with 10uM of selected compounds shows that compounds 6.1, 6.2, and 6.3 increase ThT fluorescence in a similar manner as compound 6. ThT data shows mean + SEM of 3 technical repeats and is representative of 3 independent experiments. B) Gel densitometry of end-point samples illustrate that the amount of aggregation is not significantly different from vehicle samples, as would be suggested by ThT data.

References

(1) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and Structure of Inherently Disordered Proteins. *Curr. Opin. Struct. Biol.* **2008**, *18* (6), 756–764.

(2) Uversky, V. N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics*. Frontiers Media S.A. 2019.

(3) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D ^{2} Concept. *Annu. Rev. Biophys.* **2008**, *37* (1), 215–246.

(4) Khanam, H.; Ali, A.; Asif, M.; Shamsuzzaman. Neurodegenerative Diseases Linked to Misfolded Proteins and Their Therapeutic Approaches: A Review. *Eur. J. Med. Chem.* **2016**, *124*, 1121–1141.

(5) Outeiro, T. F.; Putcha, P.; Tetzlaff, J. E.; Spoelgen, R.; Koker, M.; Carvalho, F.; Hyman, B. T.; McLean, P. J. Formation of Toxic Oligomeric α-Synuclein Species in Living Cells. *PLoS One* **2008**, *3* (4), 1–9.

(6) Tóth, G.; Gardai, S. J.; Zago, W.; Bertoncini, C. W.; Cremades, N.; Roy, S. L.; Tambe, M. A.; Rochet, J.-C. C.; Galvagnion, C.; Skibinski, G.; et al. Targeting the Intrinsically Disordered Structural Ensemble of α-Synuclein by Small Molecules as a Potential Therapeutic Strategy for Parkinson's Disease. *PLoS One* **2014**, *9* (2), e87133.

(7) Metallo, S. J. Intrinsically Disordered Proteins Are Potential Drug Targets. *Curr. Opin. Chem. Biol.* **2010**, *14* (4), 481–488.

(8) Yin, X.; Giap, C.; Lazo, J. S.; Prochownik, E. V. Low Molecular Weight Inhibitors of Myc–Max Interaction and Function. *Oncogene* **2003**, *22* (40), 6151–6159.

(9) Wang, H.; Hammoudeh, D. I.; Follis, A. V.; Reese, B. E.; Lazo, J. S.; Metallo, S. J.; Prochownik, E. V. Improved Low Molecular Weight Myc-Max Inhibitors. *Mol. Cancer Ther.* **2007**, *6* (9), 2399–2408.

(10) Mustata, G.; Follis, A. V.; Hammoudeh, D. I.; Metallo, S. J.; Wang, H.; Prochownik, E. V; Lazo, J. S.; Bahar, I. Discovery of Novel Myc-Max Heterodimer Disruptors with a Three-Dimensional Pharmacophore Model. *J. Med. Chem.* **2009**, *52* (5), 1247–1250.

(11) Drubin, D. G.; Kirschner, M. W. Tau Protein Function in Living Cells. *J. Cell Biol.* **1986**, *103* (6 Pt 2), 2739–2746.

(12) Gustke, N.; Trinczek, B.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. Domains of τ Protein and Interactions with Microtubules. *Biochemistry* **1994**, *33* (32), 9511–9522.

(13) Wang, Y.; Mandelkow, E. Tau in Physiology and Pathology. *Nat. Rev. Neurosci.* **2015**, *17* (1), 22–35.

(14) Goedert, M.; Eisenberg, D. S.; Crowther, R. A. Propagation of Tau Aggregates and Neurodegeneration. **2017**.

(15) Arendt, T.; Stieler, J. T.; Holzer, M. Tau and Tauopathies. *Brain Res. Bull.* **2016**, *126*, 238–292.

(16) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213.

(17) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*.

(18) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (10).

(19) Trott, O.; Olson, A. J. AutoDock Vina. *J. Comput. Chem.* **2010**, *31*, 445–461.
(20) Iconaru, L. I.; Ban, D.; Bharatham, K.; Ramanathan, A.; Zhang, W.; Shelat, A. A.; Zuo, J.; Kriwacki, R. W. Discovery of Small Molecules That Inhibit the Disordered Protein, P27 Kip1. *Sci. Rep.* **2015**, *5*.

Chapter 4
Novel Modulators of Phenol Soluble Modulin Aggregation and Membrane Disruption

Introduction

Phenol soluble modulins (PSMs) are a recently discovered class of short peptide involved in the virulence of staphylococci bacterial including *S. aureus*.[1] *S. aureus* colonies are very common in humans, and benignly colonize the skin of about 1/3 of the population.[2,3] However, *S. aureus* can also display antibacterial resistance and develop into serious infections that constitute a global health concern.[4] In particular, methicillin-resistant *S. aureus* (MRSA) accounts for 120,000 bloodstream infections and nearly 20,000 deaths per year.[5] Understanding the factors that facilitate the ability of these opportunistic infections to evade treatment options would expedite the formulation of more effective therapies and is of great importance.

PSMs were discovered and named by the Kebanoff group in 1990 during a hot phenol extraction of *S. epidermidis* culture. In the phenol-soluble portion of that extraction, they identified 3 peptides that they thought might contribute to gram-positive bacteria sepsis.[6] Since then, multiple PSMs have been identified and are classified as one of two types: the shorter (20-24 residue) α-type, and the longer β-type (44 residues). They have multiple functions in pathology, including the ability to attract and lyse neutrophils[7], a role in biofilm formation and structure, [8] and a role inciting a host's inflammatory immune response when at low concentrations.[9]

PSM expression is subject to regulation by the *agr* system. The *agr* system is responsible for recognizing the local concentration of bacteria and eliciting the proper cellular response, a process called quorum sensing.[10] This regulation includes the formation of biofilms, a mode of bacterial growth characterized by surface-attached organisms and an extracellular matrix that

helps protect the colony. Biofilm formation offers a resistance to mechanical interference, host defenses, as well as antibiotic treatment, and 65% of all microbial infections and 80% of chronic infections involve biofilms.[11] All known PSM classes have some level of involvement in biofilm formation or function. [12,13]

In vitro, PSMs have been shown to make aggregates that displays similar formation kinetics and have similar chemical properties to amyloid fibrils.[14] A notable difference from the canonical amyloid fibril is that PSMs form a fibrillar structure composed of alpha-helical proteins as opposed to the beta-sheet structure typical of amyloid.[15] It has been hypothesized that the amyloid-like PSM fibrils are structural components that help form and stabilize bacterial biofilms,[16] but there is also contradictory evidence that suggests that PSMs amyloids are not important to biofilm structure. When compared to a wild-type strain, a PSM-deficient strain had no noticeable changes in biofilm formation, suggesting PSMs are not necessary.[17] The exact role of PSMs and biofilms still remains somewhat elusive and tools are needed to further probe and elucidate this dynamic.

All PSM peptides have the potential to form amphipathic alpha helices.[1] In solution, PSMs have pronounced helical structure, but are still dynamic and display unstructured behavior. Recent solution state NMR studies utilized the ability of trifluoroethanol to stabilize secondary structure, and derived models of multiple PSMs helices.[18,19] This helicity increases when interacting with membranes as evidenced by circular dichroism (CD), and suggests a possible method of toxicity.[20] When comparing the ability to induce membrane leakage across different PSMs, those with stronger helical propensity displayed greater ability to puncture phospholipid bilyaers.[21] Phospholipid bilayers composed of DOPC/Cholesterol/sphingomyelin (chosen to

mimic mammalian plasma membranes) accelerate the fibril formation of PMSα3, and in the process the PSM embedded itself into the membranes which suggests a mechanism of toxicity.[20]

Because of their involvement in a number of aspects of S. aureus pathology, PSMs are promising targets for anti-staphylococcal drug development.[1,22] As much of their pathogenic activity is related to alpha helical structure, modulating the propensity of PSMs to adopt that structure is an appealing strategy for modifying and understanding their activity. This chapter utilizes computational and *in vitro* approaches to identify small molecule compounds that can adjust PSM behavior by targeting its alpha helical structure.

Methods

Molecular Dynamics Simulations.

The PSM starting structures: 5khb (PSMα1), 5kgy (PSMα3)[19] and 5i55 (PSMα3)[15] were downloaded from the Protein Data Bank,[23] placed in a sufficiently large cubic box with periodic boundary conditions, and solvated using the tip4p water model and enough counter-ions to neutralize the charge of the simulation. After solvation, simulations were energy-minimized using a steepest-descent algorithm and then allowed to equilibrate for 100 ps. After equilibration, molecular dynamics were performed using the Molecular dynamics simulations using GROMACS 4.6.5[24] with the AMBER99SB-ILDN force field.[25] Molecular dynamics simulations were performed using PME electrostatic calculations, NVT ensemble with v-rescale thermostat, and a timestep of 2 fs.

Clustering and Molecular Docking.

After 20ns of molecular dynamics simulations were completed, resulting structural ensembles were clustered using the GROMOS algorithm[26] available in the suite of GROMACS tools. Median structures were chosen such that selected structures had sufficient resolution and

encompassed 2-3% of the overall MD trajectory. These structures were then converted to receptor models for docking in AutoDock Vina.[27] Structures of the ChemBridge CNS collection were obtained from PubChem and parameters were calculated using OpenBabel.[28] Ligand models were built using AutoDock's ligand preparation utility. Ligand docking was done using AutoDock Vina, with a search area encompassing the space of a cube extending 2 nm beyond the receptor model in each dimension.

Peptide production and storage.

Peptides samples for PSMa1 and PSMa3 were ordered and synthesized from Genscript (GenScript Corporation, Piscataway, NJ) as 1mg lyophilized samples. Prior to use peptide samples were dissolved in DMSO to a concentration of 5 mg/mL and stored at -80°C for no more than 2 weeks before use.

Preparation and Storage of Selected Compounds.

Selected compounds were identified from the ChemBridge CNS small molecule collection, and dry samples were obtained through Hit2Lead (ChemBridge, San Diego, CA). Samples were dissolved in DMSO to 10 mM stock solutions and stored at -80°C.

Aggregation Assays.

