Spring 4-29-2022

Computational Investigations into Binding Dynamics of Tau Protein Antibodies: Using Machine Learning and Biophysical Models to Build a Better Reality

Katherine Lee
khlee467@gmail.com

Computational Investigations into Binding Dynamics of Tau Protein Antibodies: Using Machine Learning and Biophysical Models to Build a Better Reality

Katherine Lee

University of Connecticut

University Scholar Program

Structural Biology/Biophysics Thesis

Honors/Thesis Advisor: Dr. Eric May

University Scholar Advisors: Dr. Eric May, Dr. Yongku Cho, and Dr. Adam Zweifach

May 2022

Table of Contents

Abstract

Misregulation of post-translational modifications of microtubule-associated protein tau is implicated in several neurodegenerative diseases including Alzheimer's disease. Hyperphosphorylation of tau promotes aggregation of tau monomers into filaments which are common in tau-associated pathologies. Therefore, tau is a promising target for therapeutics and diagnostics. Recently, high-affinity, high-specificity single-chain variable fragment (scFv) antibodies against phosphorylated tau (pThr-231) were generated and the most promising variant (scFv 3.24) displayed 20-fold increased binding affinity to pThr-231 tau compared to the wild-type scFv. The scFv 3.24 variant contained five point mutations, and intriguingly none were in the tau binding site. The increased affinity was hypothesized to occur due to allosteric communication between the framework (distal) region and the binding site. To examine the mutational impact on the structure and dynamics of the scFv-pTau systems, multi-microsecond all-atom molecular dynamics simulations were conducted for four systems – the wild-type antibody and the 3.24 mutant, with and without tau. Correlation of All Rotameric and Dynamical States (CARDS) software was used to quantify allostery in terms of mutual information (MI), or the dependence between two variables. The mutant exhibited much higher total MI than the wild-type as well as MI relative to target sites of interest, including the four residues that bind directly to the phosphate group on tau. A recently developed machine learning method (DiffNets), which is a supervised autoencoder with a classification task, was used to distinguish the relevant motions separating wildtype from 3.24 mutant. Results showed long-range expansion within the mutant stemming from mutation Ile61. Recent work has been aimed towards quantifying optimal collective variables for discriminating wild-type from mutant ensembles for use in free energy calculations.

Introduction

Tau-Antibody binding

Neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis are rising in prevalence due to increased life expectancy as well as environmental factors (1). The potential culprit of the pathophysiology of these conditions is post-translational modifications (PTMs). PTMs are changes to a synthesized amino acid chain that form a mature protein and often have a role in protein regulation and, in certain cases, the inception of disease. Hyperphosphorylation, or an abnormal or excessive addition of phosphate groups to a protein, is one such PTM that is linked to the development of all the conditions listed above (1-4). In particular, hyperphosphorylation of tau protein at serine, threonine, and tyrosine residues is a hallmark of Alzheimer's disease (AD) and is associated with the misfolding of tau into neurofibrillary tangles (4). There is no cure for Alzheimer's disease, and current diagnostic methods leave much to be desired. While antibodies that bind to hyperphosphorylated tau protein have been used commercially as a diagnostic technique and show potential for drug development, their affinity and specificity to the desired region are too low. Creating antibodies that bind to a specific PTM is difficult due to the dynamic nature of proteins and small regions of modification compared to the rest of the molecule (5). In addition, current antibodies designed to detect PTMs often bind to protein regions without the desired hyperphosphorylation, which can lead to false diagnoses. Unfortunately, the specificity of many commercially available antibodies has not been experimentally confirmed (5). While wet lab techniques such as random mutagenesis followed by directed evolution have been used in the past to develop antibodies with increased affinity and specificity for hyperphosphorylated tau protein, a computational approach holds great promise in the field and will arguably play a major future role in drug

development. Conducting simulations of molecular systems allows one to analyze protein dynamics and gain a conceptual understanding of the biological actors on the stage of human life. With a better knowledge of not only the amino acids in a protein but also how they communicate with each other and other proteins to perform processes such as protein-ligand binding, one can alter these structures to engineer a desired response or improve upon an existing one.

Dr. Yongku Cho of UConn's Chemical and Biomolecular Engineering department generated an antibody with more than 20 times higher affinity to hyperphosphorylated tau and no detectable increase in nonspecific binding (5). The antibody consists of 231 amino acids: 104 residues in the light chain and 127 in the heavy chain (6). There are three complementary determining regions (CDRs), or hypervariable regions in an antibody that determine binding specificity, in the light chain (CDR-L1, CDR-L2, and CDR-L3) and three CDRs in the heavy chain (CDR-H1, CDR-H2, and CDR-H3). As revealed by the crystal structure determined by Shih et al, epitope recognition occurs via eight hydrogen bonds, three salt bridges, and six hydrophobic interactions (6). Contact with the phosphate group is dominated by CDR-H2, with support of nearby amino acids accomplished by CDR-H3. CDR-H1, CDR-L1, CDR-L2, and CDR-L3 assist in stabilization of the phosphoepitope (6). Refer to Figure 1 for a more detailed description of the interactions between the CDRs and tau epitope.

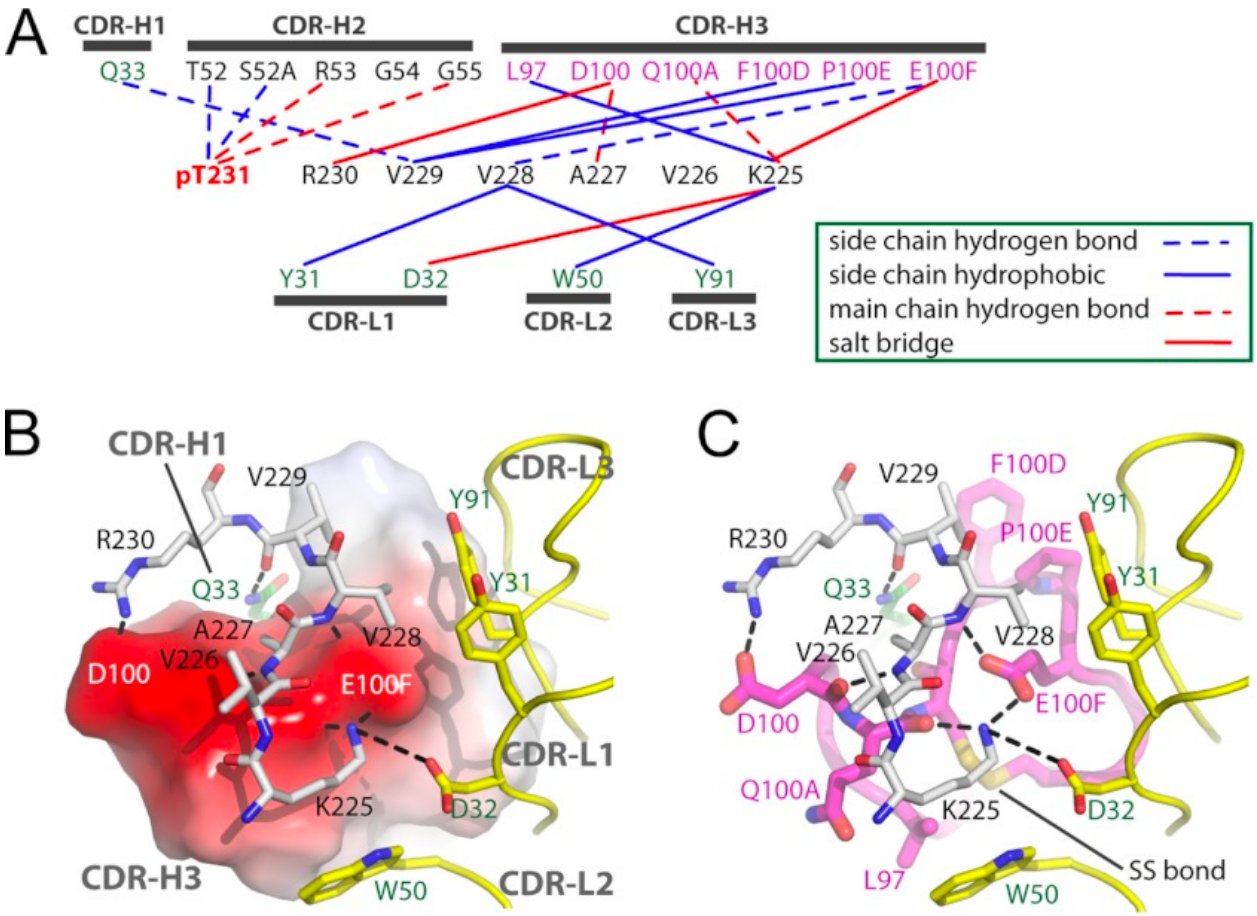
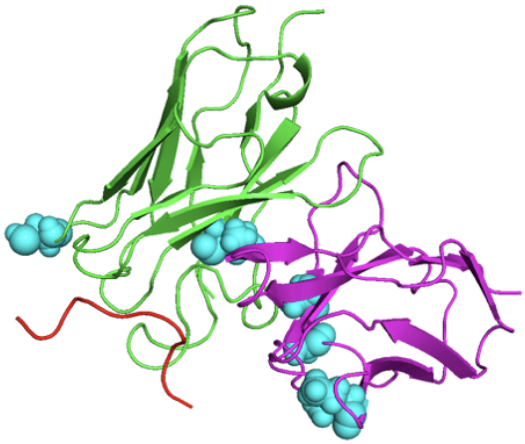