All fibrillization reactions were carried out at 37°C at pH 7.4, in buffer containing 50 mM HEPES buffer, 100 mM NaCl, and 50 μM thioflavin T (ThT). Aggregation assays of PSMα3 were carried out by mixing compound and PSM stocks in DMSO into buffer at final concentrations. Samples were then aliquoted in 100 μL increments into wells of a 96-well plate, where they were rapidly mixed with equal volumes of 3-KDa Heparin sodium. Final concentrations of peptide and heparin were 50 μg/mL and 3μM respectively. Assays examining PSMα1 were mixed at indicated concentrations, then 200 μL placed in wells of 96 well plates

without heparin. Reaction progress was monitored by increase in Thioflavin T fluorescence with λex = 440 nm and λem = 485 nm in a BioTek Synergy HTX (BioTek, Winooski, VT) plate reader. Measurements were taken every 5 minutes after 1 minute of agitation. Intrinsic fluorescence from compounds was in all cases ≤ 5% of the background, which was corrected for by baseline subtraction. Experiments were performed with at least 3 technical replicates.

Liposome Preparation. (Performed by Eleanor Vane)

Liposomes were generally prepared over the course of several days, and used within two weeks of preparation.

One hundred µL of 25 mM of lipid (50% POPC 50% POPG, purchased from Avanti, Alabaster, Alabama) were dissolved in chloroform using an appropriate glass Hamilton syringe. This solution was then evaporated with a gentle stream of nitrogen gas and left under a vacuum for at least 2 hours. Resulting films were then hydrated with 250 µL of filtered Calcein buffer:50mM HEPES, 100mM NaCL, 70mM Calcein, and 0.3mM EDTA, pH 7.4. This solution was covered with plastic film and left to equilibrate at room temperature for one hour, then freeze-thawed 10 times using liquid nitrogen and 42°C water baths. This was then extruded 23 times with a 0.1 µM PC membrane filter and extrusion kit. Free calcein was separated using a desalting column, with 50mM HEPES, 100mM NaCl pH 7.4 as the running buffer. Liposome containing fractions were transferred into a 7000 MWCO dialysis cassette and equilibrated against the above HEPES buffer for at least 3 hours. This buffer was replaced two times, with 3 hours of equilibration between each exchange. Once completed, samples were aliquoted, insulated from light, and stored at 4°C. Prior to use liposomes were examined with by phosphate assay to determine concentration.

Preliminary Leakage experiments:

A solution of compound and DMSO was prepared in the HEPES buffer detailed above at 118% of final concentration and 170 µL was placed in a well of a 96-well plate for each replicate. 20 µL of liposome at 10X final concentration was then added, and allowed to incubate for 20 minutes at 25°C with 5 seconds of shaking every minute. After incubation, 10 µL of peptide at 20X final concentration was added, the plate resealed, and continued to incubate for 80 minutes. Reaction progress was monitored by increase in calcein fluorescence with λ_{ex} = 485 nm and λ_{em} = 4520 nm in a BioTek Synergy HTX (BioTek, Winooski, VT) plate reader. Reads were taken every minute after 5 seconds of agitation. 0.5% Triton in place of peptide was used as a positive leakage control, and additional buffer was added in place of peptide and used as a negative control.

Results

We began our approach by using molecular dynamics to generate a set of relaxed conformations suitable for *in silico* docking studies. Starting structures were derived from either solution NMR or x-ray crystal structures of PSMα1 and PSMα3 available from the protein data bank (Figure 4.1 A). After energy minimization and a short equilibration with a small time step to ensure simulations started in a stable configuration, molecular dynamics simulations were run for 20 ns. Three of the simulations contain a single helix (5khb, 5kgy[19] and 5i55[15]) while the fourth simulation contains the 4 helices that comprise the 5i55 crystal structure.

The completed trajectories were then examined with clustering analysis using the GROMOS method to identify the most preferred conformations, selecting for a specific resolution with RMSD cutoff values. Groups of conformations were compiled into distinct "clusters" with each cluster represented by the median structure (the structure that had the lowest

average RMSD to every other structure in the cluster). The most populated clusters for each simulation were collected such that in aggregate they represented 2-3% of the trajectory and converted into a receptor format. Examination of the trajectory via RMSD and RMSF, as well as looking at the resulting clusters show that there were limited fluctuations in the helical nature of the peptide, and the residues most prone to movement were those closest to the ends of the peptide (Figure 4.1).

The receptors built from the clusters were then used as targets, and the entire CNS collection of the Chembridge chemical library was docked against them using Autodock vina. Resulting scores were collected, and 10 compounds selected based on these scores as well as the docking configurations they resulted from (Table 4.1 and Figure 4.2). These compounds were then examined for ability to modulate the pathological aggregation of PSMs and their ability to induce vesicle leakage. Compounds that elicited an effect on *in vitro* experiments are shown in figures 4.2 and 4.3. Compounds with notable and consistent changes in PSMα3 aggregation were then examined at different doses to examine the dose-dependent relationship of effect. Compounds PSM C2 and C8 both increased the amount of fluorescence, indicating an increase in aggregation. In contrast, compound C4 decreased the ThT fluorescence implying an inhibitory effect on aggregation. None of the compounds had a major effect on the t_{50} of amyloid formation. All three compounds exhibit a clear dose-dependent relationship and appear to have an EC_{50} in the low micromolar range although a more specific determination requires more investigation.

Discussion

The work in this chapter applies *in silico* and *in vitro* techniques that were utilized in Chapter 2 but modifies them to take advantage of the helical propensity of PSM. This helical

propensity is linked to pathological activity in a number of ways. 20 ns of molecular dynamics simulation resulted in conformational libraries for each model that show that the initial positions aren't the most suitable for computation, but they quickly settle to stable yet still somewhat dynamic structures. As would be expected for a short alpha helical peptide, the central regions are the most inert and fluctuations are mostly observed the N and C termini (Figure 4.1 D). The 4 helix simulation displays some notable changes in RMSD and RMSF for some helices, but this is largely a result of the global rearrangement from the original configuration of the 4 chains that then remains stable. This shows that the structures used for our docking procedure are at least somewhat stable, and reasonable to try and target with a small molecule.

The docking procedure we used was able to process compounds at a much faster rate than the one used in Chapter 2, largely because the number of targets was smaller. The number of receptor models in Chapter 2 was between 80 and 250 depending on the stage of docking. For this project there were a total of 15 structures and despite the fact that they were larger, the docking procedure for all of the compounds in the Chembridge CNS library (~48000) in a reasonable amount of time. With these data, we chose ten compounds for *in vitro* study from the thousands examined computationally. All of the compounds chosen displayed strong docking scores compared to the average calculated for the entire collection (-5.4), but other factors were influential in selecting these particular compounds. One of the reasons raw docking score was not the sole determining factor is that docking scores are not directly comparable across systems. Increases solvent-accessible surface area deep binding pockets capable of accommodating a ligand resulted in the 4-helix model had significantly higher average scores than any of the models that only depicted a single helix. In some cases, this was despite representing the exact same peptide ligand interaction. Other factors considered when selecting compounds included

calculated specificity (binding well to one target but not others) as well as cultivating a chemically diverse set of compounds (Table 4.1).

Small molecules that bind to these predicted alpha helices could feasibly interact in a number of different ways. Stabilization of this structure through binding would shift the energy landscape of the peptide population such that the alpha-helical population increases relative to disordered states. Such stabilization would make reactions that are dependent on this structure more favorable provided that the compound's bound position didn't interfere with the intermolecular forces that drive the interaction. In the case of fibrillization of PSMs, this would mean stabilizing the alpha helix structure without binding to the regions of the peptide that form intermolecular contacts with other PSMs. In membrane interactions, a compound could adjust the function of PSMs by stabilizing the requisite helix, but could also interfere with the proteins interface with the membrane if bound in the wrong position.

The only compound that reliably affects PSMα1 behavior in our *in vitro* assays is PSM compound 6. As shown in Table 1, PSM C6 is one of the highest scoring compounds against the PSMα1 clusters that we purchased. This suggests that the docking protocol that identified it from thousands of compounds in the ChemBridge library successfully guided us. Figure 4.3b shows that PSM C6 decreases the aggregation rate of PSM α1, but the mechanism behind this is unknown. The amyloid structure of PSMs has only been shown for PSMα3, and its possible that by stabilizing the alpha helical structure, PSM compound 6 decreases the population of aggregation viable conformations. Alternatively, the compound could be binding in such a way that it interrupts the intermolecular forces that drive aggregation. Further study including elucidation of structure upon binding would provide additional insight.

The compounds PSM C2, C4, and C6 all have dose-dependent effects on PSMα3 ThT fluorescence (Figure 4.4). As shown in Chapters 2 and 3 with the case of Compound 6, changes in ThT fluorescence are not always indicative of genuine changes in aggregation. Although the dose dependent nature of the effect is suggestive of at least a successful binding interaction, further study with orthogonal methods is warranted to gain a more complete understanding of what is occurring. Compounds PSM C2 and PSM C8 both appear to increase the amount of aggregation, and this could be done by increasing the amount of the PSM population that is aggregation viable by stabilizing helical structure. In contrast, PSM C4 seems to decrease the amount of aggregation. This could be due to it binding to the alpha helical structure in a way that prevents association with fibrils, or it could also be a result of the compound stabilizing a non-aggregation-prone conformation. Both of these mechanisms would decrease the population of PSM that can fibrilize and would effectively reduce the maximal amount of fibrilization.

The results in this chapter show that the computational methods employed were successful in guiding the selection of potential ligands of PSM peptides. Four of the ten compounds selected appear to at the very least interact with the PSMs in a meaningful way. This, in addition to data in Chapter 2 illustrates that computational tools are powerful and efficient methods of screening chemical libraries for promising compounds.