Figure 1: ***Antibody-phosphoepitope interactions.*** *A) Schematic representation of the CDRs and their interactions with the phosphate group on tau and nearby residues. B) and C) reveal the interaction details between the tau epitope and antibody fragment. In B), CDR-H3 is represented using an electrostatic surface model, with negatively charged areas in red and positively charged areas in blue. In C), CDR-H3 is colored magenta and represented as sticks, with magenta being carbon, blue being nitrogen, and red being oxygen. The CDRs in the light chain are displayed in yellow and side chains are represented as sticks, with yellow being carbon and red being oxygen. The tau epitope is represented as gray sticks, with gray being carbon, nitrogen being blue, and oxygen being red. Hydrogen bonds are displayed with dashed lines, and the disulfide bond within CDR-H3 is labeled SS bond. Figure taken from (6).*

The antibody includes three mutations in the light chain (threonine to alanine at position 41, asparagine to glycine at position 52, and threonine to isoleucine at position 61) and two mutations in the heavy chain (glycine to cysteine at position 75 and alanine to valine at position 97). The mutation locations are depicted in Figure 2. Interestingly, the mutations which led to the most effective antibodies were not in the regions that directly bind to tau protein (5).



Mutations	
THR --> ALA	AA 41, light chain
ASP --> GLY.	AA 52, light chain
THR --> ILE	AA 61, light chain
GLY --> CYS.	AA 75, heavy chain
ALA --> VAL	AA 97, heavy chain

Figure 2: ***Structure of antibody mutant (mutant 3.24) that is investigated in this study.*** *Here the light chain is in purple, the heavy chain is in green, the tau epitope is in red, and the mutations are highlighted in cyan. Image rendered in PyMOL.*

Without an easily delineated biomolecular mechanism evident that differentiates the binding of the mutant from the binding of the wildtype to hyperphosphorylated tau protein, it is difficult to predict functionality from structure or propose new antibody structures with improved binding affinity/specificity. We have hypothesized that the improved affinity of these antibodies can be explained in terms of allostery, or long-distance communication between the amino acids in the antibody. This was tested through molecular dynamics simulations of the tau-antibody system for both the wildtype and mutant in order to differentiate the two binding mechanisms and discover potential avenues for diagnostics and treatment.

Computational Approach

The future of medicine will rely heavily on the use of computer-generated biophysical models to interpret and predict molecular processes. An experimental approach allows one to see how life behaves in a natural environment, however, due to the complexity of those systems and environments, experimental methods may lack insight into the underlying mechanisms behind how molecules interact and why these interactions occur. A computational approach allows one

to put a magnifying glass to the biophysical principles dictating how molecules move and derive theories for how this dictates actions at a larger scale. In the seminal Feynman Lectures on Physics (1963), Nobel prize winning theoretical physicist Richard Feynman stated that "everything that living things do can be understood in terms of the jiggling and wiggling of atoms" (7). This jiggling of atoms is largely random but ultimately built on concrete probabilities of motion that result from the molecule's energy landscape, which is determined by the molecule's structure. If the energies that act upon a system can be quantified, one can also determine the forces that are exerted on the system as force is the negative of the derivative of the potential energy function.

At the heart of biomolecular simulation, one can utilize Newton's second law of motion: force equals mass times acceleration. In this manner forces that act upon a system can be used to find the acceleration of the system, allowing one to build a trajectory of motion for the molecule(s) in question. The potential energy function U is composed of the mathematical representation of all the factors that influence a molecule's motion, such as bonds, bond angles, torsions, Van der Waals forces, and electrostatic interactions. This equation comprises the force field of a molecular dynamics simulation. Using this force field, one can input a structure file of a molecular system, generate initial velocities, and determine a possible trajectory of motion that the molecule(s) can assume. This trajectory can then be used to analyze the dynamics of the system in question.

Analytical Techniques

Molecular dynamics simulation data was analyzed to compute root-mean-square deviation (RMSD) and root-mean-square fluctuation (RMSF) quantities, determination of mutual information to quantify allostery, and the generation of machine learning networks.

RMSD Analysis

The root-mean-square deviation of atomic coordinates is one of the most popular analytical tools for determining the differences between macromolecular structures. It is frequently used to ascertain the quality of molecular dynamics simulations and the equilibration period of a given trajectory (10). To compute the RMSD, first a least-squares fit is done to minimize the difference between two superimposed structures; in the case of determining equilibration of a simulation, these two structures are the initial and subsequent frames of the trajectory. After the two structures are superimposed, the RMSD (in length units) is calculated according to the equation

$$RMSD = \sqrt{1/N \sum_{i=1}^{N} \delta_i^2} \quad \mathbf{1}$$

Where δ_i is the distance between atom i in reference structure and atom i current structure, and N is the total number of equivalent atoms used in the calculation (11). RMSD calculations require identical numbers of atoms in both structures, and for this reason as well as to reduce noise in the results, it is normally computed using only the backbone heavy atoms of each amino acid. RMSD analysis has limitations; for example, it cannot reveal local flexible regions, in other words, it cannot distinguish between a molecule in which some regions are very rigid and some very flexible and a molecule in which all regions are semi-flexible (10). However, RMSD can be used to determine a holistic degree of difference between two structures, allowing one to

iteratively view how a given molecule in a simulation changes from an initial reference point over time.

RMSF Analysis

Root-mean-square fluctuation of atomic positions allows one to determine flexibility of a residue over the course of a simulation. It measures the fluctuation of individual residues relative to a reference position, which is typically the time-averaged position of the amino acid. RMSF analysis can thus help one determine regions of the molecule that are more flexible and regions that are more rigid, whereas RMSD provides a more global metric. RMSF is calculated according to the equation

$$RSMF = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(v_t - \bar{v})^2} \quad \mathbf{2}$$

Where T is frames, v_t is is the coordinate of an atom at time t, and $\bar{v}$ is the ensemble averaged position of the same atom (11).

Mutual Information Analysis

In probability theory, mutual information (MI) refers to the mutual dependence between two random variables. In abstract terms, it quantifies the degree of information obtained about one random variable from another random variable. MI can thus be used to determine how much the movement of one dihedral reveals about the movement of another dihedral, revealing whether these motions are correlated. In this manner, one can measure the correlated motions, or allosteric communication, within a protein. The MI between dihedrals can be summed to find the total MI of the system for a holistic sense of allostery, or MI of the dihedrals in the protein can be measured relative to a target site of interest, for example, amino acids that are crucial for binding.

CARDS software characterizes mutual information I according to the equation

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \quad (3)$$

Where $x \in X$ represents the amount of states dihedral X can adopt, $y \in Y$ represents the amount of states dihedral Y can adopt, p(x) is the probability that dihedral X adopts state x, p(y) is the probability that dihedral Y adopts state y, and p(x,y) is the joint probability that both X adopts state x and Y adopts state y (8). For more information on ordered versus disordered regimes, normalization of MI, and holistic calculations of MI, refer to the supplementary equations. For target site analysis, CARDS takes the average mutual information between two dihedral sets: dihedrals in the reference residue and nearest neighbor (within 3 Angstroms) and all dihedrals in the target site (8).

Machine Learning Analysis

In order to interpret the many degrees of freedom present in a conformational ensemble for a molecular system, several dimensionality reduction algorithms exist that condense these motions into a smaller set of understandable variables. However, one typical limitation of these techniques is that preference is given to distinguishing between large motions instead of smaller ones that may be more relevant to binding dynamics and other properties. For example, principal component analysis (PCA) determines linear combinations of features in order to keep the maximum amount of geometric variance in the original data set, which ultimately assumes that larger structural changes are more relevant than smaller ones (9). There are many molecular systems for which this assumption is faulty, for example, there are many enzymes that have a large loop that undergoes arbitrary motions, which will be interpreted as important by PCA, while subtle motions in the active site are more biochemically relevant. Machine learning

provides an opportunity to overcome this problem. Unlike PCA, which considers linear combinations of features, autoencoders consider nonlinear combinations of features (9). An autoencoder consists of an encoder, which reduces the dimensionality of the input into a latent space, and a decoder, which expands this latent space to construct the original input, in this case, a three-dimensional reconstruction of a protein's configuration (9). The latent space is optimized in order to accurately reconstruct this input. Unfortunately, autoencoders still emphasize the collection of large geometric variations. The literature suggests the use of supervised autoencoders to accurately identify features that differentiate between structures. Supervision adds the requirement that the condensed variables be able to predict a label, for example, whether the structural input came from the wildtype or a given mutant (9). This forces the machine learning network to focus on what degrees of freedom are important for distinguishing between two states rather than automatically putting emphasis on large geometric motions. The classification task in the latent space can be applied to the whole molecule or limited to a specified region, such as an active site, in order to study a particular area. For a schematic representation, see Figure 3.

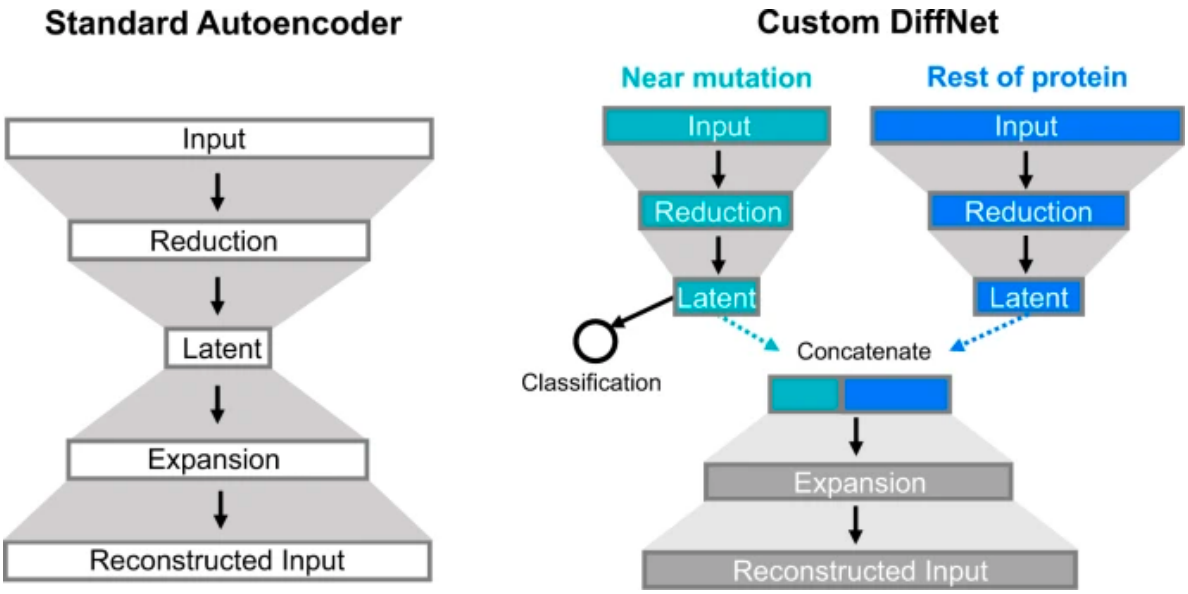


Figure 3: ***Schematic representation of a standard autoencoder on the left and a sample DiffNet on the right.*** *The encoder reduces the input into the latent space and the decoder expands it to reconstruct the input. The custom DiffNet shown in the figure contains two encoders: a supervised encoder that acts upon the atoms near a given mutation (shown in cyan) and an unsupervised encoder that operates on the other atoms in the protein (shown in blue). The latent layers constructed by the two encoders are concatenated and trained in order produce the reconstructed input.*

Input for DiffNets includes a trajectory from one structure and a trajectory from another molecular system, which are then broken down into frames that are used by the software to train a supervised autoencoder to accurately identify the variables that separate the two structures. One challenge faced by this approach is that there is likely to be overlap between the two conformational ensembles, i.e, there are structural states that the two systems may have in common. It is thus unrealistic to classify all states present in one variant as different from the states present in the other variant. In order to avoid this, DiffNets uses an expectation maximization scheme, which iteratively changes training labels to determine a set of structural states that are probabilistically more likely to occur in one variant versus another while allowing the conformational ensembles of the two variants to overlap (9). For some of the mathematics behind DiffNets theory, refer to the supplementary equations.

The Bowman lab at the University of Washington at St. Louis successfully used DiffNets to determine small structural differences that predicted the duty ratios of myosin isoforms and relative stabilities of β-lactamase variants (9). It is hypothesized that DiffNets should be able to identify the relevant structural properties that inform the biochemical variations between other molecular systems. In this study, DiffNets was used to identify the structural signatures that predict enhanced binding in the 3.24 mutant as opposed to the wildtype.

Research Aims

In the analyses of the biomolecular simulation, allosteric communication was quantified in terms of mutual information, a term in probability theory that refers to the dependence between two variables. Mutual information was measured using Correlation of All Rotameric and Dynamical States, or CARDS, software, produced by the Bowman lab at the University of Washington at St. Louis (8). CARDS can be utilized to find the total mutual information of a system and the mutual information in the system relative to a target site of one's choosing. Mutual information identifies the differences between the wildtype and mutant but fails to show whether these are biochemically important and what the functional differences between the wildtype and mutant are. Machine learning can be used to build a network that finds the relevant differences that can accurately classify wildtype versus mutant ensembles. The software DiffNets was used from the Bowman lab at University of Washington at St. Louis, which uses a supervised autoencoder that also contains a classification task (9).

The objectives of this study are as follows:

1. Generate molecular dynamics simulations of the wildtype and mutant 3.24 antibodies with and without hyperphosphorylated tau protein.
2. Use the biomolecular simulations of the wildtype and mutant 3.24 antibodies with and without hyperphosphorylated tau protein to determine the total mutual information of the antibodies and analyze the differences between mutual information of the wildtype and mutant.
3. Determine the mutual information of the wildtype and mutant antibodies relative to key binding sites in the antibody structure.

4. Utilize machine learning methods to identify relevant motions that separate the wildtype and mutant antibodies in terms of biochemical function.

The overarching goal of this study was to try to learn design rules that could inform future antibody design decisions.

Methods

System Preparation

All of the simulations in this study were performed in a Unix-based operating system with access to UConn's High Performance Computing (HPC) cluster utilized for running and processing trajectories. Access to a computer for connecting to the cluster and for simulation analysis was provided through the May lab.

In order to run a molecular dynamics simulation, one needs force field parameters, a structure file that provides the initial Cartesian coordinates of the system, and a method of solvating, relaxing, and equilibrating the molecule(s). For the simulations in this study, CHARMM-GUI (12-13) was used for system setup and solvation. For information on the contents of each simulation (wildtype vs. mutant with and without tau) and simulation length, refer to Table 1. The structure files for simulations are most often taken from the Protein DataBase (pdb) where scientists upload molecular structures that they have determined for open access. For this study, the wildtype antibody structure was taken from file 4GLR, and the mutant was generated by inserting point mutations at the appropriate residues using CHARMM-GUI software. There were 3260 atoms in the wildtype structure, 3266 in the mutant, and 175 in the tau epitope.

For all systems, the CHARMM-36m (14) force field and TIP3P (15) water model were used. Each solvation box was fit to the system with 10 Angstrom buffers on all sides. NaCl was added to reach a concentration of 150 mM, resulting in an average total system size of approximately 36,000 atoms. GROMACS (16-17) software was used to perform energy minimization over 10,000 steps. For NVT equilibration (keeping volume and temperature constant), temperature was held at 310.15 K for 25,000 steps with rvdw_switch of 1.0 nm, rvdw of 1.2 nm, and dt of 0.001 ps. The Nose-Hoover thermostat was used with full electrostatic interactions calculated via the PME method.

Production Runs

For the production runs, NPT equilibration was performed (keeping the pressure and temperature constant) using the Parrinello-Rahman barostat with isotropic scaling at 1 atm and Nose-Hoover thermostat, the same water model and force field as above, PME, and a Verlet cutoff scheme with short-range nonbonded interactions calculated with a cutoff of 12 Angstroms. A timestep of 2 fs was used with frames saved every 10 ps. For details regarding individual simulations, refer to Table 1. After the generation of the trajectories, each system was processed to correct for periodic boundary conditions.