Figures

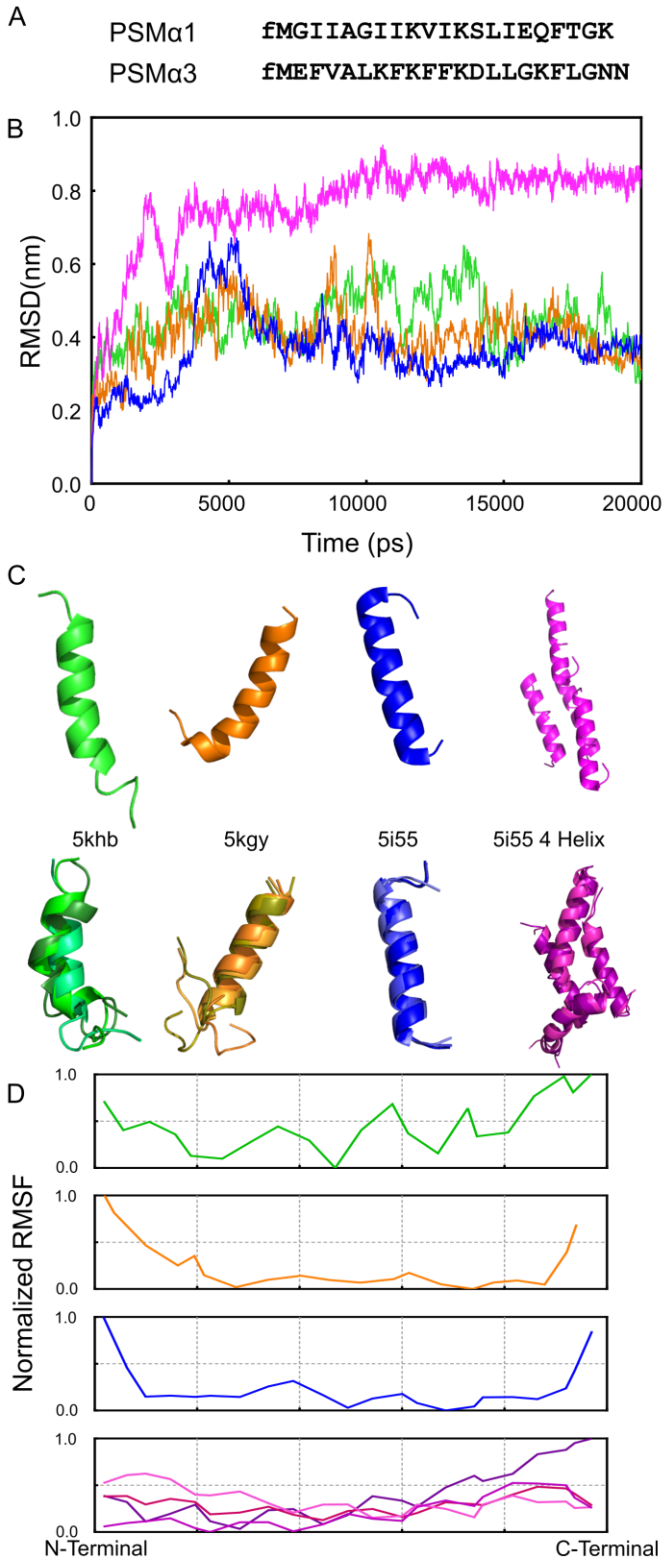


Figure 4.1: Structure and stability of PSM models *in silico.*
Different simulations are represented with different colors throughout this figure: 5KHB in Green, 5KGY in Orange, 5i55 (single helix simulation) in Blue, and 5i55 (4 helix simulation) in Purple.

A)The sequences of the two peptides examined in this chapter, PSMα1 and PSMα3.
B) RMSD plots of the molecular dynamics simulations show the amount of The molecular dynamics simulations of each helix initially relaxed from starting structures, as evidenced by an initial jump of RMSD. After this initial 5ns of settling, most changers were more gradual indicating an exploration of local conformational space.
C) Starting structures (above) and representative clusters (below) derived from PSM trajectories. PSMα1 initiated from the 5khb solution NMR structure in green. PSMα3 initiated from the 5kgy solution NMR structure is shown in orange and the single and unit cell 5i55 crystal structures in blue and purple respectively.
D) Normalized RMSF values of C-alpha carbons across the peptide quantify relative flexibility. Different helices of the 5i55 4-helix simulation are shown in different shades of purple.

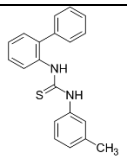
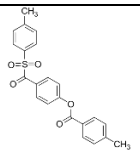
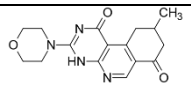
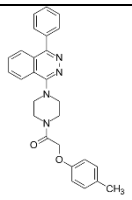
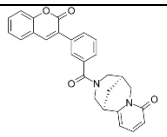
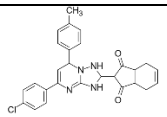
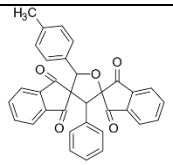
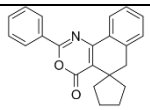
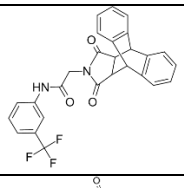
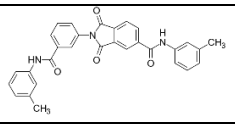
Designation	Pubchem ID	ChemBridge ID	Structure	5KHB score	5KGY score	5i55 single-helix score	5i55 4-helices score
PSM C1	976630	7727454		-6.1	-6.1	-5.8	-8.1
PSM C2	1422081	7772050		-5.9	-7.2	-6.8	-8
PSM C3	25252341	9192927		-5.3	-5.7	-5.7	-6.5
PSM C4	2951748	7747947		-6.6	-7.3	-7.9	-7.9
PSM C5	43912852	9274473		-6.6	-8	-7.5	-8.7
PSM C6	5738528	7645649		-7.2	-8.5	-8.3	-9.5
PSM C7	2840643	5309478		-6.7	-8	-7.5	-9.3
PSM C8	753933	5684107		-6.6	-8	-7.8	-8.8
PSM C9	2906394	6629705		-7	-8.2	-8.3	-9.1
PSM C10	1369732	5569062		-7.3	-8.3	-8.3	-9.3

Table 4.1: Selected compounds with their structures, and the best score against each simulation. Best scores are the best docking configuration found for that compound against any model derived from the simulation.

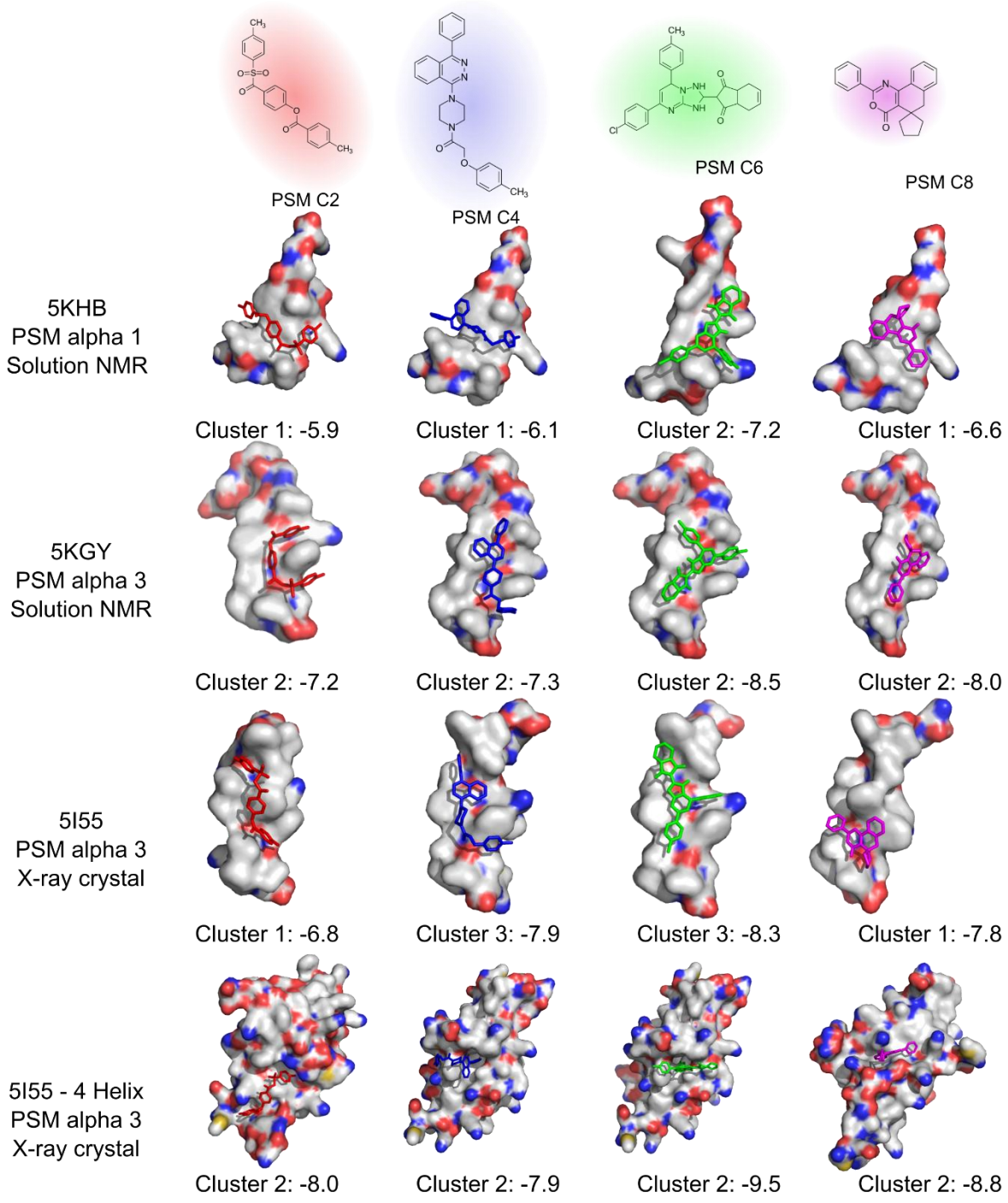


Figure 4.2: Structures of selected compounds above significant bound poses to PSM structures. The structures of active compounds is shown in the top row, above the preferred docking position against each of the 4 targets. Labels below docking poses specify which of the simulation's receptor models was most favorable, and what that interaction scored.