Table 1: ***Simulation systems and legnths.*** *All systems were run for 2 trials of 2 μs each. The trials were concatenated to build 4 μs trajectories for each system.*

Simulation Number	**Contents**	**Length**
1	Wildtype antibody and tau	2 μs
2	Wildtype antibody and tau	2 μs
3	Wildtype antibody only	2 μs
4	Wildtype antibody only	2 μs
5	Mutant antibody and tau	2 μs

6	Mutant antibody and tau	2 μs
7	Mutant antibody only	2 μs
8	Mutant antibody only	2 μs

Methodology and System Validation

In order to determine the stability of the systems, RMSD (backbone after least squares fit to backbone) and RMSF analyses were performed using concatenated trajectories. Referring to Table 1, trials 1 and 2 were combined to produce 4 μs of simulation for the wildtype and tau system, trials 3 and 4 were combined to produce 4 μs of simulation for the wildtype only system, trials 5 and 6 were combined to produce 4 μs of simulation for the mutant and tau system, and trials 7 and 8 were combined to produce 4 μs of simulation for the mutant only system.

Mutual Information

Holistic mutual information was calculated for each system's 4 μs concatenated trajectories using CARDS software. The first 100 ns of the individual trials were removed in order to give the system time to equilibrate. In order to test the convergence of MI, the 4 μs trajectory of the wildtype bound to tau was used to create smaller trajectories with lengths of 500 ns, 900 ns, 1500 ns, 1900 ns, and 3800 ns, respectively. The total holistic MI of each trajectory was determined and plotted to determine the minimum trajectory length needed to measure MI, which was determined to be 1900 ns (see Supplementary Figure 1).

MI was calculated for each 4 μs trajectory relative to key target sites. A 1900 ns trajectory for each system was used to measure the MI relative to each CDR in the antibody. Preliminary results identified three target sites for which relative mutual information was significantly different in the mutant than the wildtype: CDR-H2, the portion of CDR-H2 that directly surrounds the phosphate group, and CDR-L3 (see Supplementary Table 1). MI was then

determined for each 4 μs trajectory (with the first 100 ns removed from each trial) relative to these sites.

In order to test whether all the five mutations present were needed for enhanced mutual information, five new antibodies were generated using CHARMM-GUI software in which one point mutation was removed in each with the other 4 mutations seen in the original mutant present. For each system, a different point mutation was removed (see Table 2). 2 μs of simulation was run for each system with the same setup as the previous experiments. The first 100 ns was removed from each trajectory, total MI was calculated, and MI relative to the chosen target sites was determined.

Table 2: ***Construction of new mutants with one mutation in the 3.24 mutant removed per system.*** *In each structure, a different mutation was removed.*

System Number	Mutation Missing
1	THR to ALA, AA 41, light chain
2	ASP to GLY, AA 52, light chain
3	THR to ILE, AA 61, light chain
4	GLY to CYS, AA 75, heavy chain
5	ALA to VAL, AA 97, heavy chain

DiffNets

For the systems that included tau, the trajectories were processed so that only the antibody was included. Two DiffNets were generated with one being trained with the wildtype and mutant trajectories from the tau-bound systems and the other with the wildtype and mutant trajectories from the unbound systems. A third DiffNet was constructed with the mutant

trajectory from the tau-bound system as one set of input and the mutant simulation from the unbound system as the other. 4 μs trajectories were used as input for all systems.

Results and Discussion

System Validation

RMSD analysis of the backbone of each system fit to the backbone reveal that the structures in each trajectory were overall stable, with all system trajectories having a maximum RMSD of no more than 0.3 nm, or 3 Angstroms (see Figure 4). To observe more local fluctuations, RMSF analysis was performed for each system (wildtype bound, wildtype unbound, 3.24 mutant bound, and 3.24 mutant unbound), revealing a maximum RMSF of no more than 0.55 nm, or 5.5 Angstroms, for each residue in each system (see Figure 5). RMSD and RMSF results reveal that the antibodies in each system were stable and thus likely a reliable representation of the molecules in question.

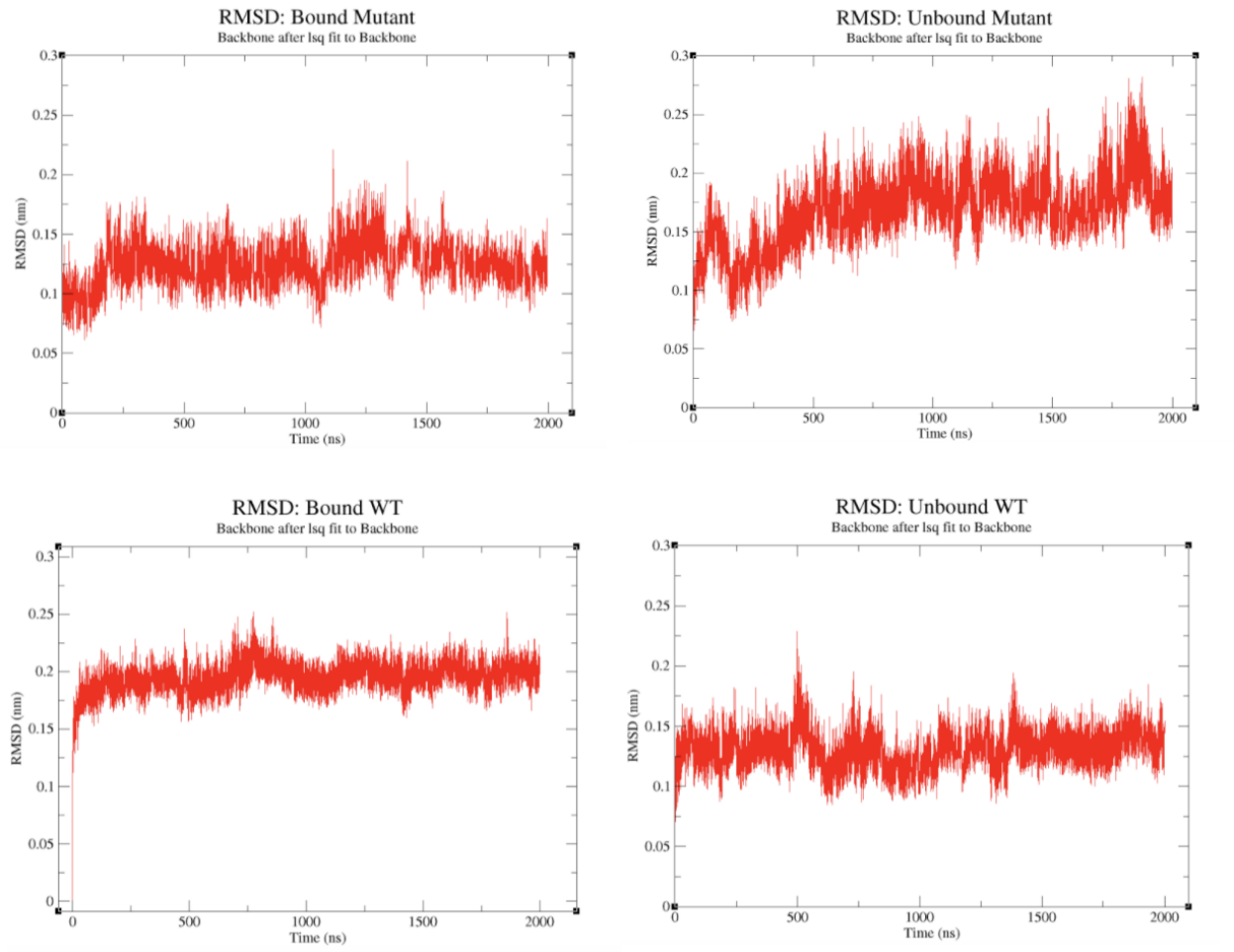


Figure 4: ***RMSD of the wildtype and mutant antibodies in Tau-bound and unbound simulations.*** *The backbone of each structure was fit to the initial structure. All structures were determined to be overall stable with a maximum RMSD of no more than 0.3 nm, or 3 Angstroms (maximum RMSD seen in the unbound mutant).*

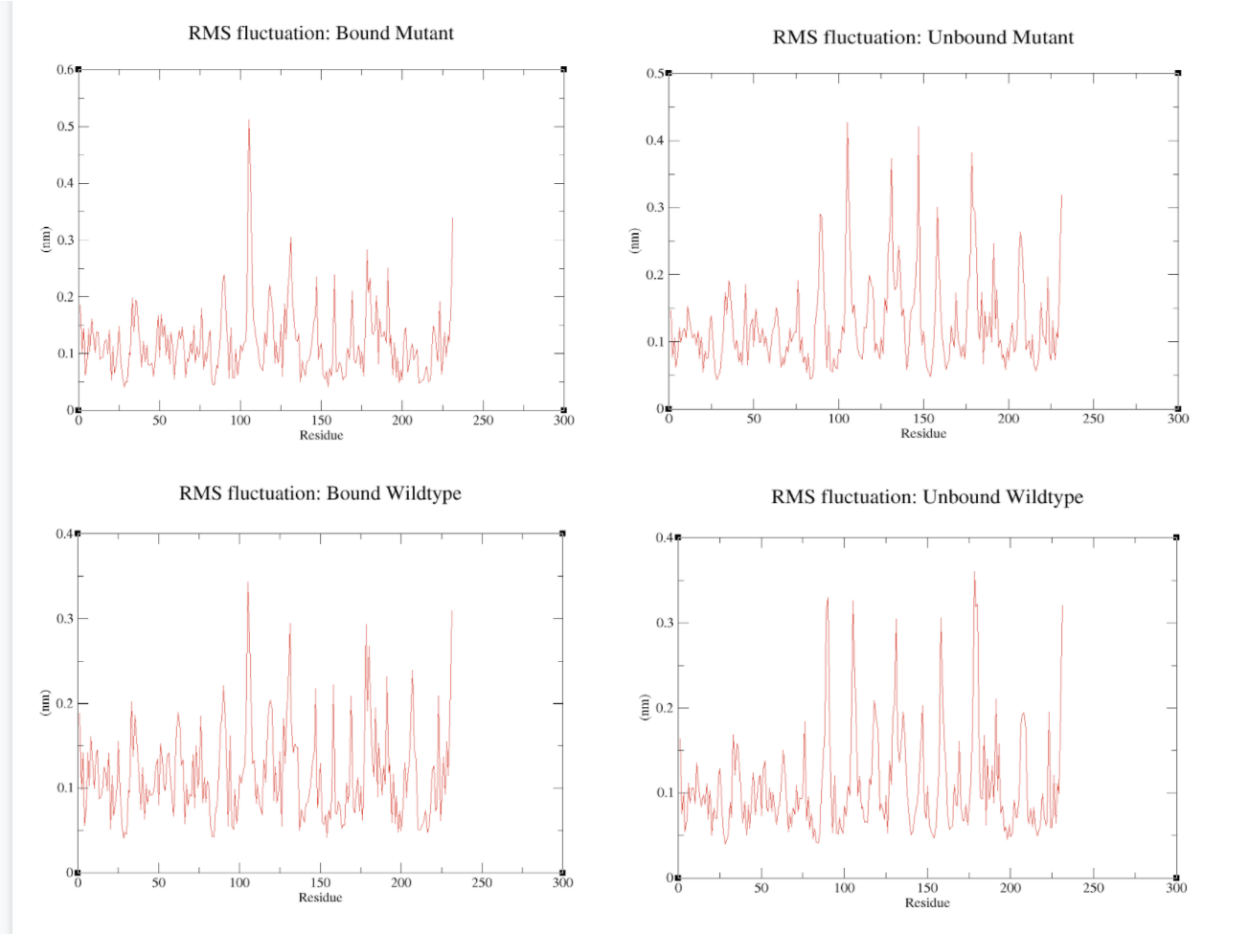


Figure 5: ***RMSF of residues in the wildtype and mutant antibody in Tau-bound and unbound simulations.*** *A maximum RMSF of no more than 0.55 nm was observed (maximum RMSF seen in the bound mutant).*

Mutual Information and Target Site Analysis

For the wildtype and the mutant bound and unbound, the holistic mutual information was summed to find the total MI of each system. The data reveals elevated MI in the mutant as opposed to the wildtype when comparing both the bound and unbound states (see Figure 6A). For each system, the total MI relative to CDR-H2, the portion of CDR-H2 that forms a cage around the phosphate group on tau, and CDR-L3 were summed. One can see elevated MI in the 3.24 mutant compared to the wildtype in the bound states for all target sites and enhanced MI in

the mutant relative to the portion of CDR-H2 that binds to the phosphate group on tau and CDR-L3 in the unbound states (see Figure 6B and 6C).

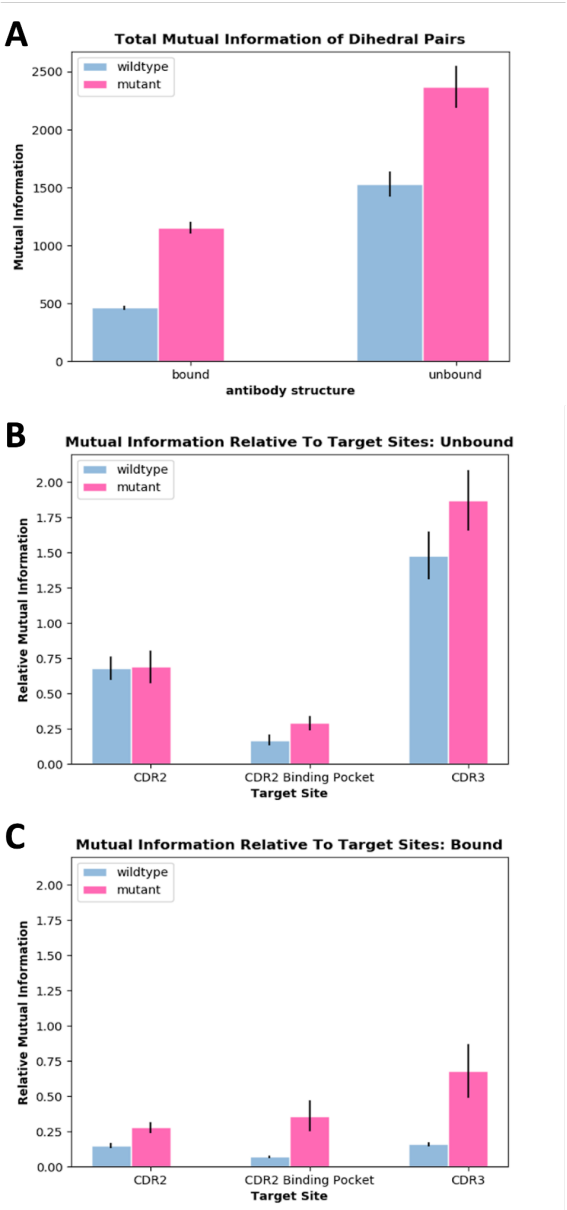


Figure 6: ***Mutual information of the wildtype and 3.24 mutant bound and unbound.*** *A) total mutual information of the wildtype and mutant bound and unbound. In the mutant, MI is elevated in both the bound and unbound states. B) MI of the wildtype and mutant in the unbound states relative to CDR-H2 (CDR2), the CDR-H2 binding pocket (CDR2 binding pocket), and CDR-L3 (CDR3). One can see a significant degree of elevated MI in the mutant compared to the wildtype relative to the CDR-H2 binding pocket and CDR-L3. C) MI of the wildtype and mutant in the bound states relative to CDR-H2 (CDR2), the CDR-H2 binding pocket (CDR2 binding pocket), and CDR-L3 (CDR3). One can see a significant degree of elevated MI in the mutant compared to*

the wildtype relative to all target sites. Values determined via a bootstrapping procedure in which each trajectory was broken up into 50 ns increments and put back together at random for 20 new trajectories. The average MI of each trajectory was reported with error bars reflecting standard deviation between the rebuilt trajectories.

The MI was also investigated for each individual residue in mutant 3.24 (bound state). With the exception of mutation THR to ILE, all the point mutations corresponded to regions of elevated MI relative to the portion of CDR-H2 that binds to the phosphate group on tau (see Figure 7), potentially revealing a network of allosteric communication.