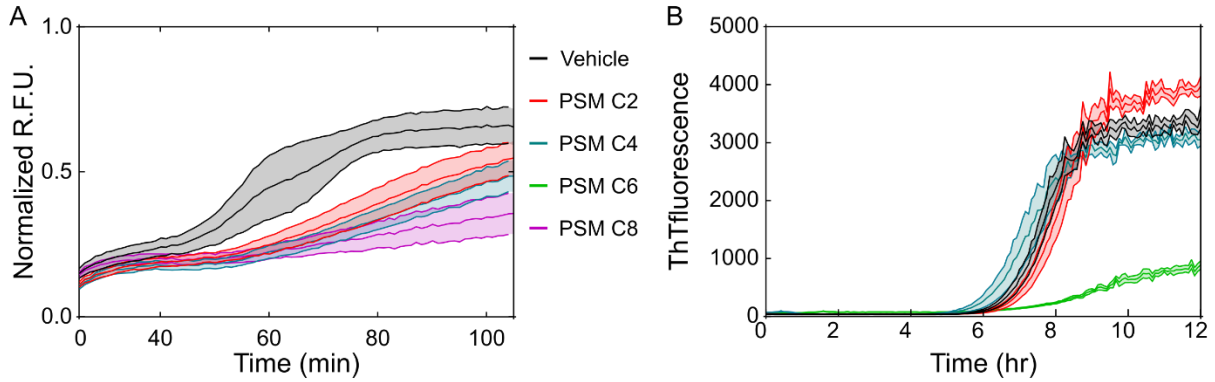


Figure 4.3: A) Preliminary results show that at 40µM, PSM compounds 1,2,and 4 impede vesicle leakage by PSM α3. B) Preliminary results show that compound 6 inhibits PSMα1 aggregation at 50µM. Traces show mean ± S.E.M of three technical repeats. Experiments planned and executed by Annie Dosey.

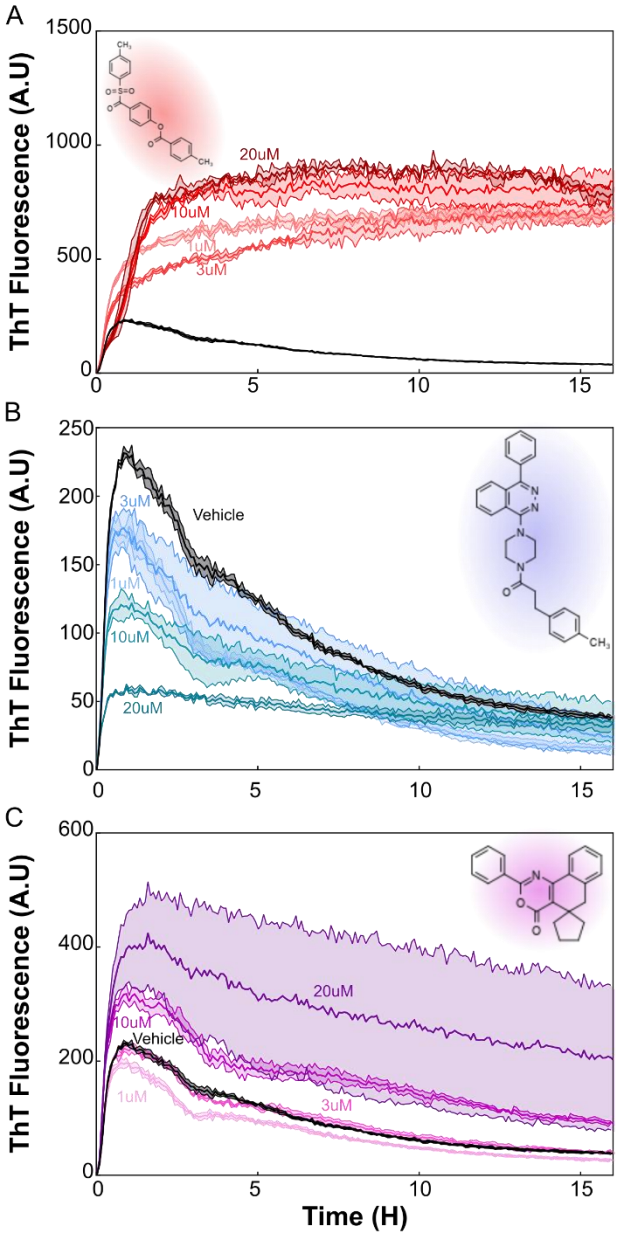


Figure 4.4: ThT-based aggregation assays of PSM Compounds 2, 4, and 8 (A, B and C respectively) all display dose dependent effects on the ThT fluorescence assay used to examine amyloid aggregation.

Acknowledgments
This work was done in part by Annie Dosey during as part of her rotation project for the Nath lab. Eleanor Vane helped plan vesicle leakage experiments, and either assisted with or directly executed all leakage experiments.

References

(1) Otto, M. Phenol-Soluble Modulins. *International Journal of Medical Microbiology*. March 2014, pp 164–169.
(2) Wertheim, H. F. L.; Melles, D. C.; Vos, M. C.; Van Leeuwen, W.; Van Belkum, A.; Verbrugh, H. A.; Nouwen, J. L. The Role of Nasal Carriage in Staphylococcus Aureus Infections. *Lancet Infectious Diseases*. 2005, pp 751–762.
(3) Rimland, D.; Roberson, B. Gastrointestinal Carriage of Methicillin-Resistant Staphylococcus Aureus. *J. Clin. Microbiol.* **1986**, *24* (1), 137–138.
(4) Lowy, F. D. Antimicrobial Resistance: The Example of Staphylococcus Aureus. *Journal of Clinical Investigation*. The American Society for Clinical Investigation 2003, pp 1265–1273.
(5) Kourtis, A. P.; Hatfield, ; Kelly; Baggs, J.; Mu, ; Yi; See, I.; Epson, E.; Nadle, ; Joelle; Kainer, M. A.; Dumyati, G.; Petit, S.; et al. *Morbidity and Mortality Weekly Report Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus Aureus Bloodstream Infections-United States*; 2017.
(6) Mehlin, C.; Headley, C. M.; Klebanoff, S. J. An Inflammatory Polypeptide Complex from Staphylococcus Epidermidis: Isolation and Characterization. *J. Exp. Med.* **1999**, *189* (6), 907–917.
(7) Kretschmer, D.; Gleske, A. K.; Rautenberg, M.; Wang, R.; Köberle, M.; Bohn, E.; Schöneberg, T.; Rabiet, M. J.; Boulay, F.; Klebanoff, S. J.; et al. Human Formyl Peptide Receptor 2 Senses Highly Pathogenic Staphylococcus Aureus. *Cell Host Microbe* **2010**.
(8) Peschel, A.; Otto, M. Phenol-Soluble Modulins and Staphylococcal Infection. *Nature Reviews Microbiology*. 2013, pp 667–673.
(9) Wang, R.; Braughton, K. R.; Kretschmer, D.; Bach, T.-H. L.; Queck, S. Y.; Li, M.; Kennedy, A. D.; Dorward, D. W.; Klebanoff, S. J.; Peschel, A.; et al. Identification of Novel Cytolytic Peptides as Key Virulence Determinants for Community-Associated MRSA. **2007**.
(10) Miller, M. B.; Bassler, B. L. Quorum Sensing in Bacteria. *Annu. Rev. Microbiol.* **2001**, *55* (1), 165–199.
(11) Jamal, M., Tasneem, U., Hussain, T., & Andleeb, and S. A. *Historical Background of Biofilm*; 2015; Vol. 4.
(12) Costerton, J. W.; Stewart, P. S.; Greenberg, E. P. *Bacterial Biofilms: A Common Cause of Persistent Infections*; 1999; Vol. 284.
(13) Periasamy, S.; Joo, H. S.; Duong, A. C.; Bach, T. H. L.; Tan, V. Y.; Chatterjee, S. S.; Cheung, G. Y. C.; Otto, M. How Staphylococcus Aureus Biofilms Develop Their Characteristic Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (4), 1281–1286.
(14) Schwartz, K.; Syed, A. K.; Stephenson, R. E.; Rickard, A. H.; Boles, B. R. Functional Amyloids Composed of Phenol Soluble Modulins Stabilize Staphylococcus Aureus Biofilms. *PLoS Pathog* **2012**, *8* (6).
(15) Tayeb-Fligelman, E.; Tabachnikov, O.; Moshe, A.; Goldshmidt-Tran, O.; Sawaya, M. R.; Coquelle, N.; Colletier, J. P.; Landau, M. The Cytotoxic Staphylococcus Aureus PSMα3

Reveals a Cross-α Amyloid-like Fibril. *Science (80-.).* **2017**, *355* (6327), 831–833.
(16) Schwartz, K.; Ganesan, M.; Payne, D. E.; Solomon, M. J.; Boles, B. R. Extracellular DNA Facilitates the Formation of Functional Amyloids in Staphylococcus Aureus Biofilms. *Mol. Microbiol.* **2016**, *99* (1), 123–134.
(17) Zheng, Y.; Joo, H.-S.; Nair, V.; Le, K. Y.; Otto, M. Do Amyloid Structures Formed by Staphylococcus Aureus Phenol-Soluble Modulins Have a Biological Function? *Int. J. Med. Microbiol.* **2018**, *308* (6), 675–682.
(18) Sónnichsen, F. D.; Van Eyk, J. E.; Hodges, R. S.; Sykes, B. D. *Effect of Trifluoroethanol on Protein Secondary Structure: An NMR and CD Study Using a Synthetic Actin Peptide* 1"*; 1992; Vol. 31.
(19) Towle, K. M.; Lohans, C. T.; Miskolzie, M.; Acedo, J. Z.; Van Belkum, M. J.; Vederas, J. C. Solution Structures of Phenol-Soluble Modulins A1, A3, and B2, Virulence Factors from Staphylococcus Aureus. *Biochemistry* **2016**, *55* (34), 4798–4806.
(20) Malishev, R.; Tayeb-Fligelman, E.; David, S.; Meijler, M. M.; Landau, M.; Jelinek, R. Reciprocal Interactions between Membrane Bilayers and S. Aureus PSMα3 Cross-α Amyloid Fibrils Account for Species-Specific Cytotoxicity. *J. Mol. Biol.* **2018**, *430* (10), 1431–1441.
(21) Laabei, M.; Jamieson, W. D.; Yang, Y.; Van Den Elsen, J.; Jenkins, A. T. A. Investigating the Lytic Activity and Structural Properties of Staphylococcus Aureus Phenol Soluble Modulin (PSM) Peptide Toxins. *Biochim. Biophys. Acta - Biomembr.* **2014**, *1838* (12), 3153–3161.
(22) Alksne, L. E.; Projan, S. J. Bacterial Virulence as a Target for Antimicrobial Chemotherapy. *Current Opinion in Biotechnology*. 2000, pp 625–636.
(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank, 1999–. In *International Tables for Crystallography*; 2006; Vol. 28, pp 675–684.
(24) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.
(25) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958.
(26) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **1999**, *38* (1–2), 236–240.
(27) Trott, O.; Olson, A. J. AutoDock Vina. *J. Comput. Chem.* **2010**, *31*, 445–461.
(28) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (10).