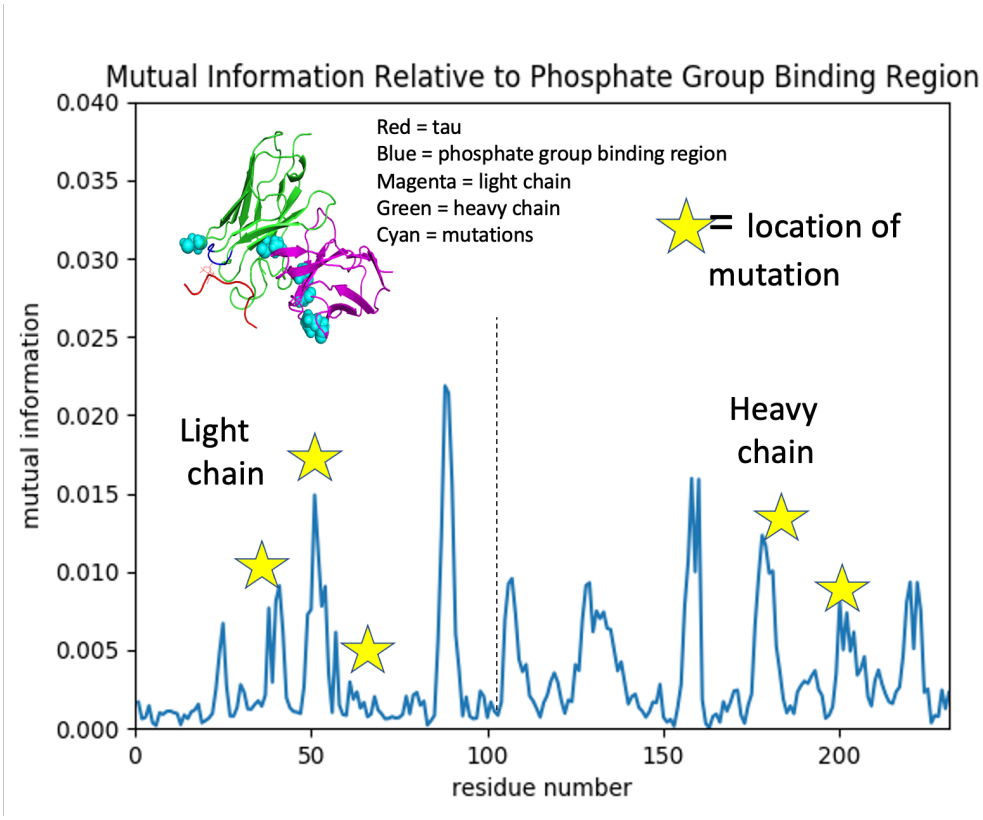


Figure 7: ***Mutual information of residues in the bound mutant relative to the portion of CDR-H2 that binds to the phosphate group.*** *Four of the five mutations (Ala41, Gly52, Cys179, and Val201) exhibit elevated MI compared to the other residues in the protein.*

The same analyses were performed on the five new mutants (where one mutation was removed) generated from the 3.24 mutant structure. The total MI of the mutants in the unbound state were all less than that of the 3.24 mutant, demonstrating that all mutations were necessary

for enhanced total MI (see Figure 8A). The MI relative to the identified target sites was also calculated for the five new systems in the unbound state (Figure 8B) and the bound state (Figure 8C). One can see that with the exception of system 1, all of the new mutants displayed less MI than the 3.24 mutant relative to the target sites that paralleled the MI levels of the wildtype. Since the total MI of system 1 in the unbound state was significantly less than that of mutant 3.24 and comparable with that of the wildtype, the data revealed that overall, all mutations were necessary for enhanced MI.

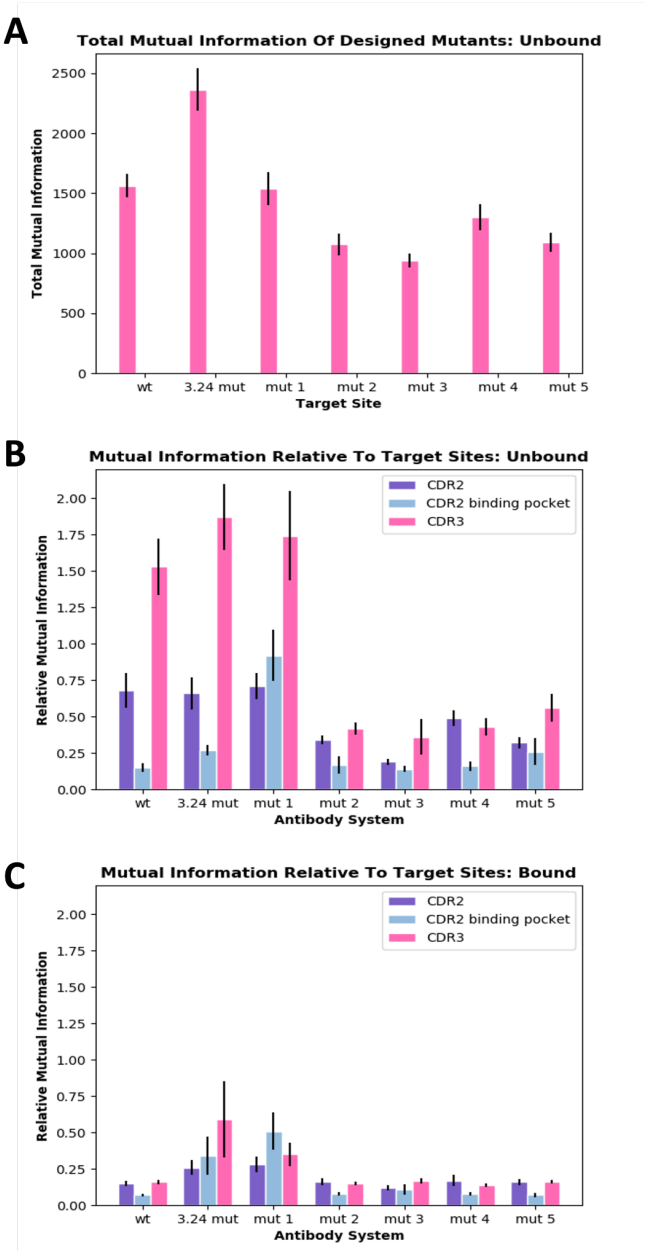


Figure 8: ***Mutual information of the wildtype, 3.24 mutant and new mutants.*** *A) total MI of antibody structures in the unbound state. B) MI of antibody structures in the unbound state relative to target sites CDR2 (CDR-H2), the CDR2 binding pocket (CDR-H2 binding pocket), and CDR3 (CDR-L3). C) MI of antibody structures in the bound state relative to the aforementioned target sites.*

DiffNets Analysis

Construction of DiffNets revealed key expansions in the 3.24 mutant compared to the wildtype, allowing one to visualize potential allosteric communication that is enhanced in the mutant. When the DiffNets architecture determined relevant motions distinguishing mutant from wildtype, it revealed expansions emanating from Ile61 (one of the mutations) in both the bound and the unbound states (see Figure 9). In the bound state, expansion also occurred near another mutation (Ala201). These motions may aid in the identification of a binding mechanism that links the distant framework region to the binding site, which could inform antibody design.

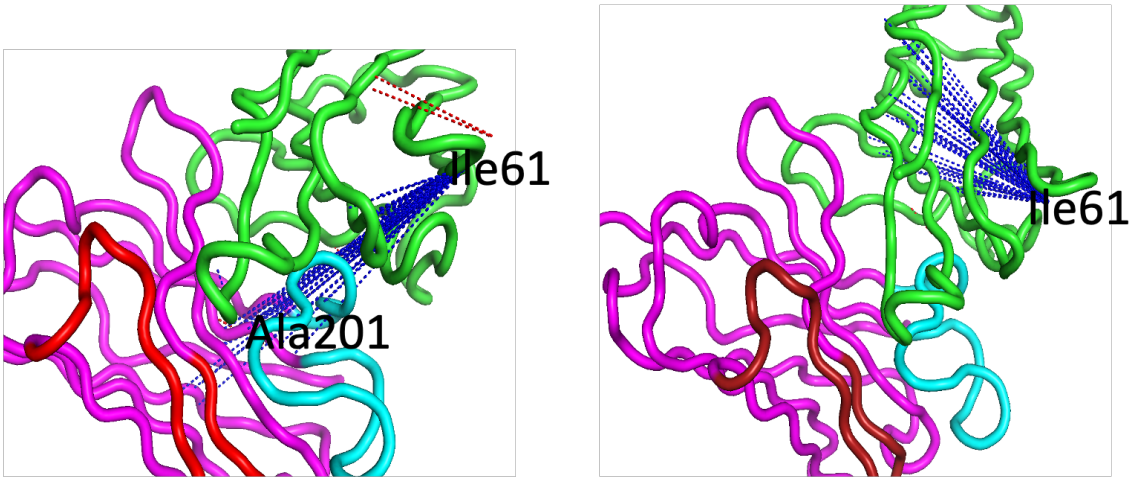


Figure 9: ***Relevant motions distinguishing 3.24 mutant from wildtype.*** *The DiffNets detected expansion in the mutant relative to the wildtype emanating from Ile61 in the bound (left) and unbound (right) states. The blue lines signify expansion, and the red lines signify compression.*

Another DiffNet was constructed in which the 3.24 mutant in the bound state was used as one set of input and the mutant in the unbound state was used as the other. A morphed structure of the mutant from the unbound to bound state revealed high variation in the region closest to the tau epitope, indicating that the configuration of these binding loops is different in the unbound state to the bound state (see Figure 10). If one plays a trajectory of ten frames moving from most unbound character to most bound character, one can see the loops closing in like a clamp,

potentially revealing one portion of the binding mechanism that, if discovered in its entirety and linked to the expansion seen in the rest of the protein, could inform intentional drug design.

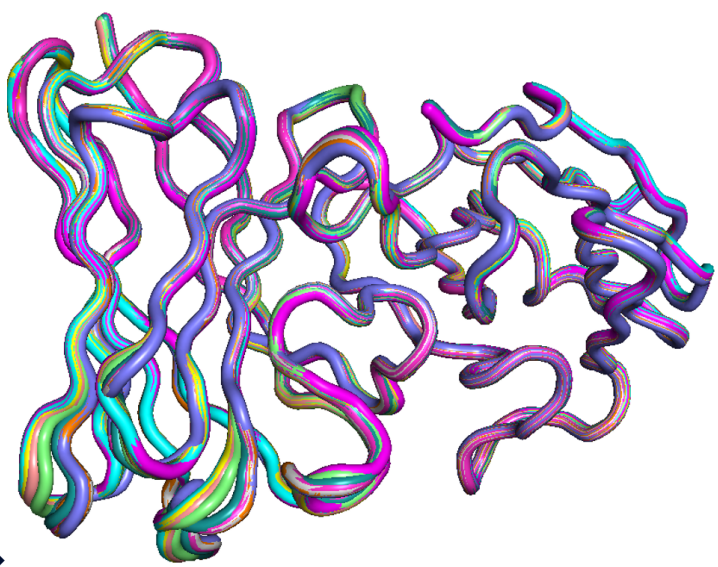


Figure 10: ***Morphed structure of the mutant from the unbound to the bound state.*** *Ten structures were aligned that were classified as various degrees of "1", or bound, character as opposed to "0", or unbound, character. Regions of high variation can be seen in the bottom loops, which appear to form a binding clamp.*

Future Directions

Results from this study identify enhanced allosteric communication in the 3.24 mutant that may explain higher binding affinity to hyperphosphorylated tau protein and reveal potential expansions that could comprise a step in the network linking the framework region with the point mutations to the binding site. Future steps would be to further characterize these allosteric communication networks in order to identify collective variables that could be used in free energy calculations. Steered molecular dynamics simulations, or simulations that exert a force on a region of interest rather than simply equilibrating the system, can reveal the binding energetics between two areas and could be used to identify a residue that can be changed in order to increase the binding affinity of the system. Once a new mutant is designed, steered MD can be

used to investigate its binding affinity to hyperphosphorylated tau. In order to test the specificity of the antibody, steered MD can be used to determine the binding affinity to normal tau, which should remain low in order to be a good candidate.

Conclusions

It can be concluded that the 3.24 mutant contains increased total allosteric communication compared to the wildtype for both the bound and unbound states as determined by mutual information calculations. It can also be concluded that allosteric communication was increased in the mutant relative to key binding sites compared to the wildtype for both the bound and unbound states. All five of the point mutations seen in the 3.24 mutant were determined to be necessary for increased allosteric communication. Finally, it can be concluded that increased allosteric communication can be seen as long-range expansions in the mutant compared to the wildtype in both the bound and unbound states. This expansion was shown to be emanating from mutation Ile61.

Acknowledgements

I would like to thank my Honors and University Scholar advisor, thesis advisor, and mentor Dr. Eric May for his guidance, instruction, and encouragement as well as the other members of the May lab for their advice and support. In addition, I would like to thank the other members of my University Scholar committee: Dr. Yongku Cho and Dr. Adam Zweifach.

I would also like to acknowledge the computational resources provided through the UConn HPC cluster as well as the Office of Undergraduate Research for their generous funding through a SURF Award.

Finally, I would like to thank Florence, Daniel, Jacob, and Harley Lee, Joshua King, Shivani Padhi, Anika Veeraraghav, and Karen Haskins for their support.

References

(1)
Oueslati, A.; Fournier, M.; Lashuel, H. A. Role of Post-Translational Modifications in Modulating the Structure, Function and Toxicity of α-Synuclein. In *Progress in Brain Research*; Elsevier, 2010; Vol. 183, pp 115–145. https://doi.org/10.1016/S0079-6123(10)83007-9.
(2)
Neumann, M.; Kwong, L. K.; Sampathu, D. M.; Trojanowski, J. Q.; Lee, V. M.-Y. TDP-43 Proteinopathy in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis: Protein Misfolding Diseases Without Amyloidosis. *Arch Neurol* **2007**, *64* (10), 1388. https://doi.org/10.1001/archneur.64.10.1388.
(3)
Wang, Y.; Lin, F.; Qin, Z.-H. The Role of Post-Translational Modifications of Huntingtin in the Pathogenesis of Huntington's Disease. *Neurosci. Bull.* **2010**, *26* (2), 153–162. https://doi.org/10.1007/s12264-010-1118-6.
(4)
Gong, C.-X.; Liu, F.; Grundke-Iqbal, I.; Iqbal, K. Post-Translational Modifications of Tau Protein in Alzheimer's Disease. *J Neural Transm* **2005**, *112* (6), 813–838. https://doi.org/10.1007/s00702-004-0221-0.
(5)
Li, D.; Wang, L.; Maziuk, B. F.; Yao, X.; Wolozin, B.; Cho, Y. K. Directed Evolution of a Picomolar-Affinity, High-Specificity Antibody Targeting Phosphorylated Tau. *J. Biol. Chem.* **2018**, *293* (31), 12081–12094. https://doi.org/10.1074/jbc.RA118.003557.
(6)
Shih, H. H.; Tu, C.; Cao, W.; Klein, A.; Ramsey, R.; Fennell, B. J.; Lambert, M.; Ní Shúilleabháin, D.; Autin, B.; Kouranova, E.; Laxmanan, S.; Braithwaite, S.; Wu, L.; Ait-Zahra, M.; Milici, A. J.; Dumin, J. A.; LaVallie, E. R.; Arai, M.; Corcoran, C.; Paulsen, J. E.; Gill, D.; Cunningham, O.; Bard, J.; Mosyak, L.; Finlay, W. J. J. An Ultra-Specific Avian Antibody to Phosphorylated Tau Protein Reveals a Unique Mechanism for Phosphoepitope Recognition. *Journal of Biological Chemistry* **2012**, *287* (53), 44425–44434. https://doi.org/10.1074/jbc.M112.415935.
(7)
Mulholland, A. J. Introduction. Biomolecular Simulation. *J. R. Soc. Interface.* **2008**, *5* (suppl_3), 169–172. https://doi.org/10.1098/rsif.2008.0385.focus.
(8)
Singh, S.; Bowman, G. R. Quantifying Allosteric Communication via Both Concerted Structural Changes and Conformational Disorder with CARDS. *J. Chem. Theory Comput.* **2017**, *13* (4), 1509–1517. https://doi.org/10.1021/acs.jctc.6b01181.
(9)

Ward, M. D.; Zimmerman, M. I.; Meller, A.; Chung, M.; Swamidass, S. J.; Bowman, G. R. Deep Learning the Structural Determinants of Protein Biochemical Properties by Comparing Structural Ensembles with DiffNets. *Nat Commun* **2021**, *12* (1), 3023. https://doi.org/10.1038/s41467-021-23246-1.

(10)

Sargsyan, K.; Grauffel, C.; Lim, C. How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2017**, *13* (4), 1518–1524. https://doi.org/10.1021/acs.jctc.7b00028.

(11)

Khezri, A.; Karimi, A.; Yazdian, F.; Jokar, M.; Mofradnia, S. R.; Rashedi, H.; Tavakoli, Z. Molecular Dynamic of Curcumin/Chitosan Interaction Using a Computational Molecular Approach: Emphasis on Biofilm Reduction. *International Journal of Biological Macromolecules* **2018**, *114*, 972–978. https://doi.org/10.1016/j.ijbiomac.2018.03.100.