Appendix:

The code that I developed for this project went through numerous iterations, so what is here is the most recent versions of them as of this writing. The code was written for python 2.7, and may fail if used with earlier or later versions.

Some notes on the code: Often times there are print functions that have been commented out. These are there for the purposes of debugging. If something is failing or not behaving properly, I found it useful to check the code's progress with print functions. Working versions of code have most of these print functions commented out.

Many of these programs a main() function that the core of the code calls. This code structure is useful when running parallel code such as the autodock binding code.

ReSA Code:

This code is intended to take a starting structure that has already been energy minimized and equilibrated, then perform alternating molecular dynamics simulations (as defined by the mdp files) on that system. This code is currently set up for pre-defined terms such as mdp files or total run time. There are also segments of code that have been commented out that allow for an interactive input, but were taken out for the purposes of automation.

```
import os
import re
import shutil
import sys
import subprocess
import time

## calls pdb2gmx and runs with the conditions
def pdb (file, conditions):
  command = ['pdb2gmx', '-f', file + '.pdb', '-o', file + '.gro', '-p', file
+ '.top', '-ignh', '-vsite', 'aromatics']
  for condition in conditions:
    command.append(condition)
  error = open('standarderror.txt','w')
  out = open('standardout.txt','w')
  subprocess.call(command,stdout=out,stderr=error)
  error.close()
  out.close()
  error = open('standarderror.txt','r')
  out = open('standardout.txt','r')
  output = open('pdb2gmxoutput.txt','w')
  output.write(error.read())
  output.write(out.read())
  output.close()
  error.close()
  out.close()
  os.remove('standarderror.txt')
  os.remove('standardout.txt')

  return (file + '.gro', file + '.top')

## calls grompp for mdrun prepwork
def grompp(mfile,protein_gro,protein_top,runcount,first,iteration):
  print "GROMPP"
  print mfile
  print protein_gro
  print protein_top
  print runcount
  directory = os.getcwd()
  found = 0
  if mfile in os.listdir(directory):
    found = 1
    runnumber = str(runcount)
    command = ['grompp', '-f', mfile,'-c', protein_gro, '-p',protein_top,'-
o', protein_top[:-4]+'_'+ mfile[:-4]+'_'+ runnumber +'.tpr']
    error = open('standarderror.txt','w')
    out = open('standardout.txt','w')
```

```
    subprocess.call(command,stdout=out,stderr=error)
    error.close()
    out.close()
    error = open('standarderror.txt','r')
    out = open('standardout.txt','r')
    output = open('gromppoutput'+str(iteration)+'.txt','w')
    output.write(error.read())
    output.write(out.read())
    output.close()
    error.close()
    out.close()
    os.remove('standarderror.txt')
    os.remove('standardout.txt')
  else:
    print " the mdp file was not found!"
  if found == 1:
    print "FOUND"
    return ( protein_top[:-4]+'_'+ mfile[:-4]+'_'+ runnumber +'.tpr')

## same as above output code, but for mdrun
def mdrun(tprfile, iteration):
  print ('mdrun ' + str(tprfile) + ' ' + str(iteration))
  mdoutput = tprfile + '_mdrunoutput.txt'
  error = open('standarderror.txt','w')
  out = open('standardout.txt','w')
  mdrun = subprocess.call(['mdrun','-v', '-s', tprfile, '-deffnm', tprfile[:-
4]],stdout=out,stderr=error)
  error.close()
  out.close()
  error = open('standarderror.txt','r')
  out = open('standardout.txt','r')
  output = open('mdrunoutput'+str(iteration)+'.txt','w')
  output.write(error.read())
  output.write(out.read())
  output.close()
  error.close()
  out.close()
  os.remove('standarderror.txt')
  os.remove('standardout.txt')
  return tprfile[:-4] + '.gro'

# Trjcat will append one trajectory to another, allowing us to make
# trajectories of the entire procedure, or of all of the segments from # a
particular set of conditions
def trjcat(firsttraj, nexttraj, time,iteration):
  print 'this is the first trajectory: ' + firsttraj
  print 'this will be added to it: ' + nexttraj
  trjcommand = ['trjcat', '-f', firsttraj, nexttraj, '-o', firsttraj, '-cat']
  print trjcommand
  error = open('standarderror.txt','w')
  out = open('standardout.txt','w')
  proxy =
subprocess.call(trjcommand,stdin=subprocess.PIPE,stdout=out,stderr=error)
  error.close()
  out.close()
  error = open('standarderror.txt','r')
  out = open('standardout.txt','r')
```

```
  output = open('trjcatoutput'+str(iteration)+'.txt','w')
  output.write(error.read())
  output.write(out.read())
  output.close()
  error.close()
  out.close()
  os.remove('standarderror.txt')
  os.remove('standardout.txt')
  print "TRAJECTORY UPDATED"


def main():
## will prompt for number of mdp files, then ask for each one, order matters
here.
#  totaltime = input ('How long (in ps) do you want your simulation to be?')
  totaltime = 100000
  num_mdp = 2
  mdpcount = 0
  mdp = []
  for mdpfilex in ['MDrun_300_PME.mdp', 'MDrun_500_PME.mdp']:
    mdpfile = open(mdpfilex, 'r')
    for line in mdpfile:
      if re.findall('nsteps',line):
        nsteps = int(re.split('\W+',line)[1])
      elif re.findall('dt',line):
        print re.findall('\d+\.\d+',line)
        dt = float(re.findall('\d+\.\d+',line)[0])
    mdp.append((mdpfilex,nsteps*dt))
#  mdp = []
#  num_mdp = input ('How many different sets of conditions do you want to
cycle through?')
#  for mdpcount in range (0, num_mdp):
#    mdpinput1 = raw_input ('What is mdp file number '+ str(mdpcount) + '?')
#    mdpinput1file = open(mdpinput1, 'r')
## reads the mdp file and then pulls the dt and number of steps to determine
how long (in ps) each iteration of that mdrun will generate
## this may need to be updated if the mdp files become more complicated
#    for line in mdpinput1file:
#      if re.findall('nsteps',line):
#        nsteps = int(re.split('\W+',line)[1])
#      elif re.findall('dt',line):
#        print re.findall('\d+\.\d+',line)
#        dt = float(re.findall('\d+\.\d+',line)[0])
#    mdp.append((mdpinput1,nsteps*dt))
#  starting_gro_file = raw_input('What is the starting .gro file?')
  starting_gro_file = 'R_02_settled.gro'
  current_gro=str(starting_gro_file)
  print current_gro
#  protein_top=raw_input('What is the protein topography?')
  protein_top='topol_R_02.top'
  time = 0
  runcount = 1
  first = 0
  firstmdp=[]
  mdptime = []
  wholetrajectory = protein_top[:-4] + '_complete_trajectory.trr'
  trajectoryfile = []
```

```
  for mdpcount in range (0,num_mdp):
    trajectoryfile.append(protein_top[:-4] + '_' + mdp[mdpcount][0][:-4] +
'_concatonated.trr')
    firstmdp.append(0)
    mdptime.append(0)
## this next section will run until the total time is reached or
## exceeded

  while time <= totaltime:
    for mdpcount in range (0, num_mdp):
      print runcount
      tprfile = grompp(mdp[mdpcount][0],current_gro,protein_top,runcount,
first,runcount)
      print tprfile
      current_gro = mdrun(tprfile, runcount)
##  if this is the first runthrough for this simulation it will copy
##  the latest trajectory and rename it to be the beginning of the
##  overall trajectory
      if first == 1:
        trjcat(wholetrajectory, current_gro[:-4]+'.trr', time,runcount)
      else:
        copycommand = ['cp', current_gro[:-4]+'.trr', wholetrajectory]
        subprocess.call(copycommand)
        first = 1
##  if this is the first runthrough for this mdp it will copy the
##  latest trajectory and rename it to be the mdp-specific trajectory
      if firstmdp[mdpcount]==1:
        trjcat(trajectoryfile[mdpcount],current_gro[:-4]+'.trr',
mdptime[mdpcount],runcount)
      else:
        copycommand = ['cp', current_gro[:-4]+'.trr',
trajectoryfile[mdpcount]]
        subprocess.call(copycommand)
        firstmdp[mdpcount] = 1

##    Updates the counts and time parameters, then prints them
      mdptime[mdpcount]=mdptime[mdpcount]+mdp[mdpcount][1]
      time = time + mdp[mdpcount][1]
      mdptime[mdpcount]=mdptime[mdpcount]+mdp[mdpcount][1]
      print 'total time'
      print time
      print 'mdp specific time'
      print mdptime[mdpcount]
      first = 1
      runcount = runcount+1

if __name__ == '__main__':
  main()
```

Automation of converting pdb files to autodock receptors

This code takes the pdb files generated by g_cluster and convert them to audotock's pdbqt format for receptor use. Things that may need to be adjusted in the code are specific pathnames for the autodock utility code such as prepare_receptor4.py.

This program requires the name for the receptors to be based off of, for example "Tau4RD"

```
import os
import re
import shutil
import sys
import subprocess
import tarfile
import time

##"cleans" a pdb file by removing the "M" atoms that occur in some
## frames in pdb=files converted from gromacs trajectories.
## These "M" atoms cause issues in later stages of conversion to pdbqt
## files needed for autodock and appear to be unneccicary as all of
## the atoms from the original pdb appear present.
## possibly these 'M' atoms are artifacts gromacs uses for some MD
##  purposes?
def clean_pdb(pdbfile):
      pdbread = open(pdbfile,'rU')
      cleanpdb = open(pdbfile[:-4]+'_clean'+'.pdb','w')
      for line in pdbread:
            if not "M" in line[12:16]:
                  cleanpdb.write(line)
      cleanpdb.close()
      return(pdbfile[:-4]+'_clean'+'.pdb')

def get_immediate_subdirectories(a_dir):
    return [name for name in os.listdir(a_dir)
            if os.path.isdir(os.path.join(a_dir, name))]

## Utelizes the autodock prepare_receptor4.py script to create
## receptors from pdbfiles.
## if this code is moved off of the nathlab Deepthought machine, or
## the architecture is changed, the path to the autodock tools folder
## will have to be altered accordingly.
def receptor(cleanpdb):
      print cleanpdb
      command =
['pythonsh','/home/dbaggett/Data/Dbaggett/Autodock_Vina/mgltools_x86_64Linux2
_1.5.6/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_receptor4.py','-
r',cleanpdb,'-o',cleanpdb[:-10]+'_receptor.pdbqt']
      print command
      error = open('standarderror.txt','w')
```

```
	out = open('standardout.txt','w')
	subprocess.call(command,stdout=out,stderr=error)
	error.close()
	out.close()
	error = open('standarderror.txt','r')
	out = open('standardout.txt','r')
	output = open('receptormakeoutput.txt','w')
	output.write(error.read())
	output.write(out.read())
	output.close()
	error.close()
	out.close()
	os.remove('standarderror.txt')
	os.remove('standardout.txt')
	return(cleanpdb[:-10]+'_receptor.pdbqt')


def split_10_pdbs(representative_clusters_pdb,output_directory):
	rep_clusters = open(representative_clusters_pdb,'rU')
	rep_clust_lines = rep_clusters.readlines()
	i = 1
	line = 0
	individual_pdbs = []
	while i < 11:
		endpdb = 0
		cluster_receptor_directory = output_directory +'/Cluster_'+str(i)
		if not os.path.isdir(cluster_receptor_directory):
			os.mkdir(cluster_receptor_directory)
		cluster_pdb =
cluster_receptor_directory+'/Cluster_'+str(i)+".pdb"
		cluster_i = open(cluster_pdb,'w')
		while endpdb ==0:
#			print "this is line number:"+str(line)
#			print rep_clust_lines[line]
			cluster_i.write(rep_clust_lines[line])
			if re.search("ENDMDL",rep_clust_lines[line]):
				endpdb = 1
			line = line + 1

	individual_pdbs.append((cluster_pdb,cluster_receptor_directory,str(i)))
		i = i+1
		cluster_i.close()

	return(individual_pdbs)


def main():
	basename = sys.argv[1]
	basefolder = os.getcwd()
```

```
	receptor_directory = basefolder + "/"+basename+"_receptors"
	if not os.path.isdir(receptor_directory):
		os.mkdir(receptor_directory)
	segment_directories = []
	for directory in get_immediate_subdirectories(basefolder):
		if re.search("Segment",directory):
			segment_directories.append(directory)
	for segment_directory in segment_directories:
		segment_receptor_directory = receptor_directory + "/" +
segment_directory
		original_segment_directory = basefolder + '/'+segment_directory
		if not re.search(basename+"_receptors",segment_directory):
			segment_files = os.listdir(original_segment_directory)
			if not os.path.isdir(segment_receptor_directory):
				os.mkdir(segment_receptor_directory)
			for segment_file in segment_files:
				if re.search("clusters.pdb$",segment_file):
					all_clusters_pdb =
original_segment_directory+'/'+segment_file
					individual_pdbs =
split_10_pdbs(all_clusters_pdb,segment_receptor_directory)
					for pdb in individual_pdbs:
						for segment_filex in segment_files:
							if
re.search('pdb\d+'+pdb[2]+'.pdb',segment_filex):

	shutil.copyfile(original_segment_directory+'/'+segment_filex,pdb[1]+"/C
luster_"+pdb[2]+"_allframes.pdb")
					cluster_clean =clean_pdb(pdb[0])
					cluster_receptor =
receptor(cluster_clean)

if __name__ == '__main__':
	main()
```

Converting sdf files to pdbqt

This code takes sdf files obtained either from pubchem or generated with obabel's –gen3D command, and then converts them to autodock's format. The arguments after calling the code is a list of sdf files to convert, and this does take entries using the glob utility (example: *.sdf). As with the other code that converts to pdbqt, it requires prepare_ligand4.py to be installed and its location specified.

```
import os
import re
import shutil
import sys
import subprocess
import time

def convert_to_pdb(sdf_file):
      command = ['babel','-i','sdf',sdf_file, '-
o','pdb',re.split('[\./]',sdf_file)[-2]+'.pdb','-m']
      print command
      error = open('standarderror.txt','w')
      out = open('standardout.txt','w')
      subprocess.call(command,stdout=out,stderr=error)
      error.close()
      out.close()
      error = open('standarderror.txt','r')
      out = open('standardout.txt','r')
      output = open(re.split('[\./]',sdf_file)[-2]+'babeloutput.txt','w')
      output.write(error.read())
      output.write(out.read())
      output.close()
      error.close()
      out.close()
      os.remove('standarderror.txt')
      os.remove('standardout.txt')
      return re.split('[\./]',sdf_file)[-2]

def AutoDock_Conv(pdb_file):
      ligand_name = pdb_file
      read_pdb = open(pdb_file,'rU')
      for line in read_pdb:
            if re.search('COMPND',line):
                  ligand_name = re.findall('\S+',line)[-1]
      print ligand_name
      command = ['pythonsh','prepare_ligand4.py','-l',pdb_file, '-F','-
o',ligand_name+'.pdbqt']
      print command
      error = open('standarderror.txt','w')
      out = open('standardout.txt','w')
      subprocess.call(command,stdout=out,stderr=error)
```

```
	error.close()
	out.close()
	error = open('standarderror.txt','r')
	out = open('standardout.txt','r')
	output = open(re.split('[\./]',pdb_file)[-
2]+'prepare_ligand4_output.txt','w')
	output.write(error.read())
	output.write(out.read())
	output.close()
	error.close()
	out.close()
	os.remove('standarderror.txt')
	os.remove('standardout.txt')

def main():
	sdffiles = sys.argv[1:]
	for sdffile in sdffiles:
		pdbfiles = []
		pdb_name = convert_to_pdb(sdffile)
		for filex in os.listdir(os.getcwd()):
			if re.search(pdb_name,filex) and re.search('pdb',filex):
				pdbfiles.append(filex)
		for pdbfile in pdbfiles:
			AutoDock_Conv(pdbfile)

if __name__ == '__main__':
	main()
```

Docking ligands to receptors:

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
#
#  160701_AD_Bind.py
#
#  Copyright 2016 DavidB Nathlab <dbaggett@deepthought>
#
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
#  the Free Software Foundation; either version 2 of the License, or
#  (at your option) any later version.
#
#  This program is distributed in the hope that it will be useful,
#  but WITHOUT ANY WARRANTY; without even the implied warranty of
#  MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#  GNU General Public License for more details.
#
#  You should have received a copy of the GNU General Public License
#  along with this program; if not, write to the Free Software
#  Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston,
#  MA 02110-1301, USA.
#
#  This code is called via: python 181204_AD_Bind.py  <ligand
# directory> <Structure directory> <number of processors>
# <results_directory>
#  It requries ligands to be pre-converted to the pdbqt format that
# Autodock Vina uses
#  For each structure being examined make sure that in the "receptor"
# directory there is a structure file converted using
# prepare_receptor.py, in the pdbqt format

import os
import re
import shutil
import sys
import subprocess
import time
import multiprocessing as mp

def get_immediate_subdirectories(a_dir):
    return [name for name in os.listdir(a_dir)
            if os.path.isdir(os.path.join(a_dir, name))]

def makebox(receptorname, lowx, lowy, lowz, highx, highy, highz):
      ##This function makes a box using the parameters that determine
acceptable binding areas in the autodock program.
      ##While not directly useful, it's beneficial to have this when
examining binding to ensure that the search area was appropriate
```

```
  box = open(receptorname+'_box.pdb','w')
  box.write('HETATM    1  N1  UNL     1      '+ str(lowx)+'  ' + str(lowy)+'
'+str(lowz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(highx)+'  ' + str(lowy)+'
'+str(lowz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(lowx)+'  ' + str(highy)+'
'+str(lowz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(highx)+'  ' +
str(highy)+'  '+str(lowz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(lowx)+'  ' + str(lowy)+'
'+str(highz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(highx)+'  ' + str(lowy)+'
'+str(highz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(lowx)+'  ' + str(highy)+'
'+str(highz) + '  1.00 20.00           N \n')
  box.write('HETATM    1  N1  UNL     1      '+ str(highx)+'  ' +
str(highy)+'  '+str(highz) + '  1.00 20.00           N \n')
  box.write('TER       9      UNL     1 \n')
  box.write('CONECT    1    2    3    5 \n')
  box.write('CONECT    2    1    4    6 \n')
  box.write('CONECT    3    1    4    7 \n')
  box.write('CONECT    4    2    3    8 \n')
  box.write('CONECT    5    1    6    7 \n')
  box.write('CONECT    6    2    5    8 \n')
  box.write('CONECT    7    3    5    8 \n')
  box.write('CONECT    8    4    6    7 \n')
  box.write('END \n')
  box.close()
  return(receptorname+'_box.pdb')

def bind_vina(receptor,ligand,result_directory):
## This code calls the various vina parameters together and executes a
## single instance of docking.
## It first reads through the receptor file to make the search box,
## extending out from the furthest extents of the receptor.
      receptorname = re.split('[/\.]',receptor)[-2]
      ligandname = re.split('[/\.]',ligand)[-2]
      low_x = 0
      low_y = 0
      low_z = 0
      high_x = 0
      high_y = 0
      high_z = 0
      pdbread = open(receptor,'rU')
      lowx = lowy = lowz = highx = highy = highz = None
      for line in pdbread:
            if re.match('ATOM',line):
                  x_value = float(re.split('\s+',line)[-8])
                  y_value = float(re.split('\s+',line)[-7])
```