(12)

Jo, S.; Cheng, X.; Lee, J.; Kim, S.; Park, S.; Patel, D. S.; Beaven, A. H.; Lee, K. I.; Rui, H.; Park, S.; Lee, H. S.; Roux, B.; MacKerell, A. D.; Klauda, J. B.; Qi, Y.; Im, W. CHARMM-GUI 10 Years for Biomolecular Modeling and Simulation. *J. Comput. Chem.* **2017**, *38* (15), 1114–1124. https://doi.org/10.1002/jcc.24660.

(13)

Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29* (11), 1859–1865. https://doi.org/10.1002/jcc.20945.

(14)

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* **2017**, *14* (1), 71–73. https://doi.org/10.1038/nmeth.4067.

(15)

Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105* (43), 9954–9960. https://doi.org/10.1021/jp003020w.

(16)

Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718. https://doi.org/10.1002/jcc.20291.

(17)

Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

Supplementary Material

Mutual Information: see (8) for reference

CARDS software calculated dihedral angles with MDTraj, assigned them to rotameric states gauche+, gauche-, and trans for most χ angles and cis or trans for backbone dihedrals. Transition-Based Assignment (TBA) was used to delineate lasting transitions from transient motions. CARDS determines if a length of time between two transitions is more ordered or disordered using a liklihood ratio L(t) in which a value of L(t) greater than 3 to characterize disorder and any value of L(t) less than 3 is considered ordered.

See Introduction to Analytical Techniques for the calculation of mutual information I(X,Y).

Mutual information is normalized through the equation

$$\overline{I(X,Y)} = I(X,Y)/C(X,Y) \quad \mathbf{4}$$

Where C(x,y) is the channel capacity, or maximum MI between two dihedrals. It accounts for the fact that different dihedrals have a different number of available states.

The holistic correlations between two dihedrals are characterized by the equation

$$I_H(X,Y) = \overline{I_{ss}(X,Y)} + \overline{I_{sd}(X,Y)} + \overline{I_{ds}(X,Y)} + \overline{I_{dd}(X,Y)} \quad \mathbf{5}$$

Where (moving from left to right) the first term is the normalized MI between two rotameric (ordered) states, the second term is the normalized MI between dihedral X's romateric state and dihedral Y's dynamical (disordered) state, the third is the normalized MI between the dynamical state of X and rotameric state of Y, and the fourth term is the normalized MI between the dynamical state of X and that of Y. For calculating the MI relative to a target site, CARDS takes the average MI between two dihedral sets: dihedrals in the reference residue and nearest neighbor (within 3 Angstroms) and all dihedrals that are present in the target site.

Diffnets: see (9) for reference

For data pre-processing, the trajectories and structure files are stripped down to the backbone coordinates without carbonyl oxygens. The trajectories are centered and then aligned to the crystal structure. The coordinates are mean-shifted to 0 via the equation

$$x^{mean-free} = \sum_{i=1}^{N_t} x_i - \bar{x} \quad (6)$$

Where $x^{mean\text{-}free}$ is the mean-shifted trajectory, x_i is a single frame, $\bar{x}$ is the mean of the XYZ coordinates, and N_t is the number of frames in the trajectory.

Data was whitened using the equation

$$\tilde{x} = C_{00}^{-\frac{1}{2}} x^{mean-free} \quad (7)$$

Where $\tilde{x}$ is the whitened trajectory and C_{00} is the covariance matrix for the XYZ coordinates.

For the training step, three loss functions were used to minimize reconstruction error, classification error, and correlation of latent space variables, respectively.

$$\mathcal{L}_{DiffNet} = \ell_{Recon} + \ell_{Class} + \ell_{Corr} \quad (8)$$

Reconstruction error loss (which discourages outliers and directs reconstructions to the proper XYZ coordinates) is expressed by the equation

$$\ell_{Recon} = \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{N_n} \sum_{j=1}^{N_n} \left[\left| x_{ij} - \hat{x}_{ij} \right| + \left(x_{ij} - \hat{x}_{ij} \right)^2 \right] \quad (9)$$

Where N_n is the number of output nodes, N_b is the number of examples in a training batch, x_{ij} is a target XYZ coordinate, and. $\widehat{x_{ij}}$ is the output value.

The classification term, which penalizes errors made by the latent space, is expressed as

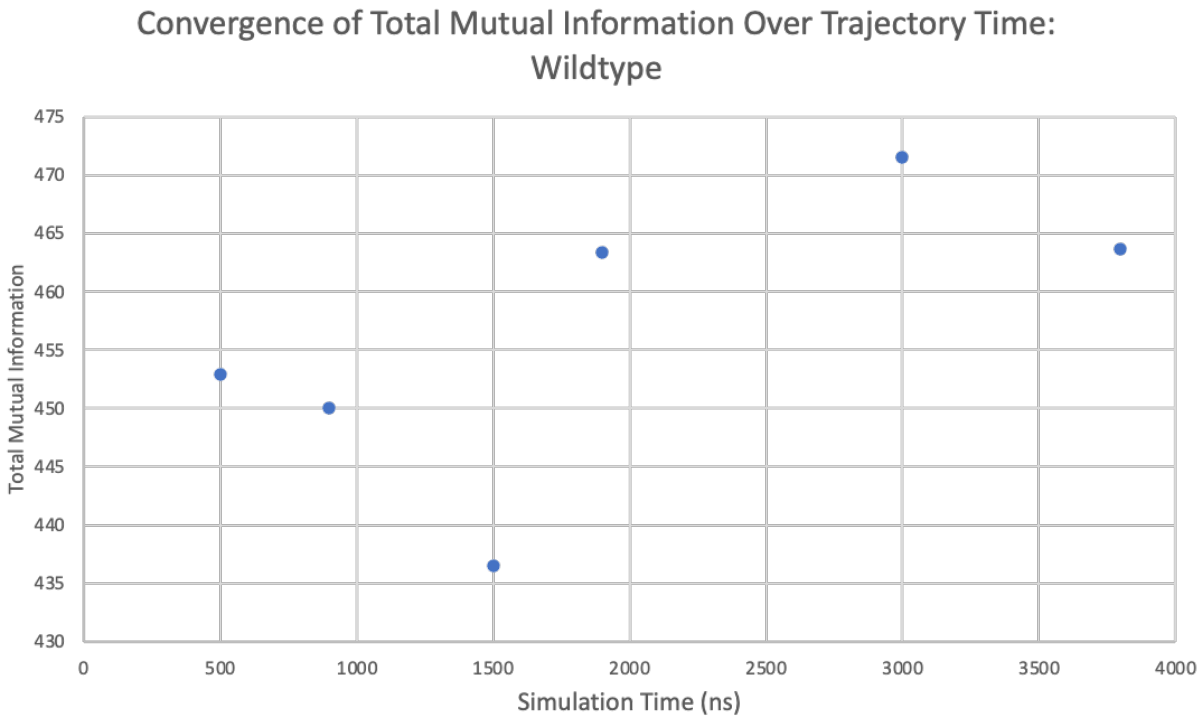
$$\ell_{Class} = \frac{1}{N_b}\sum_{i=1}^{N_b} y_i * \log(\hat{y}_i) + (1 - y_i) * \log\,(1 - \hat{y}_i)$$

Where N_b is the number of examples in a training batch, y_i is the target value, and $\hat{y}_i$ is the output of the classification layer.

The last term, which minimizes the covariance between the latent space variables, is expressed via the equation

$$\ell_{Corr} = \sum_{i \neq j} Cov(z_i, z_j)^2$$

Where $Cov(Z_i, Z_f)$ is the covariance matrix of Z (the latent vector) across all samples in the training batch.



Supplementary Figure 1: ***Convergence of total MI for bound wildtype.*** *One can see that the total MI converges at approximately 1900 ns.*

Supplementary Table 1: ***Mutual information of the wildtype and 3.24 mutant bound and unbound for key target sites.*** *The data reveals that the greatest differences in MI for the bound states are relative to CDR-H2 and CDR-L3, which informed future target site analysis. MI was calculated over 1900 ns trajectories for each system.*

Mutual Information Relative to Target Sites						
	mutant bound	wildtype bound	difference	mutant unbound	wildtype unbound	difference
CDR-H2	0.26	0.33	-0.07	0.66	0.68	-0.02
CDR-H2 binding pocket	0.48	0.16	0.32	1.71	1.69	0.02
CDR-H3	0.28	0.23	0.05	1.14	0.4	0.74
CDR-H3 epitope stabilizers	0.2	0.23	-0.03	0.79	0.17	0.62
CDR-L1	0.25	0.42	-0.17	0.37	0.24	0.13
CDR-L2	0.25	0.31	-0.06	0.56	0.25	0.31
CDR-L3	0.59	0.15	0.44	1.87	1.53	0.34