```
                  z_value = float(re.split('\s+',line)[-6])
                  if low_x == low_y == low_z == high_x == high_y == high_z ==
None:
                        low_x = high_x = x_value
                        low_y = high_y = y_value
                        low_z = high_z = z_value
                  else:
                        if x_value < low_x:
                              low_x = x_value
                        if x_value > high_x:
                              high_x = x_value
                        if y_value < low_y:
                              low_y = y_value
                        if y_value > high_y:
                              high_y = y_value
                        if z_value < low_z:
                              low_z = z_value
                        if z_value > high_z:
                              high_z = z_value
      print [low_x, low_y, low_z,high_x, high_y, high_z]
      low_x = low_x - 3
      low_y = low_y - 3
      low_z = low_z - 3
      high_x = high_x + 3
      high_y = high_y + 3
      high_z = high_z + 3
      print [low_x, low_y, low_z,high_x, high_y, high_z]
      pdbread.close()
      makebox(result_directory+'/'+receptorname, low_x, low_y, low_z, high_x,
high_y, high_z)
      mid_x = (high_x + low_x)/2
      mid_y = (high_y + low_y)/2
      mid_z = (high_z + low_z)/2
      range_x = high_x - low_x
      range_y = high_y - low_y
      range_z = high_z - low_z
      posefile = receptorname + '_'+ligandname + '.pdbqt'
      logfile = receptorname + '_'+ligandname+'.log'
      command = ['vina','--cpu' , '1', '--receptor',receptor,'--
ligand',ligand,'--center_x',str(mid_x),'--center_y',str(mid_y),'--
center_z',str(mid_z),'--size_x',str(range_x+2),'--size_y',str(range_y+2),'--
size_z',str(range_z+2),'--out',result_directory+'/'+posefile,'--
log',result_directory+'/'+logfile]
      print command
      error = open(result_directory+'/'+'standarderror.txt','w')
      out = open(result_directory+'/'+'standardout.txt','w')
      subprocess.call(command,stdout=out,stderr=error)
      error.close()
      out.close()
```

```
      error = open(result_directory+'/'+'standarderror.txt','r')
      out = open(result_directory+'/'+'standardout.txt','r')
      output =
open(result_directory+'/'+receptorname+ligandname+'vinaoutput.txt','w')
      output.write(error.read())
      output.write(out.read())
      output.close()
      error.close()
      out.close()
      os.remove(result_directory+'/'+'standarderror.txt')
      os.remove(result_directory+'/'+'standardout.txt')
      return ((posefile,logfile))

def binding_spree(result_directory,structure_directory,ligand_file):
## This function recursively moves through the folders to hunt for
## "receptor.pdbqt" files and then dock ligands against them.
## to designate a file as a "receptor" file, simply end the name of
## the file with "receptor.pdbqt"
      print "looking in " +structure_directory+" for receptors"
      print "results_directory:"+result_directory
      receptor_files = []
      for filex in next(os.walk(structure_directory))[2]:
            if re.search('pdbqt$',filex):
                  receptor_files.append(structure_directory+'/'+filex)
      if len(receptor_files)>0:
            for receptor_file in receptor_files:
                  bind_vina(receptor_file,ligand_file,result_directory)
      print "folders in the " + structure_directory +" directory :"
      print  next(os.walk(structure_directory,))[1]
      for folderx in next(os.walk(structure_directory,))[1]:
            print folderx
            if not folderx in next(os.walk(result_directory))[1]:
                  os.mkdir(result_directory+'/'+folderx)

      binding_spree(result_directory+'/'+folderx,os.path.abspath(structure_di
rectory+'/'+folderx),ligand_file)



def ligand_bind(tuplex):
      (ligand_file,basefolder,structure_directory,progress_file,results_folde
r)=tuplex
      print "results folder: " + results_folder
      ligand_name = re.split('[\/\.]',ligand_file)[-2]
      ligand_results_folder = results_folder + '/'+ligand_name +
'_binding_results'
      if not ligand_name + '_binding_results$' in os.listdir(results_folder):
            os.mkdir(ligand_results_folder)
      binding_spree(ligand_results_folder,structure_directory,ligand_file)
```

```
	progress_report = open(progress_file,'a')
	progress_report.write(ligand_name + ' complete \n')
	progress_report.close()
	os.chdir(basefolder)


def main():
	basedirectory = os.getcwd()
	ligands = []
	structure_directory= os.path.abspath(sys.argv[2])
	np = int(sys.argv[3])
	result_name = sys.argv[4]+"_results"
	results_directory = basedirectory+'/'+result_name
	if not result_name in next(os.walk(basedirectory))[1]:
		os.mkdir(basedirectory+'/'+result_name)
	if not 'progress.log' in os.listdir(results_directory):
		progress_report = open(results_directory + '/progress.log','w')
		progress_report.close()
		completed_ligands = []
	else:
		completed_ligands = []
		progress_read = open(results_directory + '/progress.log','rU')
		for line in progress_read:
			if re.search('complete',line):
				completed_ligands.append(re.findall('\S+',line)[-2])
		progress_read.close()
## Identifies the ligands that haven't been run already, and places
## them in a list of tuples.  The tuples are necessary because the
## parallel processing utility only takes one input.
	for filex in next(os.walk(sys.argv[1]))[2]:
		if re.search('\.pdbqt',filex):
			if re.split('[\/\.]',filex)[-2] not in completed_ligands:

	ligands.append((os.path.abspath(sys.argv[1]+'/'+filex),basedirectory,st
ructure_directory,basedirectory + '/progress.log',results_directory))
## For the following 4 lines, uncomment either the first two or the
## second two. the pool lines are for parallel processing and are much
## faster. the for-loop is the same code implemented in a loop and is
## significantly easier to debug
	pool = mp.Pool(processes=np)
	pool.map(ligand_bind,ligands)
#	for ligand in ligands:
#		ligand_bind(ligand)
	return 0

if __name__ == '__main__':
	main()
```

Score Compilation

The next segment of code is for compiling all of the results of the autodock binding procedure. This is useful in data analysis, and is also one of the inputs (along with fingerprints) for the PLSR prediction code.

```
## Call with python *score_compile.py <receptor directory used for
## docking> <directory in which docking results lie> <directory where
## to put the analysis>
import os
import re
import shutil
import sys
import subprocess
import time
import math

def pull_results(scorefilex):
      scorefile = open(scorefilex,'rU')
      score = "N/A"
      time = "N/A"
      receptor = "N/A"
      lines = []
      for line in scorefile:
            lines.append(line)
#     print lines
      for line in lines:
            if re.findall('\S+',line) != [] and re.findall('\S+',line)[0]
=='1':
#                 print line
                  score = re.findall('\S+',line)[1]
#                 print score
      return(score)

def tailnumbersort(folder_name):
      if re.findall('\d+',folder_name)==[]:
            return 0
      else:
            return int(re.findall('\d+',folder_name)[-1])

def find_receptors(call_directory,base_directory,summaryfile,receptors):
#     print base_directory
#     print sorted(next(os.walk(base_directory))[1],key=tailnumbersort)
      for folderx in
sorted(next(os.walk(base_directory))[1],key=tailnumbersort):
            receptors =
find_receptors(call_directory,base_directory+'/'+folderx,summaryfile,receptor
s)
      for filex in next(os.walk(base_directory))[2]:
            if re.search('receptor.pdbqt',filex):
```

```
#                 print os.path.abspath(base_directory +'/'+filex)
                  receptor =
re.split('\.pdbqt',re.split(call_directory,base_directory +'/'+filex)[-1])[0]
                  receptors.append(receptor)
                  summaryfile.write(","+receptor)
      return receptors

def main():
      basepath = os.getcwd()
      receptor_directory = os.path.abspath(sys.argv[1])
      results_directory = os.path.abspath(sys.argv[2])
      output_directory = sys.argv[3]
      if not output_directory in next(os.walk(basepath))[1]:
            os.mkdir(basepath+'/'+output_directory)
      output_directory = os.path.abspath(output_directory)
      summaryfile = open(output_directory+'/'+'Compound_scores.txt','w')
      summaryfile.write('compound')
      receptors = []
      receptors =
find_receptors(basepath,receptor_directory,summaryfile,receptors)
      #print receptors
      summaryfile.write('\n')
      for folderx in next(os.walk(results_directory))[1]:
            #print folderx
            compound_directory = results_directory+'/'+folderx
            #print compound_directory
            writeline = ''
            compound_name = re.split('_',folderx)[0]
            print compound_name
            writeline = writeline + compound_name
            for receptor in receptors:
                  score = "N/A"
                  scorefile = compound_directory
                  for directory in re.split('/',receptor)[2:]:
                        scorefile=scorefile+'/'+directory
                  scorefile=scorefile+'_'+compound_name+'.log'
                  score = pull_results(scorefile)
#                                 print "receptor:"+receptor
                  writeline = writeline + ','+score
#                 print writeline
            summaryfile.write(writeline + '\n')
      summaryfile.close()


if __name__ == '__main__':
      main()
```

Fingerprint Prediction

This code utilizes the sklearn module. It takes multiple arguments, detailed at the beginning of the main() function.

```
import math
import scipy
import numpy as np
import sys
import os
from sklearn.cross_decomposition import PLSRegression as PLSR
import re

def overall_affinity_scoring(list_of_scores):
      if len(list_of_scores)>0:
            score_affinities = []
            for score in list_of_scores:
#                 print score
                  if score == 'N/A':
                        score = 0
                  score = float(score)
                  score_affinity = math.exp(score)
                  score_affinities.append(score_affinity)
            inverse_apparent_affinity = 0
            for score_affinity in score_affinities:
                  inverse_apparent_affinity = inverse_apparent_affinity +
score_affinity ** -1
            apparent_affinity = inverse_apparent_affinity ** -1
            overall_affinity_score = math.log(apparent_affinity)
      else:
            overall_affinity_score = 'N/A'
      return(overall_affinity_score)

def hex2bin(hexcode):
      binary = ''
      for s in hexcode:
            bytes = bin(int(s,16))[2:].zfill(4)
            binary = binary + bytes
      return binary

## Fingerprints are stored in compressed files and need to be
## decompressed to interpret and use
def convert_to_binary(hexfile):
      print "CONVERTING THIS FILE:" + hexfile
      fingerprint_file = hexfile
      binary_file = re.split('\.',re.split('/',fingerprint_file)[-
1])[0]+'_Binary.'+re.split('\.',fingerprint_file)[-1]
      binary_out = open(binary_file,'w')
      fingerprints = open(fingerprint_file,'rU')
      fingerlines = []
```

```
	num_bits = 0
	for line in fingerprints:
		fingerlines.append(line)
	for line in fingerlines:
		if line[0] =='#':
			if re.search('num_bits',line):
				bits=int(re.findall('\d+',line)[-1])
				num_bits = bits + 4 - bits%4
				#print num_bits
			binary_out.write(line)
		else:
			hexcode = re.findall('\S+',line)[0]
			title = re.findall('\S+',line)[1]
			bincode = hex2bin(hexcode)
			binary_out.write(title+ '\t' + bincode  + '\n')
	fingerprints.close()
	binary_out.close()
	return binary_file

## moves the fingerprint data to a dict for easier use

def fingerprint_to_dict(fingerprint_file):
	opened_file = open(fingerprint_file,'rU')
	fingerprint_dict = {}
	for line in opened_file:
		if line[0] != '#':
			fingerprint_list = []
			ID = re.findall('\S+',line)[0]
			fingerprint = re.findall('\d+',line)[-1]
			for s in fingerprint:
				fingerprint_list.append(float(s))
			fingerprint_dict[ID]=fingerprint_list
	opened_file.close()
#	print fingerprint_dict
	return fingerprint_dict

def pull_results(filex):
	results_file = open(filex,'rU')
	content = results_file.readlines()
	results_dict = {}
	for line in content[1:]:
		if not re.search('N/A',line):
			ID = re.split(',',line)[0]
			data_list = []
			for data in re.split(',',line)[1:]:
				data_list.append(float(data))
			results_dict[ID]=data_list
#	print results_dict
	return results_dict
```

```
def main():
## Program takes three post call arguments, a training fingerprint
## file (written in hexcode), followed by the results file containing
## the selected compounds, lastly a test set of fingerprints again in
## hexcode.  The results file should lead with compound ID
      training_fingerprint_file = convert_to_binary(sys.argv[1])
      fingerprint_name=re.split('\.',re.split('/',sys.argv[1])[-1])[0]
      test_fingerprint_file = convert_to_binary(sys.argv[3])
      print test_fingerprint_file
      test_name = re.split('\.',re.split('/',sys.argv[3])[-1])[0]
#     print fingerprint_file
      training_fingerprint_dict =
fingerprint_to_dict(training_fingerprint_file)
      test_fingerprint_dict = fingerprint_to_dict(test_fingerprint_file)
      results_file = sys.argv[2]
      dependent_list = []
      independent_list = []
      results_dict = pull_results(results_file)
      x_values = open(fingerprint_name+"_X_matrix.csv",'w')
      y_values = open(fingerprint_name+"_Y_matrix.csv",'w')
      rownames = []
      for key in results_dict.keys():
            if key in training_fingerprint_dict.keys():
                  y_values.write(str(key))
                  x_values.write(str(key))
                  rownames.append(str(key))
                  independent_list.append(training_fingerprint_dict[key])
#                 print fingerprint_dict[key]
                  for item in training_fingerprint_dict[key]:
                        x_values.write(','+str(item))
                  dependent_list.append(results_dict[key])
                  for item in results_dict[key]:
                        y_values.write(','+str(item))
                  x_values.write('\n')
                  y_values.write('\n')
      x_values.close()
      y_values.close()
      test_list = []
      test_names = []
      for key in test_fingerprint_dict.keys():
            test_names.append(str(key))
            test_list.append(test_fingerprint_dict[key])
      print len(test_list)
      print len(test_names)
      independent_matrix = np.matrix(np.array(independent_list))
      dependent_matrix = np.matrix(np.array(dependent_list))
      test_matrix = np.matrix(np.array(test_list))
      pls2 = PLSR(n_components=2)
```

```
	pls2.fit(independent_matrix,dependent_matrix)
	Y_pred = pls2.predict(independent_matrix)
	test_pred = pls2.predict(test_matrix)
	estimate = open(fingerprint_name+"_estimate.csv",'w')
	error = open(fingerprint_name+"_errors.csv",'w')
	bestscores = open(fingerprint_name+"_input_PLSR_bestscores.csv",'w')
	OAscores = open(fingerprint_name+"_PLSR_Overall_Affinity.csv",'w')
	test_estimate =
open('PLSR_from_'+fingerprint_name+'_on_'+test_name+"_estimate.csv",'w')
	test_error =
open('PLSR_from_'+fingerprint_name+'_on_'+test_name+"_errors.csv",'w')
	test_bestscores =
open('PLSR_from_'+fingerprint_name+'_on_'+test_name+"_PLSR_bestscores.csv",'w
')
	test_OAscores =
open('PLSR_from_'+fingerprint_name+'_on_'+test_name+"_PLSR_Overall_Affinity.c
sv",'w')
	i = 0
	best = 'null'
	for row in test_pred:
		best = 'null'
		print test_names[i]
		j = 1
		segmentbest = 'null'
		test_estimate.write(test_names[i])
		bestscores.write(test_names[i])
		test_bestscores.write(test_names[i])
		test_OAscores.write(test_names[i])
		segmentscores= []
		for entry in row:
			if best == 'null':
				best = float(entry)
			elif best > float(entry):
				best = float(entry)
#				print entry
			if segmentbest == 'null':
				segmentbest = float(entry)
			elif segmentbest > float(entry):
				segmentbest = float(entry)
			test_estimate.write(','+str(entry))
			if j < 10:
				j = j+1
				segmentscores.append(segmentbest)
				segmentbest = 'null'
			else:
				j = j+1
		i =i+1
		test_estimate.write('\n')
		test_bestscores.write(','+str(best)+'\n')
```

```
	test_OAscores.write(','+str(overall_affinity_scoring(segmentscores))+'\
n')
	errors = np.matrix(np.matrix(Y_pred) - np.matrix(dependent_matrix))
	i = 0
	for row in errors:
		test_error.write(rownames[i])
		for entry in row:
			for item in entry:
				test_error.write(','+str(item))
		i = i+1
		test_error.write('\n')
	test_estimate.close()
	test_error.close()
	test_bestscores.close()
	test_OAscores.close()
	i = 0
	for row in Y_pred:
		best = 'null'
		print rownames[i]
		j = 1
		segmentbest = 'null'
		estimate.write(rownames[i]+',')
		bestscores.write(rownames[i] + ',')
		OAscores.write(rownames[i]+',')
		segmentscores= []
		for entry in row:
			if best == 'null':
				best = float(entry)
			elif best > float(entry):
				best = float(entry)
#				print entry
			if segmentbest == 'null':
				segmentbest = float(entry)
			elif segmentbest > float(entry):
				segmentbest = float(entry)
			estimate.write(str(entry)+',')
			if j < 10:
				j = j+1
				segmentscores.append(segmentbest)
				segmentbest = 'null'
			else:
				j = j+1
		i =i+1
		estimate.write('\n')
		bestscores.write(str(best)+'\n')
		OAscores.write(str(overall_affinity_scoring(segmentscores))+'\n')
	errors = np.matrix(np.matrix(Y_pred) - np.matrix(dependent_matrix))
	i = 0
```

```
	for row in errors:
		error.write(rownames[i]+',')
		for entry in row:
			for item in entry:
				error.write(str(item)+',')
		i = i+1
		error.write('\n')
	estimate.close()
	error.close()


if __name__ == '__main__':
	main()
```

Afterword

This thesis project has spanned 6-years and required more hours than I care to estimate. When I joined the Nath lab I was optimistic about using relatively new computational techniques to approach challenges in the relatively new field of disordered proteins. At the time, I do not think I fully appreciated the hurdles I would have to overcome to incorporate, understand, and utilize so many new ideas. While my ignorance often led to struggles, it also was at times what led me to make innovations in my projects. Without an established background that told me how things are done in the field, I was unencumbered and explored new ideas instead of relying on established methods.

This isn't to say that ignorance of what's going on in the field is something that should be aspired for. The development of the enhanced sampling technique, and the method of identifying locally persistent structure were both inspired by established techniques but differed in execution. Without an understanding of what other people had attempted, I never would have been able to create my own approaches. Additionally, while innovative thinking helps create new solutions, it is also capable of creating at least as many challenges, and established methodologies are often the most reliable ways to get things done.

I'm interested to see where this research advances. There are numerous areas to investigate and improve upon. Possible ways to adjust and improve the ligand discovery method include adjusting the molecular dynamics, both the base algorithms, as well as improving the ReSA protocol. In our computational work, we used the best force-fields and molecular models available to us, but even in the short amount of time since then there have been notable improvements in both. Additionally, optimization of the ReSA protocol could provide an even more efficient method of generating conformational ensembles of disordered proteins. By adjusting the amount of time in different conditions, or by adjusting the conditions used, it's possible to generate a set of conditions that explores the conformational library of a disordered protein more efficiently that what was done in this thesis. Furthermore, there are more advanced machine learning algorithms beyond PLSR that can likely identify more complex patterns that dictate docking, and could be used to not only expedite the ligand search, but also provide a better understanding of favorable or unfavorable chemical features for ligands of a specific target.

In addition to improvements on methodology, I'm interested in further applications of it. The framework outlined in this work is easily translatable to other systems, which may be able to utilize methods of creating more refined structural ensembles. By incorporating SAXS, NMR, or FRET data into the molecular dynamics force-fields, the structural ensembles generated by ReSA can be guided to be more representative of in vitro data.

I believe there is still more work that can be done with the compounds identified. At some point I had hoped to do some *in vivo* studies in simple organism models such as *C. elegans* or use a neuroblastoma cell line to observe how modulation of tau aggregation with compound 1 identified in chapter 2 could affect disease outcomes. Unfortunately circumstances did not allow for those studies, but I am very interested to see what happens when they do. Separate from *in vivo* studies, further biophysical characterization detailing the interaction between active compounds and tau isoforms would be very insightful. For instance, any method of elucidation of the binding positions of these compounds in relation to tau would be useful in confirming if the computational methods were accurate in those predictions.